Knowledge-Intensive, High-Performance Relation Extraction

vorgelegt von
Dipl.-Inf.
Sebastian Krause
geb. in Halle (Saale)

von der Fakultät IV - Elektrotechnik und Informatik
der Technischen Universität Berlin
zur Erlangung des akademischen Grades

Doktor der Ingenieurwissenschaften
- Dr.-Ing. -

genehmigte Dissertation

Promotionsausschuss:

Vorsitzender:	Prof. Dr. Odej Kao
Gutachter:	Prof. Dr. Volker Markl
Gutachter:	Prof. Dr. Hans Uszkoreit
Gutachter:	Dr. Shinichi Nakajima

Tag der wissenschaftlichen Aussprache: 27. September 2017

Berlin 2018

Acknowledgements

This dissertation would not have been possible without the encouragement, frequent feedback, and friendly support of many people. First and foremost, I would like to thank Prof. Hans Uszkoreit and Feiyu Xu, who have supervised this thesis. They have escorted me on various stages of my academic education and introduced me to computational linguistics in 2009. Furthermore, I would like to express my gratitude to Prof. Volker Markl, who kindly agreed to review this thesis and to be my supervisor at TU Berlin. Also, I would like to thank Shinichi Nakajima for consenting to be another reviewer of my dissertation and to serve as a doctoral-committee member.

More people that need to be mentioned here are my colleagues at the DFKI Language Technology Lab in Berlin, with whom I genuinely enjoyed working together. It is rare that one can be a part of a team that offers such an excellent research environment, with plenty of people to learn from in various aspects of a researcher's work. I particularly benefitted from the many fruitful discussions on interesting topics, which often emerged in the context of projects on which we collaborated. Many of these people have also provided valuable comments and actionable feedback for draft chapters of this thesis. The following is a likely incomplete and not particularly ordered list of these colleagues: Renlong Ai, Aleksandra Gabryszak, Leonhard Hennig, Hong Li, Philippe Thomas, Sarah Weichert, Dirk Weissenborn.

Over the years as a doctoral student, I have had the chance to collaborate with several people outside of DFKI, the interaction with whom helped me shape this thesis. In particular, thanks are due to Enrique Alfonseca, Daniele Pighin, Katja Filippova, and Mikhail Kozhevnikov, who hosted me during two research internships at Google's Zurich site. Also, I must mention Andrea Moro and Roberto Navigli from the Sapienza University of Rome, who were project partners at a Google-sponsored research project on topics closely related to this thesis. Finally, I would like to acknowledge the financial support of the Software Campus Initiative, who funded a one-year research project that allowed me to hire student assistants for annotation work.

The content of Chapters 3–7 is based on research projects conducted in collaboration with several colleagues, who intellectually contributed to the research presented here. I would like to thank the following persons for their contributions: Enrique Alfonseca, Katja Filippova, Aleksandra Gabryszak, Leonhard Hennig, Hong Li, Andrea Moro, Roberto Navigli, Daniele Pighin, Hans Uszkoreit, Dirk Weissenborn, and Feiyu Xu.

Abstract

Research on information extraction (IE) from texts has attracted much attention for at least the past two decades. This is not surprising given its significance for applications such as personal digital assistants. Information extraction and its subtask relation extraction play a central role in data processing pipelines that make hidden knowledge such as the content of news articles available to downstream users. This thesis presents four main contributions to important questions of the corresponding research field.

The first two contributions deal with various aspects of the automatic discovery of linguistic patterns, which we use for the detection of relations. We initially look at scenarios with predefined relations of interest. Here, state-of-the-art methods employ simplistic assumptions at training time, which has a drastic negative effect on both precision and coverage. We propose methods for the production and filtering of patterns that mitigate this shortcoming by leveraging existing knowledge about the target domains. Next, we address scenarios without a-priori relation definitions. Here, produced linguistic patterns need to be disambiguated to resolve their meaning, which is particularly hard for patterns in the long tail, which tend to get misinterpreted. Our proposed solution for this issue is the implementation of a global model that can generalize over many pattern occurrences and thus manages to handle rare patterns as well.

The third contribution of this thesis focuses on the versatility of linguistic patterns beyond their designated use for extraction purposes. The patterns convey interesting information about the actual usage of language expressions, which is exactly what is missing in the current landscape of IE-relevant resources. More specifically, the relational information from world-knowledge graphs is not at all grounded in the language information present in lexical-semantic resources. We aim to remedy this deficit by proposing a construction methodology for a new kind of resource that is created by transforming many linguistic patterns into a single graph of language expressions.

Finally, in the fourth contribution, we consider a fundamental shortcoming in the construction of systems for relation extraction, be they based on linguistic patterns or a different methodology. This flaw is the invalid premise that relational information is mostly contained within the boundaries of individual sentences. We initially address this problem with an analysis of its severity and follow-up by designing an approach that can easily be used to post-process the output of existing extraction systems and that allows them to produce cross-sentence relation mentions, and thereby resolves the design flaw.

Zusammenfassung

Die Forschung zur Informationsgewinnung aus Texten erregt seit mindestens zwei Jahrzehnten viel Aufmerksamkeit. Dies ist nicht überraschend angesichts des praktischen Nutzens, den sie für Anwendungen wie digitale Assistenzsysteme mit sich bringen. Informationsextraktion (IE) und das Teilgebiet Relationsextraktion spielen eine zentrale Rolle in Datenverarbeitungspipelines, welche strukturiertes Wissen aus unstrukturierten Quellen wie Nachrichtenartikeln gewinnen. Die vorliegende Arbeit präsentiert vier Hauptbeiträge zu wichtigen Fragen dieses Forschungsfeldes.

Die ersten zwei Beiträge beschäftigen sich mit der automatischen Entdeckung sprachlicher Muster, welche für die Erkennung von Relationen verwendet werden. Wir betrachten zunächst Szenarien mit vorgegebenen Zieldomänen. Der Lernprozess aktueller Systeme in diesem Gebiet basiert auf stark vereinfachenden Abstraktionen, was Präzision und Abdeckung negativ beeinflusst. In dieser Arbeit beschreiben wir Methoden für die Generierung und Filterung von sprachlichen Mustern, die diesen Mangel beseitigen, indem sie vorhandenes Wissen über die Zieldomänen ausnutzen. Als nächstes behandeln wir Szenarien mit flexiblen Zieldomänen. Hier müssen gefundene sprachliche Muster gegeneinander disambiguiert werden, was mit heutigen Methoden besonders für Muster im Long-Tail der Häufigkeitsverteilung zu schlechten Resultaten führt. Zur Lösung dieses Problems schlagen wir die Verwendung eines globalen Modells vor, das über viele Mustererwähnungen verallgemeinert und dem es somit gelingt, seltene Muster korrekt zu interpretieren.

Der dritte Beitrag dieser Arbeit konzentriert sich auf Verwendungszwecke der sprachlichen Muster, die über deren originäre Verwendung hinausgehen. Beispielsweise können die Muster als Quelle für Statistiken über den tatsächlichen Sprachgebrauch von Ausdrücken dienen. Ferner existiert derzeit unter den verfügbaren IE-relevanten Ressourcen nur wenig sprachliches Wissen. Hierzu präsentieren wir eine Konstruktionsmethodik für eine neue Art von Ressource, die durch die Umwandlung vieler linguistischer Muster in einen zusammenhängenden Graphen von sprachlichen Ausdrücken geschaffen wird. Der abschließende vierte Beitrag adressiert einen fundamentalen Konstruktionsfehler von heutigen Relationsextraktionssystemen: Die ungültige Prämisse, dass sich relationale Informationen auf einzelne Sätze beschränken. Wir analysieren zunächst die Relevanz dieses Problems und entwickeln dann einen Ansatz, der es bestehenden Extraktionssystemen erlaubt, satzübergreifende Relationsextraktion auf der Ebene von Dokumenten durchzuführen.

Contents

List of Abbreviations

ACE . . . . . Automatic Content Extraction

AI . . . . . Artificial Intelligence

BLANC . . . . Bilateral Assessment of Noun-Phrase Coreference

BN . . . . . BabelNet

CNN . . . . . Convolutional Neural Network

CR . . . . . Co-reference Resolution

CS-RE . . . . Cross-Sentence Relation Extraction

DL . . . . . Deep Learning

DS . . . . . Distant Supervision

EE . . . . . Event Extraction

EEC . . . . . Extracted-Event-Candidate Set

ICC . . . . . Intraclass Correlation

IE . . . . . Information Extraction

FO filter . . . . Frequency-Overlap filter (for patterns)

KB . . . . . Knowledge Base

KBP . . . . . Knowledge Base Population

LTW . . . . . Los Angeles Times/Washington Post part of the Gigaword 5 corpus

MUC . . . . . Message Understanding Conference

ML . . . . . Machine Learning

NER . . . . . Named-Entity Recognition

NLP . . . . . Natural Language Processing

NN . . . . . Neural Network

NYT . . . . . New York Times part of the Gigaword 5 corpus

POS . . . . . Part(s) of Speech

RE . . . . . Relation Extraction

Sar-graphs . . Graphs of Semantically Associated Relations

S filter . . . . . Semantic filter (for patterns)

TAC . . . . . Text Analysis Conference

WN . . . . . WordNet

WSD . . . . . Word-Sense Disambiguation

Part-of-Speech Tags and Dependency-Relation Labels

This section lists abbreviations of part-of-speech tags and dependency-relation labels. We only present here abbreviations actually used in figures of this thesis.

Universal Part-of-Speech Tags

See Petrov et al. (2012), `http://universaldependencies.org/` (last access: 2017-04-26).

ADJ . . . . . adjective

ADP . . . . . adposition (prepositions and postpositions)

ADV . . . . . adverb

DET . . . . . determiner

NOUN . . . . . noun (common and proper)

NUM . . . . . cardinal number

P(A)RT . . . . particle

PRON . . . . . pronoun

VERB . . . . . verb (all tenses and modes)

. . . . . . punctuation

Penn Treebank (Part-of-Speech) Tags

See M. P. Marcus et al. (1993) and Taylor et al. (2003).

CC . . . . . coordinating conjunction

IN . . . . . preposition or subordinating conjunction

NN . . . . . noun, singular or mass

NNP . . . . . proper noun, singular

VBD . . . . . verb, past tense

VBN . . . . . verb, past participle

Universal Dependency Relations

See Nivre et al. (2016) and `http://universaldependencies.org/` (last access: 2017-04-26).

acl . . . . . clausal modifier of noun

advmod . . . . adverbial modifier

amod . . . . . adjectival modifier

aux . . . . . auxiliary

case . . . . . case marking

compound . . . compound noun

det . . . . . determiner

dobj . . . . . direct object

mark . . . . . marker

nmod . . . . . nominal modifier

nsubj . . . . . nominal subject

nummod . . . numeric modifier

punct . . . . . punctuation

Stanford Dependency Relations

See de Marneffe and Manning (2008).

amod . . . . . adjectival modifier

appos . . . . . appositional modifier

auxpass . . . . passive auxiliary

cc . . . . . coordination

conj . . . . . conjunct

conj_and . . . conjunct via “and”

cop . . . . . copula

dep . . . . . dependent

det . . . . . determiner

dobj . . . . . direct object

neg . . . . . negation modifier

nn . . . . . noun compound modifier

nsubj . . . . . nominal subject

nsubjpass . . . passive nominal subject

pobj . . . . . object of a preposition

poss . . . . . possession modifier

possessive . . . possessive modifier

prep . . . . . prepositional modifier

prep_from . . prepositional modifier via “from”

prep_in . . . . prepositional modifier via “in”

prep_of . . . . prepositional modifier via “of”

prep_since . . prepositional modifier via “since”

punct . . . . . punctuation

rcmod . . . . . relative clause modifier

tmod . . . . . temporal modifier

List of Algorithms

List of Figures

List of Tables

Chapter 1

Introduction

Contents

1.1 Motivation

Modern-day IT systems are faced with an ever-growing flood of data. This data mass occurs in a manifold of different shapes and sizes and creates high demands on technology aiming to process it. The term *big data* has been coined to describe both the emergence of the information overload (Snijders et al., 2012) and the technology needed to cope with it (Hashem et al., 2015). It is widely accepted in industry and academia that mastering big data is a crucial step towards the future of economy (Labrinidis and Jagadish, 2012). One of the main obstacles in handling today's masses of data is its great variety. Most data occurs in semi-structured and unstructured forms (Gandomi and Haider, 2015), with textual data being a major contributor, besides audio and video formats. In particular, textual documents on the world wide web are a huge source of knowledge, as they contain unstructured information in the shape of natural language. Harvesting this information and reducing it to (database) tuples is an integral part of big data processing (Agrawal et al., 2012; Hu et al., 2014). Another major trend of the past years has been the *semantic*

web, an intermediate step to a machine-readable internet. Berners-Lee (1999, p. 177) defines it as "a web of data that can be processed directly and indirectly by machines." Today, most information available on the web is still encoded in unstructured textual forms and there are no indications that this will change in the (near) future. As internet activist Swartz (2013, p. 3) puts it, "it's hard enough to getting people to share data as it is, harder to get them to share it in a particular format, and completely impossible to get them to store it and manage it in a completely new system." This, too, establishes a need for reliable and high-performance text mining.

The research area of computational linguistics and its branch *natural-language processing* (NLP) address this need by developing methods for the automatic processing of human language, i.e., language technology. Such technology has been very present in the public perception throughout the last decade, with rapid progress and the development of all kinds of business, enterprise, and intelligence applications, for example, in areas such as social-media monitoring, knowledge organization, content-based advertising, and algorithmic trading. Furthermore, this area has seen the release of many consumer-facing products with impact from the NLP tasks machine translation, text summarization, question answering, and conversational agents (Jurafsky and Martin, 2009, chapters 23–25). Famous examples include digital personal assistants with spoken- and written-language interfaces, like Apple's Siri[1], Microsoft's Cortana[2] (MSR, 2014; Sarikaya et al., 2016), Google Now[3], Google Assistant[4], as well as Amazon's Echo/Alexa[5], and Facebook's M (D. Marcus, 2015) and accompanying bot engine (D. Marcus, 2016; Rosenberg, 2016; Lebrun and Team Wit, 2016). Another prominent example is IBM's Watson[6] (Ferrucci et al., 2010), a question-answering system capable of beating the best human contestants in the quiz show "Jeopardy!".

Astonishing advancements in NLP could be recorded in recent years. These achievements would not have been possible without the extensive use of machine-learning (ML) techniques in combination with processing increasing amounts of textual data. ML technology is widely used today in both academia and industry, in particular, the recent revival of neural networks (NNs) (LeCun et al., 2015) has had a great impact. Before this revival, it was widely believed that NNs were unsuitable for a wide range of problems, despite a few successful neural architectures (e.g., LeCun et al., 1998). This has changed

[1] `http://www.apple.com/ios/siri/`, last access: 2016-05-02.
[2] `https://www.windowsphone.com/en-us/how-to/wp8/cortana/meet-cortana`, last access: 2016-05-02.
[3] `https://www.google.com/landing/now/`, last access: 2016-05-02.
[4] `https://assistant.google.com/`, last access: 2017-03-16.
[5] `http://www.amazon.com/gp/product/B00X4WHP5E`, last access: 2016-05-02.
[6] `http://www.ibm.com/smarterplanet/us/en/ibmwatson/`, last access: 2016-05-02.

due to the growing availability of data, the increased computing capabilities, as well as algorithmic advances (Goodfellow et al., 2016, pp. 18–21), which have turned NNs into the tool of choice for a manifold of use cases in NLP (Manning, 2015). Amongst the most popular strands of neural methods for language are the utilization of so-called *distributed representations* of linguistic items (words, etc.) and the composition of very *deep* (↪ deep learning, DL) model architectures for high-level abstraction of input data (e.g., in speech generation, Oord et al., 2016; in machine translation, Y. Wu et al., 2016).[7] Another perspective on the interplay of big data, NLP, and ML is provided by leading artificial-intelligence (AI) experts who believe that a more generic AI can be achieved by enabling machines to learn by communicating with their environment (Mikolov et al., 2015). This requires the automatic processing of vast data amounts, particularly language. Without well-functioning language understanding components, such machines would not be able to properly interact with human instructors and hence would not get direct feedback on their decisions and actions.

The impact of data size on the prediction performance has long been a subject of investigation in NLP (Banko and Brill, 2001a). Many studies have found that collecting large text corpora from the web can boost the performance for simple and more complex language problems (Keller et al., 2002; Sasano et al., 2009). Some studies also show a roughly log-linear relationship between performance improvements and an increase in training data size (Banko and Brill, 2001b; Brants et al., 2007). It has, however, also been acknowledged that simply adding more data without new language processing methodologies is only of limited help (Curran and Osborne, 2002; Lapata and Keller, 2004). Another noteworthy trend is the employment of cheap crowd-sourced annotation of texts to produce large amounts of labeled textual data, albeit at a lower quality than would be expected of linguistics experts (Snow et al., 2004). In this thesis, we employ classic ML as well as DL techniques to process large repositories of text, which allows us to advance the state-of-the-art in NLP.

The NLP research area of *information extraction* (IE) develops approaches which distill structured information from textual data. The generated output can be instances of concepts (persons, locations, organizations, etc.) or relations among these. Relations can appear in different manners, i.e., one-time events, general facts, or even opinions. For example, biographic information about people may include the marriage relation between two persons, or kinship relationships. IE systems are often implemented as processing pipelines (Cardie, 1997; Piskorski and Yangarber, 2013), where the individual components analyze language on different granularity levels and build upon the results of previous

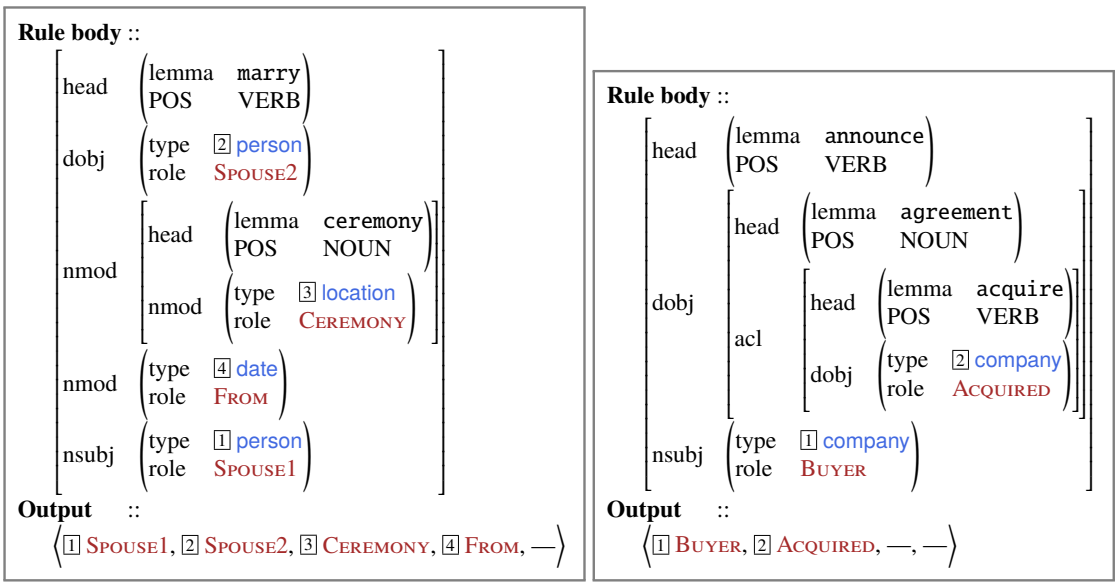
[7] Section 2.6 presents an introduction to neural architectures for NLP.

pipeline modules. Frequently used elements in such pipelines focus on the segmentation of words and sentences, on grammatical analysis, on the recognition of concepts, on the extraction of relations (relation extraction; RE) between concept mentions, and, finally, on the identification of relations that refer to the same real-world instances. The latter two steps are of utmost importance for applications and thus receive particular consideration in this thesis. To this end, novel methods for the detection of relations both within and across sentences (cross-sentence RE) are presented. Furthermore, we lay our attention on the intersection of IE methods and the semantic web with its repositories of linked data. Specifically, we report how factual knowledge from databases can be linked to linguistic knowledge, i.e., language expressions.

1.2 Research Problems

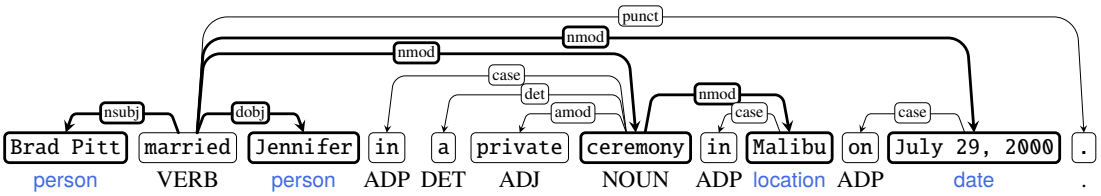
In order to motivate the specific contributions of the present thesis, this section introduces relevant fundamental research problems of IE. These problems are concerned with the utilization of so-called *linguistic patterns* for the detection of relational information in texts. Patterns, also known as *rules*, are templates for sentences, which are used to determine which types of relations are mentioned in texts and where relational arguments are located in sentences. In literature, RE systems often exploit extraction rules with different underlying sentence models. Early works (Brin, 1998; Agichtein and Gravano, 2000; Etzioni et al., 2005; Pantel and Pennacchiotti, 2006) used *lexico-syntactic patterns* (Hearst, 1992), which are regular expressions containing surface-level strings of words. They can also be placeholders for named entities of certain types or wildcards for noun phrases. In order to handle relations with more semantic arguments as well as sentences with a more complex structure, current approaches additionally employ grammatical analysis, making the patterns more expressive. A popular approach is to utilize *dependency relations* between words (de Marneffe and Manning, 2008; Nivre et al., 2016). These binary relations allow a rule formalism to skip irrelevant parts of a sentence and to easily connect semantic arguments. Figure 1.1 (p. 5) depicts two examples of extraction rules/patterns, both based on this formalism.

Rules are graph-based templates for sentences: nodes specify properties of individual words/phrases and edges connect nodes based on adjacency (neighboring words) or grammatical functions. At extraction time, RE systems produce corresponding graph representations for input sentences, which are subsequently compared with the rules. Finding the part of a sentence which matches a rule is equivalent to solving a sub-graph isomorphism problem, the solution to which has the information on relation mentions

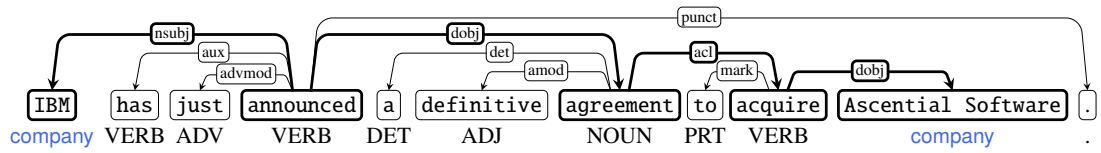


a) *marriage* relation between people. **b)** *acquisition* relation between companies.

Figure 1.1 – Example extraction rules for two semantic relations. Rules are presented in the form of attribute-value matrices matching parts of a syntactic analysis of a sentence. *POS* is short for *part of speech*, tags correspond to Petrov et al. (2012). Blue typesetting corresponds to abstract types of real-world concepts; brown color indicates output semantic roles in relations. Left-hand abbreviations like *dobj*, *acl*, etc. correspond to syntactic links between words, see the framework presented by Nivre et al. (2016). Long dashes in the output field of the rules indicate that some arguments of the respective semantic relation are not covered by the rules.



a) Sentence describing the wedding event of a celebrity couple.



b) Sentence indicating that one company is bought by another.

Figure 1.2 – Two example sentences with mentions of semantic relations. This figure also depicts the results of running automatic components for syntax analysis (above/below the sentence) and recognition of concept references (below the sentence). The grammatical framework used here is the same as in Figure 1.1. For the meaning of abbreviated tags and edge labels in this and other figures of this thesis, see the List of Abbreviations (p. ix). Highlighted parts match the respective extraction rule in Figure 1.1.

contained in the text. The rule in Figure 1.1a covers four semantic arguments of a marriage relation, i.e., two spouses, date, and location of the associated wedding event. Applied to a sentence as shown in Figure 1.2a (p. 5), this rule matches the syntactic structure of the sentence and locates the arguments of the mentioned semantic relation. Analogously, the rule in Figure 1.1b describes one particular way of expressing the acquisition relation between companies. This expression is used in the sentence of Figure 1.2b (p. 5) with the relation instance ⟨`IBM`, `Ascential Software`⟩. Approaches building their systems on dependency-relation analysis or related linguistic formalisms are presented, e.g., by Yangarber (2001), Stevenson and Greenwood (2005), Greenwood and Stevenson (2006), Suchanek et al. (2006b), and Adolphs et al. (2011).[8]

The main goal of this thesis is the development of novel strategies for text analysis and information extraction. This goal is approached from different angles, with linguistic patterns being a central building block. A number of questions naturally arise when following a pattern-based approach to IE, which are not sufficiently addressed by current methods:

1. How can we collect a large set of linguistic patterns and automatically assess their quality, relevance, and diversity?

2. What is a good way to represent the collected patterns so that the linguistic knowledge in them is available for applications other than immediate relation extraction?

3. What are the limitations of a pattern-focused extraction methodology and how can we overcome them?

We formulate three research problems based on these questions that reflect shortcomings of the current techniques in text analysis and information extraction.

Research Problem 1: Automatic Acquisition of Linguistic Patterns and Confidence Estimation In recent approaches, linguistic patterns are typically extracted from large text collections automatically. Inevitably, sets of acquired patterns contain noise, which is why accurate methods for confidence estimation are needed. RE approaches in literature commonly employ statistics about the distribution of linguistic patterns in texts, e.g., how often does the phrase "... *married* ... *in a festive wedding event last weekend*" co-occur with references to persons a-priori known to be spouses. Extraction systems rely on accurate confidence estimation of automatically discovered linguistic patterns.

[8] We review pattern-based RE approaches in Section 2.3.

Deficiencies in this step result in a combination of both low quality of extracted knowledge and, at the same time, little coverage of this extracted knowledge with respect to what is mentioned in texts.

RE performance of state-of-the-art systems is below the level that is required for production systems. For example, in the 2013 cycle of the RE shared task at the Text Analysis Conference (TAC),[9] the best performing system scored only 37.28 % F1[10] (Surdeanu, 2013). The winner of the follow-up task in 2014 only reached 36.77 % F1 (Surdeanu and Ji, 2014). This level of performance is merely half of what human experts achieve manually in similar experimental conditions (also Surdeanu, 2013; Surdeanu and Ji, 2014); results of other IE competitions (e.g., Sundheim, 1995; Hendrickx et al., 2010) confirm this discrepancy between human text understanding and the capacity of automatic analysis methods. To date, both accurate and high-coverage extraction of facts from textual sources is far from being reality, although it is desperately needed.[11]

Research Problem 2: Large-Scale Linking of Linguistic and World-Knowledge Resources Knowledge graphs are vast networks which store entities as well as their semantic types, properties, and relations. In recent years considerable effort has been made to construct these large knowledge bases (KBs) in academic research, community-driven projects and industrial development (Bollacker et al., 2008; Suchanek et al., 2007; 2008; Lehmann et al., 2015; Carlson et al., 2010b; T. M. Mitchell et al., 2015; Nakashole et al., 2011; Dong et al., 2014). Internet communities and industrial applications often employ this type of resource to support user-facing services which at least in part display relational information in response to user queries. Examples include the Google Knowledge Graph, which backs Google's search engine (Singhal, 2012), Satori, supporting Microsoft's Bing (Qian, 2013), and Wikidata (Vrandečić and Krötzsch, 2014), an attempt to provide a homogeneous data source for Wikipedia's infoboxes.

A parallel development is the emergence of several large-scale linguistic-knowledge resources with a focus on language (Melo and Weikum, 2009; Navigli and Ponzetto, 2012; Gurevych et al., 2012; Speer and Havasi, 2013). At their core, these resources are lexical databases with dictionary and thesaurus functionality. Most of them also capture richer semantic information about words, for example, their semantic and syntactic combinatorial properties. Both knowledge graphs and linguistic resources are commonly published as linked data. The concept of linked data can be considered as an instantiation

[9] More precisely, the Slot Filling task in the Knowledge Base Population track of TAC.

[10] F1 score is calculated as the harmonic mean of precision and recall. It is a measure to aggregate information about both coverage and accuracy aspects of an extraction system.

[11] Section 2.7 reports results of further shared tasks.

of the idea of the semantic web, which gives detailed practical instructions on how resources should be formatted in order to become a proper member of this web of data (Bizer et al., 2009a). Today, repositories of world-knowledge and linguistic knowledge complement each other in the so-called linked-data cloud. What is missing, however, is an explicit link between these two types of repositories, a link that builds a bridge from the semantic relations of knowledge graphs to their linguistic representations. In particular, a resource is needed which lists the different possibilities of how particular relations can be expressed in natural language. Such a linking of repositories would enable new types of applications and would increase the impact that individual knowledge resources already have, e.g., for tasks like IE.[12]

Research Problem 3: Overcoming the Sentential Barrier The third problem addressed in this thesis is the occurrence of native discourse-level information in RE, i.e., relational information which crosses the sentence boundary. The restriction of the RE task to the sentence level only was originally introduced to foster progress on RE, as this simpler variant of the task grasps at least part of the relational information mentioned in a document. However, more than 20 years have passed since the publication of the first RE approaches (e.g., Hearst, 1992), so it is time to tackle this issue.

Although cross-sentence RE has received some attention in literature, most systems nevertheless restrict themselves to processing only one sentence at a time, thereby imposing an upper limit on the coverage any RE system can reach. For example, Swampillai and Stevenson (2010) find that 28.5% of binary relation mentions in the RE dataset from the sixth iteration of the Message Understanding Conference are cross-sentential, as are 9.4% of relation mentions in the 2003 dataset of the IE competition called Automatic Content Extraction. More researchers reported similar numbers, e.g., Ji and Grishman (2011) estimate that 15% of slot fills in the training data for the RE task at TAC 2010 require cross-sentential inference. This clearly emphasizes the importance of methods for cross-sentential RE. Yet, it is largely unexplored how this problem can be effectively approached, and which new challenges (compared to traditional, intra-sentential RE) need to be faced.[13]

[12] In Section 2.5, we discuss the different types of resources in greater detail.
[13] See also Section 2.4.

1.3 Contributions of this Thesis

The following section summarizes the main contributions with which the three research problems stated above can be resolved. The first two contributions deal with linguistic-pattern discovery and pattern-based relation extraction (**Research Problem 1**). The language-resources aspect of the thesis is covered by the third contribution (**Research Problem 2**). This is complemented with research on discourse-level, i.e., cross-sentential, aspects of RE in the fourth item (**Research Problem 3**).

Contribution 1: Pattern Discovery in Schema-Based IE The starting point of this research is a RE system that follows a pattern-based methodology. This system learns grammar-based extraction rules from web documents by utilizing a large number of instances of semantic relations as seed knowledge. To deal with the large amount of noisy patterns that can potentially hamper extraction quality, the system also contains a pattern filter that is based on inter-relation constraints. We show that the precision of RE can be improved by employing wide-coverage, general-purpose lexical-semantic networks like BabelNet (Navigli and Ponzetto, 2012) for effective semantic rule filtering. This process results in higher precision at any given recall level, compared to existing baseline methods. Subsequent improvements are achieved by employing a combination of the initial filter and the semantic filter.

In addition, we design a new method for the discovery of patterns in sentences with relation mentions, compatible with the distantly supervised learning scheme. Employing knowledge from lexical-semantic repositories ensures that induced extraction rules cover all semantically relevant material, even if part of it is situated outside the shortest path which connects the relation arguments in a sentence. This new method significantly raises both recall and precision with roughly 20% F1 score boost in comparison to a previous approach, which does not consider lexical semantic information during pattern extraction. We supplement the above work with the description of an annotation effort leading to a new RE dataset, which constitutes an additional testbed for competing RE approaches, thereby increasing comparability of RE methods in general.

Contribution 2: Distributed Representations for Patterns We conclude the work on the first research problem with a new unsupervised method for the learning of distributed representations of linguistic patterns. This contribution is concerned with an *open* IE scenario in which all patterns observed in a large text corpus are at first collected independently from one another, while ignoring the aspect of semantic relations which

patterns can express. The patterns are then disambiguated in a downstream step in order to identify clusters of paraphrastic patterns, i.e., groups of expressions which correspond to the same real-world relations. Compared to traditional IE methods, this open approach has advantages when dealing with text domains with continuously evolving information, where fixed background ontologies of relations quickly become outdated and hence the relations to be extracted need to be discovered themselves in the texts.[14]

Our method for learning paraphrases of event patterns is based on a neural-network architecture. The training of this network is guided by a weak supervision signal coming from the conformity of patterns with respect to publication dates of newspaper articles and mentions of real-world entities. The network can generalize across extractions from different dates to produce a robust paraphrase model for event patterns, which also captures meaningful representations for rare patterns. The proposed model is evaluated on both small-scale and large-scale datasets and shows superior performance compared to a strong ML baseline system on all of them.

Contribution 3: A Linked Language Resource from Linguistic Patterns Here we address the second research problem, i.e., the problem of transforming a set of automatically constructed patterns into a widely useful linguistic resource, connected at a fine-grained level to other resources. To this end, we develop a new type of knowledge repository, called *graphs of semantically associated relations* (sar-graphs), which link semantic relations from factual knowledge graphs to the linguistic patterns with which a language expresses instances of these relations. Compared to lexical-semantic resources, sar-graphs model syntactic and semantic information at the level of relations. Hence, they are useful for tasks such as knowledge-base population and relation extraction. A language-independent method to construct sar-graph instances is presented, which is complemented by manual pattern verification work to enhance the quality of the resource. The graphs are linked at the lexical level to existing resources, most prominently to BabelNet (Navigli and Ponzetto, 2012). Furthermore, links are established on the pattern- and relation-level to a frame-based lexical-semantic repository called FrameNet (Baker et al., 1998). Finally, it is described how the graphs have been processed in order to publish them as part of the linked data cloud via the *lexicon model for ontologies* (Lemon; McCrae et al., 2011).

[14] See the introduction of Chapter 5 for more details on the application settings that benefit from (or even require) open-IE techniques.

Contribution 4: Cross-sentential Relation Extraction We approach the third research problem by conducting a study which identifies shortcomings of intra-sentential extraction approaches and the obstacles that have to be faced in order to broaden a standard IE system's scope to multiple sentences. We investigate which properties of current IE systems are particularly important for high-coverage extraction of document-level facts. The study is conducted on several publicly available gold-standard corpora from the IE area. In addition, a dataset has been specifically created for the purpose of this study. This dataset is dedicated to the detection of events and facts spread across sentence boundaries, it features annotation of entities and relations, of co-reference relations among these entities and events, as well as of terms that are semantically relevant for the relations and events.

We then select the sub-problem of event-mention linking as a means to overcome the single-sentence limitation by developing a new kind of information extraction approach. This method learns to model sentence semantics via convolutional neural networks, and in a second step uses these representations to identify co-referential event mentions in different parts of a document. Co-referential event mentions, along with their arguments, are merged, which facilitates true document-level information extraction. The approach is thoroughly tested against competing systems from literature, with the result that our model reaches state-of-the-art results albeit using less domain-dependent features than the other systems.

1.4 Research Context

The research presented in this thesis has been performed as part of the work of the Text Analytics group at the Berlin site of the German Research Center for Artificial Intelligence (DFKI). It took place in the context of the following research projects:

2013–2014: *Deependance*, funded by the German Federal Ministry of Education and Research (BMBF; contract 01IW11003)
2013–2014: *Intellektix*, a sub-project of the *Software Campus* initiative financed by BMBF to support PhD-level researchers (contract 01IS12050)
2013–2015: *LUcKY*, supported by Google through a Focused Research Award
2014–2017: *BBDC, Berlin Big Data Center*, funded by BMBF (contract 01IS14013A)
2015–2017: *All-Sides*, funded by BMBF (contract 01IW14002)

All of these projects include the research areas of information extraction and semantic-web technologies in their research programs.

1.5 Thesis Overview

- This thesis continues with an introduction of the field of text analytics and information extraction in Chapter 2. Various common sub-tasks of IE are introduced and their interplay is explained. We also take a brief look at prominent approaches from literature, and explain the terminology used in subsequent chapters.
- Chapter 3 describes the distantly-supervised system for RE-pattern discovery, which operates directly on web documents; this chapter also discusses the pattern filter based on inter-relation constraints. Chapter 4 then reports how a lexical-semantics resource is used to increase both precision and recall of the patterns.
- Chapter 5 describes the unsupervised approach for the construction of distributed representations for linguistic patterns. This chapter focuses on the open-IE paradigm, which is important in particular for emerging relations.
- Chapter 6 elaborates on the newly constructed linguistic resource 'sar-graphs', which is built from sets of automatically gathered linguistic patterns. This chapter comments on the various processing steps that are applied in order to integrate the individual patterns into a larger graph, e.g., the manual curation of linguistic constructions. Further discussed aspects are the native representation format of the sar-graphs, as well as comparisons with related resources.
- Chapter 7 focuses on the problem of cross-sentential relation extraction. The first part of this chapter presents the analysis of IE systems and manually annotated RE datasets, as well as a description of the annotation process of the newly created dataset. Afterwards, a new method that is capable of bridging the sentence gap is presented.
- Chapter 8 concludes this thesis and describes potential directions for future work.

1.6 Prior Work

Parts of the research presented in this thesis have been already published. The work on schema-based IE from Chapters 3 & 4 has been reported in three full conference papers (Krause et al., 2012; Moro, H. Li, and Krause, et al., 2013; H. Li and Krause et al., 2015) and a journal article (Krause et al., 2016a). A predecessor of the system portrayed in Chapter 3 has been discussed in a *Diplom* thesis (Krause, 2012). The dataset whose

construction is briefly described in Chapter 4 has been initially presented in a full paper at a language-resources conference (H. Li and Krause et al., 2014). The unsupervised approach for pattern-representation learning in Chapter 5 has appeared as another full conference paper (Krause et al., 2015a). The work on sar-graphs (Chapter 6) has been covered in the aforementioned journal article (Krause et al., 2016a), one full conference paper (Gabryszak and Krause et al., 2016), and one workshop paper (Krause et al., 2015b). The dataset created for the analysis of cross-sentential RE phenomena (Chapter 7) has been documented in another conference paper (Krause et al., 2014); same as the proposed approach for co-reference resolution of event mentions in the same chapter (Krause et al., 2016b). The work reported in Chapters 3–7 was conducted jointly with the authors of the publications listed above.

Chapter 2

Knowledge Acquisition from Texts

Contents

2.1 Introduction

The research field of natural-language processing and text analytics has a long history, going back as far as to the middle of the 20th century (Jurafsky and Martin, 2009, pp. 43–47). The field has seen the emergence of many different tasks concerned with various aspects of language and their numerous applications in current technologies. In this chapter, we review approaches and tasks which are relevant to the central issues of this thesis. A broader overview of the developments in NLP in the past decades is provided by the works of Allen (1987), Manning and Schütze (1999), and Jurafsky and Martin (2009).

This thesis presents methods for the problem of information extraction (IE), which involves the acquisition of structured information from texts. The first part of this chapter (Section 2.2) is dedicated to the introduction of IE, its sub-tasks, and related problems. A common practice in IE approaches in literature is the utilization of a pipeline architecture. By adding further annotation layers with increasingly abstract linguistic information, the components of such a pipeline subsequently build up more complex representations of an input text. At the end of this pipeline, processors produce as output factual statements (so-called *relations* and *events*) that were originally hidden in text. Section 2.3 discusses various approaches to this final pipeline step.

Many of the systems in literature which deal with the recognition of factual knowledge handle texts only at the level of individual sentences. In Section 2.4, we broaden this scope and explain methods that take into account a wider discourse context, either for improved sentence-level extraction performance, or for solving discourse-level tasks. We then move on to a different type of information source that is available to modern information systems, i.e., structured knowledge resources, in Section 2.5. We attempt to describe only a part of the vast language and world-knowledge resource landscape already in existence and will particularly focus on the resources that play a role for the methods described in the remainder of this thesis.

The chapter continues with an elaboration of the recent revival of neural methods for NLP in Section 2.6 and with a section on evaluation principles for IE tasks (Section 2.7). The final section (2.8) of this chapter summarizes the state-of-the-art in IE and connects it to the research problems outlined in the previous chapter.

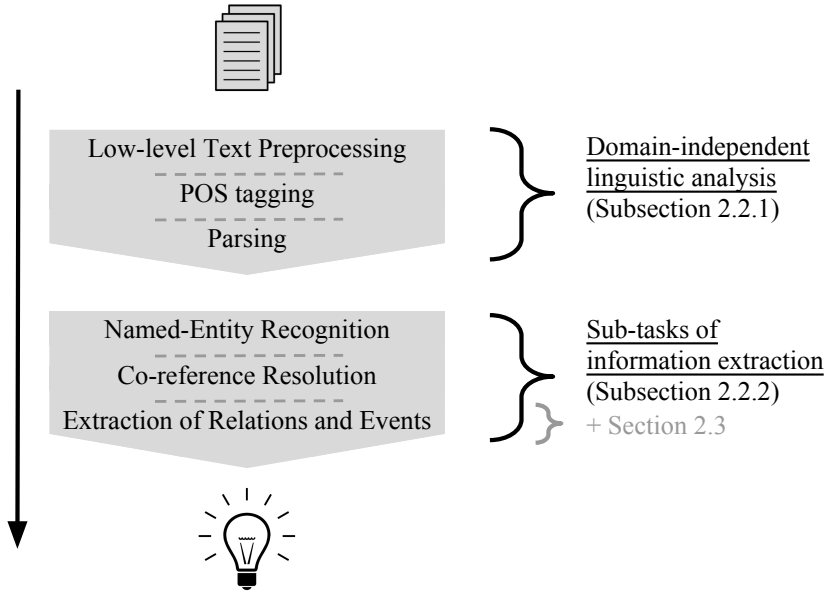


Figure 2.1 – A prototypical IE pipeline. The listed processing steps are discussed in Section 2.2.

2.2 A Prototypical Pipeline for Information Extraction

The Message Understanding Conference defines IE as "the extraction or pulling out of pertinent information from large volumes of texts."[15] In other words, the goal of IE is to "turn the unstructured information embedded in texts into structured data" (Jurafsky and Martin, 2009, p. 759). The type of information regarded as relevant and pertinent can either be defined a-priori by users, in which case a deep analysis of specific semantic aspects of a text is conducted and a closed-schema database is populated. It can also be open-ended (*open IE*), in which case a comprehensive extraction of knowledge, including emerging information types, is possible (Nastase et al., 2013, p. 24). The next two subsections discuss common linguistic preprocessing techniques (Subsection 2.2.1) and review the main components of standard IE pipelines (Subsection 2.2.2). See also Figure 2.1, which presents an overview of the topics in this section.

The great diversity of published IE research renders it impossible to give a complete overview of the field, hence this section highlights only a few research directions which are particularly relevant for the work presented here. Alternative introductions to and overviews of various aspects of IE were presented by Grishman (1997), Muslea (1999), Siefkes and Siniakov (2005), Chang et al. (2006), Cunningham (2006), Turmo et al.

[15] See `http://www-nlpir.nist.gov/related_projects/muc/` ↪ "Information Extraction" ↪ "Information Extraction Definitions". Last access: 2016-06-24.

(2006), Moens (2006), Bach and Badaskar (2007), Sarawagi (2008), D. Zhou and He (2008), Weikum and Theobald (2010), Grishman (2012), Nastase et al. (2013), C. Li et al. (2013), Piskorski and Yangarber (2013), Hirschberg and Manning (2015), and Nickel et al. (2016).

2.2.1 Domain-Independent Linguistic Analysis

A prerequisite step to the collection of factual information from text is the execution of several linguistic preprocessors which make raw text accessible to downstream analysis components. While the specific selection of preprocessors varies greatly in literature, the ones described in the following are relatively common.

Low-level Text Preprocessing The initial steps of text processing happen on a low level (Manning and Schütze, 1999, pp. 124–136) and aim to identify very basic blocks of meaning in (written) human language. The process of *tokenization* divides an input text into a series of so-called tokens, mostly words, but also other items like punctuation characters and numbers. At least for the English language, tokenization is generally considered to be a solved problem, for which rule-based methods provide acceptable performance (see Dridan and Oepen, 2012), albeit more sophisticated methods were proposed as well (e.g., Evang et al., 2013).

After tokenization, words are mapped from their inflected forms (e.g., `married`)[16] to a less sparse representation by either removing affixes (*stemming*, see the famous Porter Stemming Algorithm, Porter, 1980; `married` → `marri`) or identifying their corresponding semantic base forms in a dictionary (e.g., Chrupała, 2006; Gesmundo and Samardzic, 2012), so-called lemmas (`married` → `(to) marry`). What follows is the segmentation of a text into paragraphs and sentences. While there are pitfalls in this seemingly simple task—most prominently that not all periods in a text mark the end of a sentence—heuristic boundary detection algorithms achieve good results already (Manning and Schütze, 1999, p. 135). Read et al. (2012) recently presented a survey of various available segmentation toolkits and found that most of them give very good performance out-of-the-box on a selection of standard datasets from NLP. Yet, the authors warn that noisy web text still poses a challenge.

POS Tagging After a text document has been segmented, the next processing step deals with the tagging of words with syntactic categories (*part-of-speech (POS) tagging*). Different inventories of grammatical tags exist that define the categories into which the

[16] Text examples are displayed in `typewriter font` throughout this thesis.

Tag	Category name	Example (English)
VERB	verbs (all tenses and modes)	`assigned`
NOUN	nouns (common and proper)	`machine`
PRON	pronouns	`he`
ADJ	adjectives	`clean`
ADV	adverbs	`quickly`
ADP	adpositions (prepositions and postpositions)	`whether`
CONJ	conjunctions	`and`
DET	determiners	`the`
NUM	cardinal numbers	`371`
PRT	particles or other function words	`'s`
X	other: foreign words, typos, abbreviations	`oops`, `jour`
.	punctuation	`!`, `?`, `.`

Table 2.1 – Universal part-of-speech tags (Petrov et al., 2012).

words of a text are grouped. The Penn Treebank tag set (M. P. Marcus et al., 1993; Taylor et al., 2003) emerged as a standard scheme in the 1990s because of the widespread use of the corresponding dataset in the community. This tag set features 45 different categories; even more detailed schemes exist (Manning and Schütze, 1999, pp. 139–145). More recently, Petrov et al. (2012) proposed a "universal" POS tagset of twelve tags that aims at capturing the most common word categories across 22 languages; Table 2.1 lists these categories.

There are two main problems with which automatic methods for POS tagging have to deal. The first one is the existence of *open* classes like nouns and verbs, which have a large set of members and which change over time. The second problem are ambiguous words that can have multiple functions. For example, `well` can be a noun, verb, adjective, and adverb. Disambiguating a particular use of a word with respect to its grammatical class requires looking at the context of this word (the remaining sentence). Most published approaches can be categorized into one of two classes: rule-based or stochastic (Jurafsky and Martin, 2009, pp. 169–184). An example for the stochastic class is Brants (2000), who presented a Hidden Markov model for sequence classification of the tags in a sentence.

Parsing A common step that follows POS tagging is the analysis of the grammatical structure of sentences (their syntax). The syntax plays a major role in determining the meaning of a sentence and addresses the construction of sentences from words. Grammars are formalisms that control this composition process. The procedure of matching a particular grammar against a given sentence is called *parsing*, the result of which is a

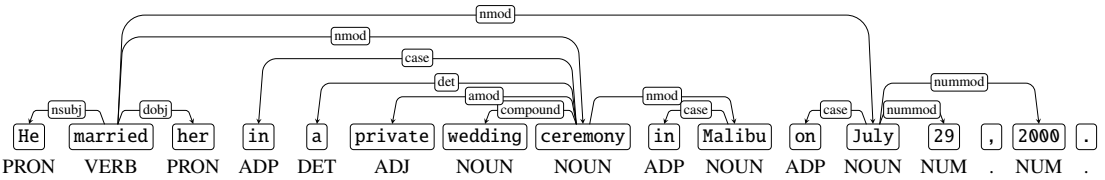


Figure 2.2 – Example sentence with POS tags (below the sentence) and dependency analysis (above the sentence) in the frameworks of Petrov et al. (2012) and Nivre et al. (2016).

parse.[17] Several theoretical frameworks for grammars exist. One such framework that is of particular importance for this thesis builds on the concept of *dependencies* (Manning and Schütze, 1999, pp. 101–106); grammars in this framework are called dependency grammars (Jurafsky and Martin, 2009, pp. 448–451).[18] Dependencies are relations between pairs of words in a sentence. Examples include the relation between a verb and its subject argument, and the relation between two nouns in a noun compound.

An example formalism in the space of dependency grammars is the "Stanford typed dependencies representation" (de Marneffe et al., 2006; de Marneffe and Manning, 2008). This formalism features approximately 50 different types of relations between words, for example, one for nouns that are combined via a conjunction, or ones connecting adverbs and adjectives to the words they modify. Although dependency parsing is not considered a solved problem, current approaches are able to provide an accurate analysis for input sentences from many sources. The book by Kübler et al. (2009) gives a detailed introduction to various methods for producing dependency analyses. Petrov and McDonald (2012) summarized the results of a shared task on parsing of potentially noisy web texts. They found that, as expected, parsing web text is much harder than analyzing relatively well-formed newswire texts, resulting in parsing accuracy of only approximately 80%. Based on the universal POS tags mentioned above and in a similar line of thought, very recently a framework was presented that unifies the annotation of dependency analyses of sentences for many languages (Nivre et al., 2016). Figure 2.2 depicts an analysis of a sentence with these universal dependencies. This example is typical in that most connected words are close to one another, with the exception of a few long-distance links. For instance, the verb `married` is connected via a nominal-modifier edge to the noun `July`, which specifies a temporal attribute of the verb.

[17] The ambiguity of language can result in more than one plausible analysis of a sentence, in which case several parses may be produced.

[18] Another framework relies on the linguistic notion of word groups that form units called phrases or constituents. These groups can be nested and their layout in a sentence is analyzed with the help of probabilistic context-free grammars (Manning and Schütze, 1999, Chapter 11).

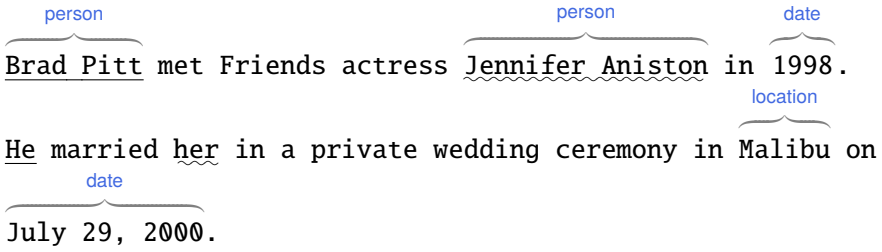


Figure 2.3 – Example sentence with highlighted entity mentions, containing a mention of a *marriage* relation between two persons. Blue colored text represents entity classes. Words underlined in the same style refer to the same persons.

2.2.2 Information Extraction and Sub-tasks

The problem of IE is typically seen as consisting of a small number of sub-tasks (Jurafsky and Martin, 2009, pp. 759–760): named-entity recognition (NER), (co-)reference resolution (CR), and the extraction of relations (RE) and events (EE). Since the early days of IE research, systems have been implemented as pipelines, which handle low-level pre-processing of texts and each of the sub-tasks in separate components (Cardie, 1997). A more recent survey by Piskorski and Yangarber (2013) confirms that pipeline architectures are still wide-spread; albeit there are instances of joint approaches in literature (Singh et al., 2013; Miwa and Bansal, 2016). These are relatively rare, but have the advantage of reduced error propagation between pipeline components, since they share information more freely between sub-tasks.

Named-Entity Recognition The task of NER is to recognize and classify strings in texts which refer to named entities (Jurafsky and Martin, 2009, p. 761). Here, an entity is a named object or concept, belonging to a certain semantic class, which are called named-entity types.[19] The granularity of these types is largely application-dependent. Many systems use a scheme with coarse-grained, general-domain classes, such as person, location, organization, and date. Some works refine these coarse types with a second more fine-grained layer (e.g., organization ↦ governmental org., commercial org., educational org.) (Linguistic Data Consortium, 2005a; Grouin et al., 2011) or add even more depth (Sekine et al., 2002; Sekine and Nobata, 2004).

Figure 2.3 contains an example text snippet. Here, named entities of types person, date, and location are highlighted with horizontal braces. When referring to a concrete text snippet representing a named entity in a document, the terms (entity) *mention* or

[19] Named-entity types are typeset sans-serif throughout this thesis.

(entity) *occurrence* are often used. NER is at its core a sequence-labeling problem, where for each token a decision has to be made on whether it constitutes a reference to an entity, and if so, to which type this entity belongs. Many resources with lists of (real-world) entities exist (see the survey by Ehrmann et al., 2016) and depending on the particular application domain, rule- and gazetteer-based approaches based on dictionary look-ups (e.g., in the framework of Drozdzynski et al., 2004) may already provide acceptable performance. More generic approaches are based on statistical sequence labeling techniques; Tjong Kim Sang and De Meulder (2003) and Nadeau and Sekine (2007) provided an overview of methods used in the field.

Co-reference Resolution This task is concerned with identifying and resolving noun phrases or named-entity mentions that refer to the *same* entities (Jurafsky and Martin, 2009, p. 730). For example, a text commonly uses a variety of expressions to refer to a person of interest, ranging from name variants (e.g., full name, first name, surname, abbreviations, etc.) to pronouns or descriptive noun phrases (e.g., `the president`). In Figure 2.3 (p. 20), words and phrases that refer to the same entities are underlined in the same way. In this example, the proper name `Brad Pitt` and the possessive form `his` refer to the real-world person *Brad Pitt*, `Jennifer Aniston` and `her` are references to another person.

Classic approaches to CR include the algorithms of Hobbs (1978) and Brennan et al. (1987), both essentially implementing algorithmic search procedures which start with an entity reference and go through the prior discourse (the previous sentences) in order to find the correct *antecedent*. This search is guided and restricted by various factors, such as the sentences' syntactic parses, gender, person, and number agreement. Similar constraints can also be encoded as features to a binary classifier, which is presented with pairs of entity occurrences from a document, and which decides for each such pair whether a co-reference relation holds between them (e.g., V. Ng and Cardie, 2002; Wiseman et al., 2015). V. Ng (2010) surveyed the different methodologies of CR work of the past decades. A series of shared tasks on CR compared and evaluated approaches (Recasens et al., 2010, at SemEval 2010; Pradhan et al., 2011, at CoNLL 2011; Pradhan et al., 2012, at CoNLL 2012). One noteworthy outcome of these competitions is the insight that rule-based methods based on hand-written heuristics can perform surprisingly well, an example is the system described by H. Lee et al. (2011), Raghunathan et al. (2010), and H. Lee et al. (2013).

A related task to CR is the problem of *disambiguating* or *linking* entity mentions against entries in a database. While this problem formulation comes with the availability

of rich entity information (e.g., extensive lists of name variants for entities in a database) and hence allows for very different approaches, the basic problem that a classifier/an algorithm has to solve remains the same, i.e., whether two given instances of an entity refer to the same real-world concept or not. The shared tasks on entity linking in context of the Text Analysis Conference's Knowledge Base Population track can give a good idea of current techniques (see the summaries by Ji et al., 2014; 2015), examples of recent approaches include Moro et al. (2014), who exploited commonalities of entity linking with word-sense disambiguation, and Y. Sun et al. (2015), who jointly embedded entity mentions from texts and entity representations from a database into a vector space and then disambiguated via vector similarity.

Extraction of Relations and Events Given the entities identified in texts, the goal of RE is to recognize and classify mentioned semantic relations between entities (Jurafsky and Martin, 2009, p. 760). For the text sample in Figure 2.3 (p. 20), one mentioned relation is the *marriage*[20] between `Brad Pitt`, `Jennifer Aniston`, `Malibu` and `July 29, 2000`. In other words, the goal of RE is to extract tuples of named entities from texts if there is textual evidence that these entities are arguments of a semantic relation. In traditional IE, the semantic relations which are to be recognized are defined in advance (Piskorski and Yangarber, 2013) and can, in general, have an arbitrary number of arguments (four in the example of Figure 2.3), although some literature exclusively works with binary relation definitions (e.g., Linguistic Data Consortium, 2005c).

The task of EE differs from RE mainly in that it focuses on finding mentions of real-world events (e.g., a terrorist attack) and their participating entities in contrast to recognizing references to static relations between entities. However, this distinction is blurry as events often initiate or end relations (e.g., a *wedding* starts a *marriage* relation). In this work, we treat the problems of RE and EE as closely related and do not view them as strictly separate tasks, meaning unless explicitly noted, statements about RE include EE and vice versa. A recent overview of relation and event definitions in literature and public datasets for RE/EE was presented by Aguilar et al. (2014).

2.3 Approaches to the Extraction of Relations and Events

This section surveys various approaches from literature for the extraction of semantic relations and events. We focus on two dimensions along which the approaches can be categorized: (a) how systems are trained, and (b) how relation mentions and sentences are

[20] The names of semantic relations are usually typeset in italics in the remainder of this thesis.

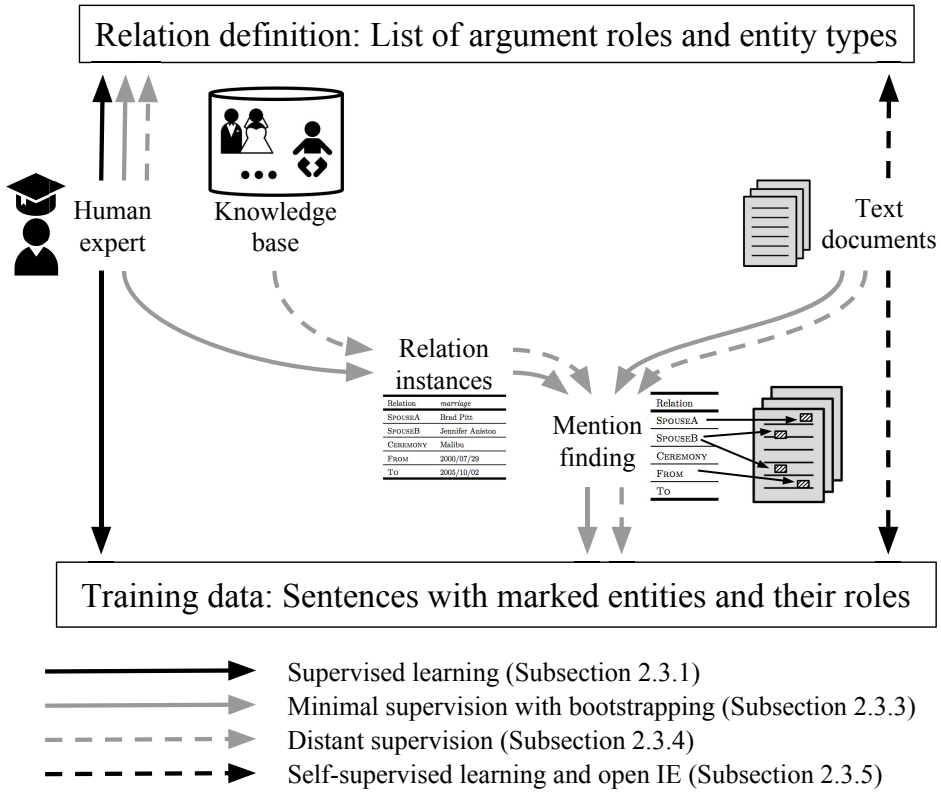


Figure 2.4 – Common learning paradigms for relation extraction. The picture outlines how training data is created (arrows ending at bottom rectangle) and how relation definitions are determined (arrows ending at top rectangle) in the paradigms. Each of the four types of arrows corresponds to one paradigm.

represented in the systems. In the first subsection (2.3.1), we describe the most common training setting and elaborate on different representations for relation mentions. This is continued in Subsection 2.3.2, which discusses a few principle ways of representing the structure of sentences in an IE system. The remaining subsections (2.3.3–2.3.5) comment on alternative training schemes which do not require as much hand-labeled data as the first training setting. Figure 2.4 compares the different training schemes; we describe the figure in the next subsections.

2.3.1 Supervised Learning

In a supervised setting, a system learns from labeled training examples how mentions of relations and events are expressed in a text. An example is typically a single linguistically-preprocessed sentence with two or more marked entity occurrences. The label of this example denotes the relation or event type that connects the entities, as well as the argument roles that the entities have in this relation or event. In Figure 2.4, the standard

supervised approach is depicted using the regular black arrows on the left: the relations for which information is to be extracted are manually defined, and training data is produced by human annotators.

Two main ways of representing mentions and sentences were proposed in literature: one uses vectors of so-called features, and the other one learns explicit sentence templates. The approaches presented in this thesis are mostly instances of the second category.[21]

Feature-based Methods IE systems in this category view an input sentence and the contained entity mentions as a feature vector, which, in classic approaches, is a set of hand-crafted properties of the sentence. G. Zhou et al. (2005) employed a somewhat prototypical set of RE features, where, besides properties of the named entities in the sentence (how they appear in text, their abstract types, etc.), most features capture the words in the sentence (their surface form, their stems and lemmas, their number, etc.) and the sentence's grammatical structure. Giving a comprehensive overview of the wide range of feature types employed in literature is beyond the scope of this chapter. For more details on various kinds of feature categories as well as examples, we refer to other works (Nastase et al., 2013, pp. 37–43; Jurafsky and Martin, 2009, pp. 770-772).

Feature-based IE methods build on many different machine-learning classifiers, which learn from training data and allow to extract information from unseen application-time texts. Prominent examples of classifiers include support-vector machines for RE (Zelenko et al., 2003; Bunescu and Mooney, 2005; G. Zhou et al., 2005; Grishman et al., 2005), logistic-regression/maximum-entropy models for RE (Chieu and H. T. Ng, 2002; Kambhatla, 2004; A. Sun et al., 2011) and EE (Grishman et al., 2005; Ahn, 2006), as well as graphical models like hidden Markov models and conditional random fields for RE (Ray and Craven, 2001; Skounakis et al., 2003; Rosario and Hearst, 2004) and EE (W. Lu and Roth, 2012). More approaches that build on statistical models for the extraction of relations and events are described in the surveys of Moens (2006), Siefkes and Siniakov (2005), and Turmo et al. (2006).

Traditionally, features that are defined intellectually and a-priori. A recent line of research aims to automatically learn which aspects of a text indicate the presence of relations and events. The features in these models are latent and represent short pieces of text in a continuous vector space, on top of which established classifiers are applied. For the construction of these representations, different neural-network architectures are used. We provide an overview of this research in Section 2.6.

[21] More precisely, this is true for Chapters 3–6, while Chapter 7 belongs to the first class.

Pattern-based Methods In contrast to the representation of sentences as feature vectors, some systems utilize explicit sentence templates called patterns or rules, which are learned from labeled examples in the same way that classifiers are trained. Patterns are often based either on surface-level representations of sentences or on dependency-grammar analyses. In the next subsection (2.3.2), we present examples for these two prominent formalisms and explain in detail how patterns are created from sentences.

In pattern-based methods, the extraction decision at inference time is made via explicit pattern matching (see for example F. Xu et al., 2007; Kim et al., 2011) and features like the types of entities etc. are explicitly encoded in the patterns. Whether a pattern-based methodology is superior or inferior to the use of feature-vector representations of sentences is a matter of active debate since both ways have pros and cons. The use of explicit patterns, for example, has the benefit of allowing humans to hand-write rules and inspect the system state more easily, which makes them the method of choice for industrial applications. Statistical extraction approaches, on the other hand, have currently more coverage in academic publications (Chiticariu et al., 2013).

2.3.2 Sentence and Pattern Representations

Existing extraction systems proposed in literature exploit various levels of syntactic analysis to generate syntax-based features or explicit extraction patterns. This subsection presents a brief survey of the most common ones.

Surface-level Representations A simple yet popular representation are the so-called *Hearst patterns* or *lexico-syntactic patterns* (Hearst, 1992). These patterns are regular expressions containing surface-level strings of words and sometimes also placeholders for named entities of certain types or wildcards for noun phrases. An example for a Hearst pattern is given in the following; here NP denotes a wildcard for a noun phrase, brackets group other elements, and the asterisk indicates an arbitrary number of repetitions:

Example 2.1 `such NP as {NP,}* {(or | and)} NP`

This rule extracts instances of the binary relation *is-a*, i.e., the hyponym-hypernym relation. Consider the following text example, also taken from (Hearst, 1992):

Example 2.2 `... works by such authors as Herrick, Goldsmith, and Shakespeare ...`

By applying the rule of Example 2.1 to the above phrase, three instances of the *is-a* relation can be extracted:

a) ⟨Herrick, author⟩

b) ⟨Goldsmith, author⟩ and

c) ⟨Shakespeare, author⟩.

This kind of rule representation requires only limited natural-language preprocessing, namely POS tagging, phrase recognition, and possibly NER. Many approaches to RE and EE are based on this representation or similar surface-oriented ones (Brin, 1998; Agichtein and Gravano, 2000; Etzioni et al., 2005; Pantel and Pennacchiotti, 2006; Pasca et al., 2006a; Kozareva et al., 2008; Hovy et al., 2009; Kozareva and Hovy, 2010a).

This rule representation is relatively robust. However, it is more suitable for dealing with relations formulated in local textual structures within small text windows. It does not work well for complex semantic relations where arguments are spread across different places in a long sentence, like in Example 2.3 (arguments are underlined):

Example 2.3 Giuliani will wed Judith, his true love for many years, in a public ceremony (the location is yet to be announced) on May 1.

Here, many words not relevant to the targeted relation separate the date and the spouses of the mentioned *marriage*. The relationship of the semantic arguments can often be more precisely inferred from the grammatical structure of a sentence.

Dependency-Grammar Patterns Another commonly employed formalism are dependency relations between words (Subsection 2.2.1). In contrast to surface-level patterns, these binary relations allow a rule formalism to skip unimportant parts of a sentence and to more directly connect the semantic arguments. For the Example 2.3 above, the dependency-relation analysis states that the verb "wed" is directly modified by the date "May 1", only separated by the preposition "on". As another example, consider the extraction pattern in Figure 2.5 (p. 27), which constitutes a template for a dependency parse of a sentence. Here, words in a sentence are either represented by lexical nodes (specified by a lemma and POS tag) or by entity placeholders with a particular named-entity type. The nodes are connected via dependency relations, and the semantic roles of entities are encoded in the rule representation. Applying this pattern to the second sentence in Figure 2.3 (p. 20) with the grammatical analysis in Figure 2.2 (p. 19) allows to extract the *marriage* instance in Example 2.4 (p. 27), with the results from co-reference resolution already taken into account.

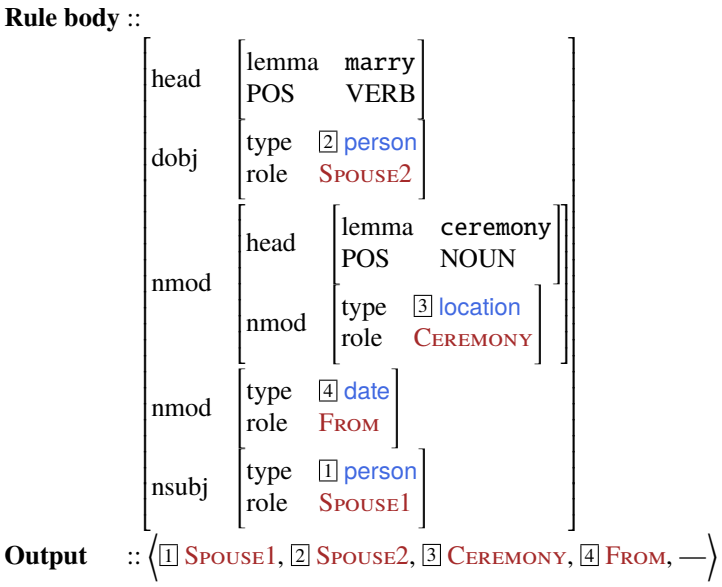


Figure 2.5 – Example extraction pattern for the *marriage* relation between people, based on the typed-dependency formalism. Identical to Figure 1.1a (p. 5).

Example 2.4

- ⟨SPOUSE1: `Brad Pitt`, SPOUSE2: `Jennifer Aniston`, CEREMONY: `Malibu`, FROM: `July 29, 2000`, TO: — ⟩

Approaches building their systems on top of dependency-relation analysis or closely related linguistic formalisms are presented, e.g., by Yangarber et al. (2000), Yangarber (2001), Stevenson and Greenwood (2005), Greenwood and Stevenson (2006), Suchanek et al. (2006a), F. Xu et al. (2007), and Adolphs et al. (2011).

Miscellaneous Representations More techniques have been employed to represent sentential structure for extraction purposes.

On the syntactic side, constituency parsing for the generation of features is applied occasionally (e.g., M. Zhang et al., 2006; Plank and Moschitti, 2013; T. H. Nguyen et al., 2015), but has lost momentum since the availability of data-driven dependency parsers (e.g., Nivre et al., 2007; Volokh, 2010) which no longer require the computation-intensive construction of full constituency parses as an intermediate step. Miyao et al. (2009) examined how the choice of syntactic representation effects the quality of an RE system. Their results indicate that while there are observable speed differences between parsers, the gap between formalisms with respect to the quality of the downstream extraction decisions is often negligible.

Illig et al. (2014) described an approach that makes use of *unsupervised* syntactic parsers well-suited for settings with either new-domain text that is very different from the domains for which datasets with manual syntax annotation are available, or for languages with generally low availability of linguistic resources. Another example of the range of employed frameworks is Abstract Meaning Representation (AMR), which captures not only a sentence's syntax but also its semantics in a graph structure. AMR parses of sentences are used by X. Li et al. (2015) to create features for EE.

2.3.3 Minimally Supervised Learning with Bootstrapping

The fully-supervised paradigm (Subsection 2.3.1) is the standard way of training ML systems, in particular in context of shared evaluation tasks (Subsection 2.7.2) where many labeled examples are available to learn from. However, for many application domains such labeled data is not readily available and costly to produce. The paradigm of *minimal supervision* in combination with an iterative learning scheme is one way to circumvent the manual annotation process, thereby allowing to use ML techniques outside of shared tasks and artificial evaluation scenarios. In the kind of minimally-supervised learning we look at, a *bootstrapping* process takes a limited number of examples as input, so-called *seeds*, and automatically labels free texts over the course of several iterations, using intermediate versions of trained classifiers or preliminary extraction-pattern sets. Figure 2.4 (p. 23) depicts minimal supervision with solid gray arrows: Instead of directly producing training data, human experts create (a few) relation instances whose automatically found mentions are then employed to generate data. This process is usually started by seeds in the shape of instances of a target relation (e.g., a few instances of kinship relations of famous people). However, seeds can also have the shape of manually written extraction patterns or manually annotated sentences containing relation mentions.

F. Xu et al. (2007) and F. Xu (2007) presented the DARE (Domain Adaptive Relation Extraction) system, which uses relation instances as seeds and a rule formalism based on dependency analysis.[22] DARE consists of two main parts: (a) extraction-rule learning and (b) relation-instance extraction. The rule-learning step and the instance extraction are coupled in a bootstrapping process, as shown in Figure 2.6 (p. 29). To start the process, the rule-learning component is provided with a small number of target-relation instances, then rules are learned from sentences which mention some of the named entities appearing as arguments in the seeds. Next, the learned rules are applied to texts to extract new instances, which in turn are employed as seeds for further iterations.

[22] The rule formalism was also extended to cover deeper linguistic methods such as head-driven phrase structure grammars (F. Xu et al., 2011; Adolphs et al., 2011; F. Xu et al., 2014).

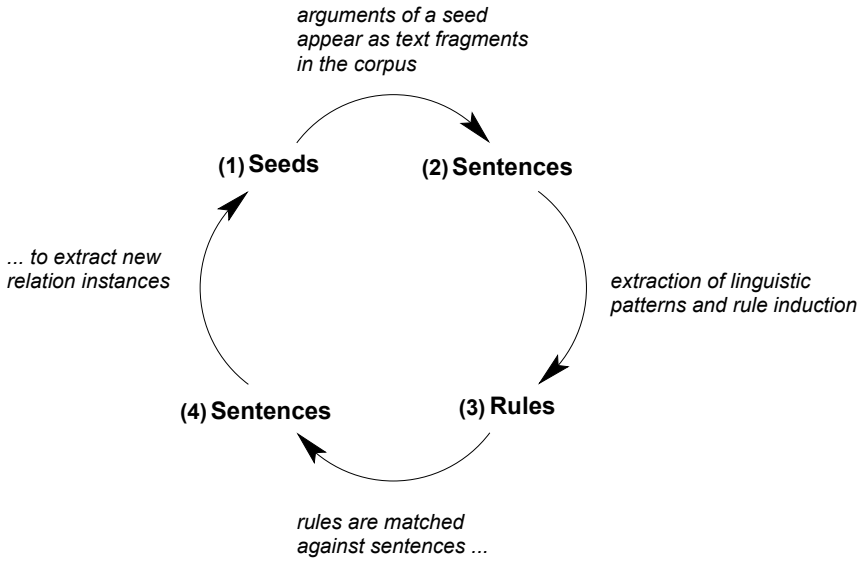


Figure 2.6 – Bootstrapping process of DARE.

The bootstrapping cycle stops as soon as no more new rules or instances are detected.

The iterative character of the rule learning, as well as other factors, lead to semantic drift and the propagation of errors across iterations, i.e., the generation of low-precision patterns. To combat this, DARE contains a component which assigns confidence values to both learned rules and extracted instances, which are calculated according to the duality principle (Brin, 1998; Agichtein and Gravano, 2000; Yangarber, 2001). This principle states that the reliability of rules is dependent on the trustworthiness of their extracted instances and on the confidence of the seed instances from which they stem. The confidence value of an extracted instance is calculated using the confidence in its ancestor seed instances. This heuristic is only able to filter out some of the incorrect rules, hence additional features for confidence estimation of rules have been implemented in DARE (F. Xu, Uszkoreit, and Krause, et al., 2010). Another caveat of minimally supervised approaches, including DARE, is that their performance is strongly dependent on the properties of the data used for training, i.e., on the specific linguistic variations in conjunction with the redundant mentioning of facts (Uszkoreit, 2011).

Many more systems aside from DARE made use of minimal supervision in the form of bootstrapping. Such approaches were not only popular for IE in the early 2000s (e.g., Agichtein and Gravano, 2000; Yangarber et al., 2000; Ravichandran and Hovy, 2002), but also continue to be employed in modern systems (e.g., Z. Zhang, 2004; Liao and Grishman, 2010a; A. Sun and Grishman, 2010; Gerber and Ngomo, 2011; Fujiwara

and Sekine, 2011; Bronstein et al., 2015). Due to the iterative character of the learning process, minimal supervision is particularly suited for settings with only little available structured information (e.g., intelligence applications). There are many other domains, however, where many knowledge resources are available (Section 2.5), from which a large number of relation instances of target relations can be created. This gives rise to a different kind of training paradigm which is discussed in the next subsection.

2.3.4 Distant Supervision for RE and EE

Another weak or distant form of supervision can be implemented by utilizing a large set of relation instances known to be true in combination with a huge corpus of texts from a domain related to the semantic target relations. The underlying assumption of this variant of supervision is similar to the one that was depicted in the previous section, namely, sentences which mention a tuple of entities known to participate in a semantic relation are assumed to express this relation. This assumption, or weaker forms, are referred to as the *distant-supervision (DS) assumption* in literature (e.g., Mintz et al., 2009; Riedel et al., 2010). DS is represented in Figure 2.4 (p. 23) with dashed gray arrows: Here, the relation instances whose mentions generate training data are retrieved automatically, in contrast to their manual creation by human experts in minimal supervision.

A number of IE systems follow this paradigm to overcome the time-consuming creation of training examples. Mintz et al. (2009), for example, proposed the idea to train a logistic-regression classifier on examples derived from mentions of Freebase relation instances in a large Wikipedia corpus. They focused on approximately 100 relations which are among the most frequent ones in their background database. The learned classifier worked on shallow features like word sequences and POS tags as well as on dependency relations between words. Their approach was succeeded by work from Yao et al. (2010), Riedel et al. (2010), and Hoffmann et al. (2011), who extended the DS approach with different kinds of probabilistic graphical models for improved RE performance, and who additionally tested their models on out-of-domain news data.

Wikipedia has not only been used as a source of texts for DS experiments, but has also been employed for relation-instance generation via their infobox system, which can potentially result in better performance due to a better alignment of texts with structured information. Examples include the Kylin system (F. Wu and Weld, 2007), the WOE system (F. Wu and Weld, 2010) and the Luchs system (Hoffmann et al., 2010). Each of these systems parsed Wikipedia infoboxes to generate relation instances, which they used as training examples for learning extractors based on conditional random fields over shallow sentence features. WOE also learned relation-independent patterns over

dependency parses of sentences. Another approach based on a weak form of supervision is called Universal schema (Riedel et al., 2013). Here, RE was cast as an inference problem with huge matrices, which represented at the same time the knowledge-base relations from Freebase and textual patterns from a news corpus, as well as the entities that participated in the relations and which were observed together with the patterns. Then, the matrices were factorized, which allowed the disambiguation of textual patterns against the knowledge-base relations and which gave information about new relation instances not present in the knowledge base but mentioned in the corpus. Toutanova et al. (2015) extended the original approach with a more sophisticated pattern formalism. Another approach that has many similarities with Universal Schema was presented by Weston et al. (2013).

Due to tuples of entities participating in more than one relation (e.g., persons dying in their birthplace), incompleteness of KBs (e.g., missing recent acquisitions of a company), and errors in linguistic preprocessing, the DS assumption is violated in many cases. This is a major cause for the low accuracy of systems trained with distant labels and various works focused on mitigating this problem. Angeli et al. (2014) employed *active learning*, i.e., they integrated human feedback into the learning process. Takamatsu et al. (2012) and Roller et al. (2015) aimed to automatically identify unreliable distant labels using learned relation paths in the background KB and a generative model assessing the quality of the weak labeling process. For Universal Schema, Rocktäschel et al. (2015) presented a way to incorporate hand-written constraints into the factorization process. Despite these attempts to reduce noise in DS, headroom for performance improvement exists (Section 2.7).

2.3.5 Self-supervised Learning and Open IE

Another type of learning scheme was developed for scenarios where no separating line between interesting relation types and uninteresting ones is defined a-priori. This might be the case because the relations types covered by a text are new or because the kind of information mentioned in texts is unknown. In such situations, anything might be useful and hence more or less *all* relational information from a text collection should be extracted. This paradigm is called *open information extraction* (open IE) because it is not restricted to pre-selected domains.[23] Figure 2.4 (p. 23) pictures this learning scheme with dashed black arrows: Both relation definitions and training data are produced directly from texts, without manual intervention or accessing of existing knowledge bases.

[23] See also the introduction of Chapter 5, which further motivates the need for open extraction methods.

A predecessor of modern open-IE systems is KnowItAll (Etzioni et al., 2004a; b; 2005), a system mainly intended for harvesting entities and their types from web pages. Using hand-crafted lexico-syntactic patterns as queries to search engines, KnowItAll was capable of extracting thousands of facts from the web. The subsequent TextRunner system (Banko et al., 2007; Yates et al., 2007) then introduced the notion of open IE. In the TextRunner's training phase, a dependency-parsed corpus was filtered for sentences containing at least two noun phrases. Lexical sentence features and POS tags from the words connecting the noun phrases on the sentence's dependency graph were then used as training examples for a classifier. This classifier was subsequently applied to a large POS-tagged corpus to determine for each sentence if it contained a mention of a relation. As there were no manually constructed examples, this approach was referred to by the authors as self-supervised. In this setting, the name of a relation was derived from the words connecting two entities (noun phrases in this case) in a sentence. One drawback of this approach was that the semantics of a given observed relation between entities was not immediately clear and had to be determined in a separate step. The Resolver system (Yates and Etzioni, 2007; 2009), as well as Soderland et al. (2010)'s work, implemented such a disambiguation step. Later systems often had this disambiguation step built into their (pattern) extraction methodology (e.g., Movshovitz-Attias and Cohen, 2015). Another shortcoming of early open-IE systems was the large amount of produced noisy extractions, which was addressed by utilizing language models to assess the quality of extractions (Downey et al., 2007), as well as by models capturing factual redundancy in the source texts (Downey et al., 2005; 2010).

The TextRunner system was later refined in many ways, e.g., by additionally implemented constraints on extracted predicates which restricted the allowed POS-tag sequences. The resulting systems ReVerb (Fader et al., 2011) and R2A2 (Etzioni et al., 2011) featured better performance and more meaningful output, and constituted the second generation of open-IE systems. Other techniques aimed at improving the open-IE approach through smarter pattern extraction algorithms and upstream sentence simplification (Schmidek and Barbosa, 2014; Angeli et al., 2015; Niklaus et al., 2016). Many more IE approaches were developed to address the need for less supervision and to allow for a discovery of new and emerging relations. Early examples include Shinyama and Sekine (2006) and Sekine (2006), but there are also recent approaches heading towards a similar direction (Akbik et al., 2012; 2013). Another line of research dealt with the problem of headline generation for clusters of news articles by applying open-IE techniques (Alfonseca et al., 2013; Pighin et al., 2014; Congle Zhang and Weld, 2013; Congle Zhang et al., 2015).

2.4 Discourse-Level Processing of Text Documents

The approaches to the extraction of relations and events which have been discussed so far in this chapter are predominantly sentence-level models, i.e., extraction decisions are made on a per-sentence basis. In the classic IE pipeline architecture, these local decisions happen after references to concepts and entities in the text have been detected and resolved, a step which naturally requires the processing of document-wide context. In contrast, this section reviews approaches which relax the sentence restriction for the downstream RE/EE step, first by incorporating different forms of beyond-sentence content for improved intra-sentential performance and finally by directly handling cross-sentential fact extraction and producing document-level output.

2.4.1 Incorporating Various Forms of Wider Context

A fundamental observation regarding the interpretation of polysemous words was made by Gale et al. (1992) and Yarowsky (1993), who state that the sense of a word is typically consistent within a given discourse and that context words give strong clues as to how a word should be interpreted. This principle is usually referred to as *one sense per discourse* and *one sense per collocation* and has since proven to be true for NLP tasks other than pure word-sense disambiguation (for entity recognition by Barrena et al., 2014; for machine translation by Carpuat, 2009). The underlying intuition of this principle was exploited for RE/EE as well, e.g., by Ji and Grishman (2008), who retrieved additional topically related documents after intra-sentential extraction in order to verify recognized trigger words and argument-role labels of entities. The idea behind Ji and Grishman's principle is that the related documents are consistent with their original versions to the extent that the same entities participate in the same relations and that they occupy the same role. This consistency intuition was implemented via nine hand-coded inference rules operating over event mentions. A less elaborate approach was presented by Huang and Riloff (2012), who also performed sentence-level extraction before wider context was exploited in a post-processing step. In their approach, this wider context was the set of sentences in a document, for which a conditional-random-field model was trained to identify non-coherent sentences. Extractions obtained from these sentence during the intra-sentential processing phase were then discarded.

Liao and Grishman (2010b) based their event extraction approach on a similar observation, namely that certain groups of event types tend to co-occur frequently, e.g., events covering different aspects of terrorist attacks. In this example, an event that reports the actual attack is likely to occur in close vicinity to events describing the injuring and

death of people. In a first step, the system of Liao and Grishman searched for event mentions using standard within-sentence extraction means and applied a strict filtering so that only the high-confidence ones would be kept. The presence and distribution of these high-confidence events (i.e., their types) served as additional input to a second pass of the system, which then made more informed extraction decisions by assessing local (within-sentence) features based on the document-level event context (i.e., the presence of particular other event types). Later, the same authors used the principles described above to enhance an iterative semi-supervised event extraction system. Eventually by applying these ideas, error propagation between iterations was kept to a minimum (Liao and Grishman, 2011).

Ji et al. worked on a task variant of event extraction intended to evaluate the impact of IE techniques for applications more directly (Ji et al., 2009; 2010). The authors aimed to align events with respect to salient entities in clusters of related documents and to measure how quickly human users could find the information they deemed interesting. By aggregating the event-relevant information for particular entities from different documents, they also produced event information which was consistent across documents, thus following the ideas outlined above.

Yao et al. (2010) learned correlations of entity types and relations (so-called selectional restrictions) in a cross-document fashion during training in a distantly supervised RE scheme. These restrictions allow to exclude low confidence extractions after the initial intra-sentential processing is completed. Similarly, Q. Li et al. (2011) learned constraints from training data about the type-wise compatibility of events. An example constraint is that politicians who reportedly met at an international summit usually do not represent the same country. At inference time, these constraints are applied after single-sentence extraction in order to identify the extractions that need to be dropped to achieve global coherence. A very different type of context was utilized by Ji (2010) and A. Lee et al. (2010), who do not only go beyond the processing of individual sentences, but also proposed to do inference in a cross-lingual and cross-media (text and video) fashion.

2.4.2 Cross-sentence Mentions of Relations and Events

The restriction of relation and event mentions to the sentence level is a simplification which allows for robust computational approaches, but does not sufficiently reflect the actual distribution of information in text. In his studies (Stevenson, 2004; 2006), Stevenson analyzed the distribution of relation arguments in datasets from a shared task which took place in context of the Message Understanding Conference series. He found that in the three analyzed datasets with the domains of terrorist reports, management

> Shortly after 6:30 on the evening of December 22, the guests were invited, without fanfare, to take their seats. Guided by the glow of hundreds of candles, Gwyneth Paltrow, Rupert Everett, Donatella Versace, a kilt-clad Sting and some 55 others gathered near the foot of the grand staircase in the Great Hall of Scotland's 19th-century Skibo Castle. As the skirls of a lone bagpiper gave way to the music of French pianist Katia Labèque and a local organist, the wedding ceremony of Madonna Louise Ciccone, 42, and film director Guy Ritchie, 32, began.

Figure 2.7 – Example of relation mention that extends over several sentences. This excerpt from a news article mentions a *marriage* relation between two persons, as well as the date and location of the wedding. The arguments (underlined) are mentioned in three different sentences, meaning they are not accessible by traditional intra-sentential extraction approaches. Retrieved from `http://people.com/premium/madonnas-wedding-story/`, last access: 2017-03-03.

succession in organizations, and rocket launch reports, up to 40% of facts from the gold-standard require inter-sentential extraction approaches, i.e., ones which cross the sentence boundary.[24]

More studies have later confirmed Stevenson findings. Swampillai and Stevenson (2010) built on the results of Stevenson (2006) and additionally investigated cross-sentential properties of relations in the 2003 dataset of the Automatic Content Extraction program. They observed that almost ten percent of relations in this dataset are inter-sentential. Ji and Grishman (2011) determined that cross-sentence analysis was required for approximately 15% of the attributes in the Knowledge Base Population task of the 2010 Text Analysis Conference. Similar observations were made for the 2013 & 2014 iterations of this task (Surdeanu, 2013; Surdeanu and Ji, 2014): Approximately one-fifth to one-third of missing attributes were due to scattering of arguments across documents, implicit arguments, and required inference with world knowledge. As an example for a relation mention which cannot be extracted with conventional intra-sentential methods, consider the excerpt of a news article in Figure 2.7. In this text, four arguments of a *marriage* event are present and spread across three sentences. In the following, we discuss methods that aim to handle such cross-sentence mentions.

[24] It must be noted that his analysis does not count anaphoric references as relation arguments and does not verify that all cases of relation-argument mentions are actually embedded in a reference to the relation, hence the amount of cross-sentence relation mentions is likely over-estimated. However, the author addresses these caveats to some extent in an additional experiment, whose results indicate that in fact a significant portion of mentions are inherently cross-sentential.

Cross-sentence Extraction through Mention Linking One way to produce document-level relation and event mentions is to execute a downstream resolution step on top of intra-sentential information fragments. This step is analogous to the application of (entity) CR after NER and is often referred to as *linking* (of mentions of relations and events). In this step, pairs of mentions or groups of them are compared in an iterative manner, in order to determine if they share relevant properties (common arguments, overlapping lexical terms or synsets in WordNet, etc.) and whether it is likely that they refer to the same real-world fact. There is a wide range of features, methods, and models that are used in literature for this comparison. Chapter 7 presents an approach in this direction and illustrates the types of features commonly used. In the following, we briefly report some representative approaches for this type of inter-sentential extraction.

Common methods that decide whether given mentions of relations and events are co-referential include random forests (Z. Liu et al., 2014) and graph-clustering algorithms (Z. Chen and Ji, 2009; Sangeetha and Arock, 2012). Other methods made use of agglomerative clustering (Z. Chen et al., 2009) and simple left-to-right clustering schemes (Ahn, 2006), as well as Markov logic networks (J. Lu et al., 2016). Cybulska and Vossen (2015) worked on co-reference resolution for mentions within and across documents. They employed decision trees, support vector machines, and Naive Bayes for classification. Bejan and Harabagiu (2010) reported to use non-parametric Bayesian models with standard lexical-level and semantic features, as well as WordNet-based ones. Later, they extended their approach (Bejan and Harabagiu, 2014), others presented similar ideas (Yang et al., 2015). Recently, Peng et al. (2016) proposed to construct vector representations for event mentions by concatenating dense embeddings (see Section 2.6) for the elements of an event mention as determined by a semantic role labeler, followed by a simple left-to-right clustering algorithm that compares these vectors via cosine similarity.

A number of approaches did not follow this standard pattern of factual co-reference resolution. For example, J. Lu and V. Ng (2015) implemented an approach, in which a hierarchy of heuristics (so-called sieves) determined whether or not a pair of event mentions should be linked. The heuristics had different trade-offs of precision and recall and were applied subsequently to the data. This is an idea first introduced and successfully applied to entity co-reference resolution (Raghunathan et al., 2010). T. Zhang et al. (2015) worked on cross-media event co-reference resolution by combining news videos and the corresponding closed captions. Sachan et al. (2015) described an active-learning based method for the problem of joint cross-document and within-document resolution, where a clustering of mentions was derived by incorporating bits of human judgment

as constraints into the objective function. H. Lee et al. (2012) handled the resolution of references to entities and events in a joint manner, motivated by the insight that noun phrases can potentially refer to either class and that joint handling can take advantage of the participation of (newly resolved) entities in semantic roles of events. Their model iteratively compared pairs of clusters and the mentions in them using relatively basic features. Then, cluster-merging decisions were made via linear regression. Araki et al. (2014) reported on experiments with more fine-grained co-reference relations between events, i.e., in addition to full co-reference, also parent-child-like and sibling-like relations were considered. Their results suggested that this more elaborate way of representing event-mention relations allowed for better resolution performance. Araki and Mitamura (2015) simultaneously identified event triggers and disambiguated them using a structured-perceptron algorithm.

Holistic Approaches with Cross-sentence Mentions Another branch of methods directly models the recognition of relation and event mentions on the document level. In these holistic approaches, the recognition step already considers entities from more than just one sentence. The common topic of these approaches is to extend the (syntactic) graph representation of single sentences in IE (e.g., dependency parses) with edge types that build bridges between sentences. Standard methods for processing the graphs can then be applied to determine whether a given piece of text contains a fact reference and, if applicable, which mentioned entities participate in it. Melli et al. (2007) proposed such an approach. They built a graph from constituency parses of sentences, with cross-sentence connections being introduced by co-reference relations between entity mentions, as well as by connecting the last and first word of adjacent sentences with a special edge.

Swampillai and Stevenson (2011) presented another instance of this idea. They focused on the prediction of binary relations and built cross-sentence syntactic graphs by connecting the root nodes of two constituency parses. Then, support-vector machines processed the shortest paths connecting the two entities of interest, additionally taking into account features like the mere distance of the entities and local features from the respective context of the entities. Recently, Quirk and Poon (2016) proposed another method in this scheme. For windows of three consecutive sentences, they built a graph representation consisting of the words of each sentence in this window. Edges were obtained from the dependency parses of the sentences and from the analysis of a discourse parser. Additional edges were inserted between adjacent words and for co-reference relations between entity mentions. Finally, dependency-parse roots of neighboring sentences were connected as well. In order to make a linking decision for a given candidate pair of

entity mentions, multiple paths connecting the entities were generated and subsequently presented to a logistic-regression classifier.

2.5 Structured Knowledge Resources in IE

Knowledge resources are of great importance for IE. Apart from being the goal of extraction attempts (in knowledge-base building), resources can provide seed information in training settings with lighter forms of supervision and can be utilized to assess the quality of extracted facts and learned patterns. While in the past, resources used to be created by hand in long-running annotation efforts, recent advances in IE and other areas facilitate the automatic creation of such resources. An important line of work with knowledge resources focuses on how they can be linked in efficient yet expressive ways, i.e., how they can be integrated into the semantic web. Berners-Lee (1999, p. 177) gave an early definition by characterizing it as "a web of data that can be processed directly and indirectly by machines." The concept of *linked data* can be seen as an instantiation of this idea, which gives detailed practical instructions on how resources should be formatted in order to become a proper member of this web of data (Bizer et al., 2009a). Linking repositories of data enables new types of applications and even increases the impact of individual knowledge resources on tasks like IE. In the following, this section introduces important resource types and instances of these.

2.5.1 Manually Created Resources

Perhaps one of the most cited examples of resources relevant for NLP is WordNet (G. A. Miller, 1995; Fellbaum, 1998). WordNet is a machine-readable lexical database with dictionary and thesaurus functionality for approximately 150,000 English words.[25] The two main components of this resource are groups of near-synonymous words (synsets) and the semantic relations between these groups. The relations are of lexical-semantic nature and include hyponymy, antonymy, meronymy, entailment, and troponomy. One of the ways in which WordNet was used for IE is as a source of *is-a* relation instances for DS training (e.g., Snow et al., 2004). Linguistic knowledge resources that go beyond the level of lexical items are scarce and of limited coverage due to significant investments of human effort and expertise required for their construction. FrameNet (Baker et al., 1998; Ruppenhofer et al., 2010) serves as an example of a resource for English that documents the range of semantic and syntactic combinatorial possibilities of words and their senses.

[25] `http://wordnet.princeton.edu/wordnet/man/wnstats.7WN.html`; last access: 2017-02-25.

FrameNet consists of schematic representations of situations (called frames), e.g., the frame *win-prize* describes an awarding situation with semantic roles like COMPETITOR, PRIZE, COMPETITION, etc. A pair of word and frame forms a lexical unit, similar to a word sense in a thesaurus. These units are connected to lexical entries, which capture the valence patterns of frames, providing information about their semantic roles and their phrase types and grammatical functions in relation to the lexical units. As of 2017, FrameNet contains information for more than 1200 frames.[26] A downside of FrameNet with respect to IE application is that it does not provide an explicit link to real-world fact types. In Chapter 6, we present a resource that mitigates this shortcoming by connecting FrameNet to relation-type-specific linguistic patterns.

Another manually created resource whose origins date back to the late last century is Cyc (Lenat, 1995; Matuszek et al., 2006). Cyc is a large KB with formalized common-sense knowledge (e.g., you have to be awake to eat; Lenat, 1995, p. 33), created with the intention of being useful for AI applications reasoning about the world. Due to the general domain of the contained information, it was extended to and used in a diverse set of areas, such as biology/medicine (Witbrock et al., 2015), education (Lenat and Durlach, 2014), and terrorism information management (Deaton et al., 2005). Cyc currently features more than 500,000 concepts, which are defined and elaborated using approximately 7,000,000 facts.[27] A further resource gathering common-sense knowledge is ConceptNet (Speer and Havasi, 2012; Speer et al., 2016), which differs from Cyc in that it focuses on knowledge needed for understanding human language. It is also less formalized, essentially being a network of words/phrases connected by predicates. For the English language, it contains approximately 11.5 million assertions (Speer and Havasi, 2012).

A third type of resource, besides repositories with lexical-semantic or common-sense knowledge, are databases which concentrate on gathering factual knowledge about real-world concepts and entities. Such databases have been used extensively for distant supervision of RE systems (e.g., Mintz et al., 2009). An instance is Freebase, a freely available graph-structured database with more than 100 million assertions (Bollacker et al., 2008), started as a collaborative annotation/editing effort. In the meantime, the data from Freebase was integrated with two other ontologies, the proprietary Knowledge Graph (Singhal, 2012) of Google, and Wikidata (Vrandečić and Krötzsch, 2014). The latter continues to gather knowledge in a collaborative manner, which resulted so far in approximately 25 million concepts represented in the database.[28] Like the other resources introduced above, Wikidata is available as linked data (Erxleben et al., 2014).

[26] `https://framenet.icsi.berkeley.edu/fndrupal/current_status`; last access: 2017-02-25.
[27] `http://www.cyc.com/kb/`; last access: 2017-02-25.
[28] `https://www.wikidata.org/wiki/Wikidata:Statistics`; last access: 2017-02-25.

2.5.2 Processing of Wikipedia and Other Resources

The online encyclopedia Wikipedia is not only a vast repository of text, but it also features much semi-structured content. In particular, the categorization of articles and the widespread utilization of infoboxes (tables of attribute-value pairs embedded into articles) are rich sources of information (see also Subsection 2.3.4). This is why Wikipedia served as the basis for the automatic construction of many resources. Hovy et al. (2013b) argued that the use of such semi-structured, collaboratively created and edited resources has many advantages over the extraction of knowledge from unsupervised sources (text) and the manual creation of structured KBs. Most notably, the semi-structured knowledge is of high quality and high coverage, and comes at low cost. DBpedia (Bizer et al., 2009b; Lehmann et al., 2015) is one of many efforts to facilitate the access to the structured information in Wikipedia. Its main motivation is the insight that structured KBs are hard and cost-intensive to maintain; at the same time, many volunteers keep Wikipedia's articles and the associated structured content up-to-date. This makes the population of KBs from Wikipedia attractive. The DBpedia content that originates from the English version of Wikipedia is an ontology with more than four million concepts.[29] Finally, DBpedia is linked to a great many resources, which is why it counts as an important example of linked data (Dojchinovski et al., 2016).

WikiNet (Nastase et al., 2010) is another resource that is built on top of Wikipedia, similar to DBpedia in its goals, but going one step further with heavier post-processing of the extracted structured information. WikiNet's concepts were extracted from articles and their categories. Relations between these concepts were obtained from infobox attributes as well as from relationships encoded in category names. As an example, consider the category name *Movies directed by Woody Allen*. This category implies that for concepts in this category, the relation *directed by* holds with the argument `Woody Allen`. The WikiNet resource contains approximately three million concepts with more than 38 million relational statements.[30] A similar methodology was followed by the long-running YAGO (Yet Another Great Ontology) project (Suchanek et al., 2007; 2008), which started around the same time as the DBpedia effort, i.e., almost a decade ago. In addition to the use of the semi-structured parts of Wikipedia, YAGO employed WordNet, with the main idea being that a combination of the vast set of individual concepts in Wikipedia with the less chaotic taxonomy from WordNet would allow to create a more accurate resource than possible by processing only Wikipedia. Later (Hoffart et al., 2013), YAGO was extended with information from a further repository with geographical data.

[29] `http://wiki.dbpedia.org/about`; last access: 2017-02-25.
[30] `https://www.h-its.org/en/research/nlp/wikinet/`; last access: 2017-02-27.

Currently, YAGO contains information from ten language-specific Wikipedias, providing 120 million facts about 10 million entities.[31] More works in the context of the YAGO project are mentioned later in Section 3.6.

Similar to YAGO, the construction process of the resource BabelNet (Navigli and Ponzetto, 2012) also drew information from Wikipedia and WordNet, but with a focus on retaining and extending the lexical-semantic knowledge in WordNet, which differentiates it from the repositories described above. BabelNet's core components are the BabelNet synsets, which are sets of multilingual synonyms, created from WordNet synsets and Wikipedia concepts, which were merged (where appropriate) in a disambiguation step. Each Babel synset is related to other Babel synsets by semantic relations such as hypernymy, meronymy and semantic relatedness, obtained from both its source resources. Since its initial creation, BabelNet was extended with information from a number of additional repositories, among them other lexical-semantic resources like FrameNet and multilingual wordnets. One major difference between BabelNet and WordNet is their considerably different size, both in terms of number of concepts and semantic relation instances. On the one hand, WordNet provides roughly 100K synsets, 150K lexicalizations, and 300K relation instances. On the other hand, BabelNet contains roughly 13.8M synsets, more than 740M lexicalizations and 380M relation instances.[32] Moreover, given the multilingual nature of BabelNet (version 3.7 considers 271 different languages), this resource can exploit multilinguality to perform state-of-the-art knowledge-based word-sense disambiguation (in contrast to WordNet which encodes only English lexicalizations) and thereby it enables new methods for the automatic understanding of the multilingual web.

UBY (Gurevych et al., 2012) is a meta-resource which covers a similar set of resources as BabelNet does, however concentrating on only two languages (English and German) and following a different methodology with respect to the integration of the source resources. While BabelNet integrates information from different repositories at a very fine-grained level, UBY retains the independence of the original resources in the newly created one. The main focus is to provide a unified representation for the original resources while the full degree of originally available information is still accessible. In contrast to superimposing semantically equivalent elements, links between them are inserted. The UBY version described by Gurevych et al. (2012) is based on ten lexical-semantic ontologies and covers more than four million lexical entries and more than five million relations between their senses. In order to allow lexical meta-repositories like UBY to

[31] `http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/`; last access: 2017-02-27.

[32] `http://babelnet.org/stats`; last access: 2017-02-27.

be expressed in terms of standard semantic-web technologies, the Lemon framework (lexicon model for ontologies; McCrae et al., 2011) was designed. Lemon is a linked-data format for language expressions, which allows to individually and unambiguously address linguistic items present in lexical-semantic repositories. This format removes the need for maintainers of meta-resources, as the creators of new repositories can integrate their resource into the cloud of linguistic data resources by publishing in this particular format. Examples include the Lemon-formatted version of UBY (LemonUBY; Eckle-Kohler et al., 2015), a BabelNet variant (Ehrmann et al., 2014), and a Lemon lexicon for representing DBpedia content (Unger et al., 2013).

2.5.3 Ontology Building through the Analysis of Text

The KBs discussed so far in this section concentrate on factual information about entities and capture the semantics of words and phrases. In addition, the focus was on repositories that have either been manually created or that have been obtained by processing mainly the semi-structured and structured elements of other databases. In contrast, the efforts described below analyze the huge body of text in Wikipedia and on the web, and align this information in pre-defined or ad-hoc taxonomies.

WiSeNet (Moro and Navigli, 2012; 2013) is a repository of language expressions in the shape of relation phrases, obtained with an open-IE methodology (Section 2.3). Phrases connecting two concept mentions within a sentence were collected from Wikipedia articles and grouped into clusters based on the distributional properties of their relational arguments. For generalization, concrete arguments were resolved into their Wikipedia categories, which in turn produced a network of specific concepts linked with various relation phrases. The PATTY resource (Nakashole et al., 2012a; b) went one step further by arranging relation phrases from Wikipedia, a news corpus, and crawled web pages in a taxonomy. The phrases were obtained by extracting shortest paths between entities from syntactic analyses of the sentences in the corpus. For a single source sentence, a large number of pattern variants was generated, e.g., by variation of the fine-grained entity types of the arguments, by syntactic generalization of the lexical items on the path, and by generalization through the insertion of placeholders. The generalization relations between phrases as well as their support set overlaps then allowed to arrange the patterns in a subsumption hierarchy. PATTY was later extended with an alignment of phrases to verbs of WordNet, as well as more intricate and accurate determination of subsumption links between relation phrases and the clever utilization of multilingual corpora for again improved taxonomy construction (Grycner and Weikum, 2014; Grycner et al., 2015; Grycner and Weikum, 2016).

More instances of phrase-centered repositories, which largely employed similar techniques during construction as already described, exist. For example, Akbik and Michael (2014) presented a repository of phrases extracted from a large dataset of syntactic n-grams occurring in English books (Goldberg and Orwant, 2013). Due to the different nature of the texts compared to Wikipedia, news texts, etc., the resulting resource is a collection of common-sense knowledge, rather than standard relational information. Yet another source of texts was exploited by Delli Bovi et al. (2015), who processed textual definitions from BabelNet. Classic distantly-supervised (non-open-IE) methods were employed by Kirschnick et al. (2014) and Akbik et al. (2014) to create databases with RE patterns that can be easily explored using web-based interfaces by users. Biberpedia (Gupta et al., 2014) is an approach utilizing text analysis in addition to the processing of query streams from a web-search engine to generate an ontology of class attributes in an automatic fashion. This is to be considered in contrast to instance-oriented approaches like the one of Dong et al. (2014), who focused on extending an existing factual KB with automatically information extracted from the web in various ways, among them the leveraging of IE methods on free text, the analysis of semi-structured information in web pages (page structure and tables), as well as human labeling.

2.6 Text Analytics with Neural Methods

So far, this chapter has reviewed classic machine-learning approaches, which can achieve remarkable classification results in many tasks. A shortcoming of these methods is their inability to process data (here: texts) in its raw form. Instead a representation as feature vectors needs to be generated. The definition of these features is mostly done by hand. The automatic learning of these representations from raw data, i.e., having a learning procedure figure out the best features, is at the heart of the research area of deep learning (LeCun et al., 2015).

The prototypical kind of models in this area are neural networks, whose basic building blocks are neurons. These neurons are simple functions that produce a scalar output value through a linear combination of a number of input values, followed by a non-linear activation function. Neurons are often arranged in layers, with the lowest layer having access to the raw input data and higher layers processing the outputs of the earlier ones. This multi-layer way of arranging the neurons allows the first layers of a network to learn to recognize low-level patterns or features in the input data, which in later levels can be combined to high-level features. An illustrative example is the work on object recognition from images. Here, the task is to determine the kind of object that a picture

contains, where the data is presented in raw pixel form to the model (Le et al., 2012). Zeiler and Fergus (2014) and others analyzed the kind of features that the different layers in neural networks learn to recognize. They found that, for the task of object recognition, successful models can learn to recognize simple structures like edges in earlier layers. The presence of these structures is then used to determine if and where corners and contours are present in the given picture, which in turn are combined to find more complicated geometric structures in the top levels.

Neural networks have a long history (going back as far as to the 1940s) and the use of them has been known by different names over time (cybernetics, connectionism, deep learning) and has passed times of less popularity (Goodfellow et al., 2016, p. 12). A number of developments caused the renewed interest in neural-network research since approximately 2006. New ideas for training particularly large networks with many layers as well as the availability of larger datasets and the computational power needed to process them made it possible to train models with more layers than before (Goodfellow et al., 2016, pp. 18–21). DL research was further fostered by the design of frameworks for large-scale distributed neural computations (e.g., DistBelief by Dean et al., 2012) and by the wide dissemination of scientific computing frameworks which also allow the implementation of deep models (Jia et al., 2014; Al-Rfou et al., 2016; Abadi et al., 2016a; b). Work considered breakthroughs from the very recent past exemplifies that DL is a very active research field (e.g., Mnih et al., 2015; Silver et al., 2016), which still has some blind spots in its theoretical foundations (e.g., Chiyuan Zhang et al., 2016).

Architectures In addition to classic feed-forward neural networks with multiple layers of stacked feature detectors, a great deal more architectures were considered over time. Schmidhuber (2015) gave a detailed account to developments in the field. Many of these architectures were used for NLP and IE; Goldberg (2016) reviewed the use of neural models for natural-language problems. One particular architecture are convolutional neural networks (CNNs), a type of network invariant to small variations of the input data which are irrelevant to the classification problem. CNNs were successfully applied to computer vision problems (e.g., Krizhevsky et al., 2012; see also Goodfellow et al., 2016, pp. 359–361), before being used for the extraction of relations (Zeng et al., 2014; T. H. Nguyen and Grishman, 2015b; Y. Liu et al., 2015; dos Santos et al., 2015; Zeng et al., 2015) and events (Y. Chen et al., 2015; T. H. Nguyen and Grishman, 2015a). An example of a CNN can be seen in Chapter 7, which proposes such a model for cross-sentence fact extraction.

A further important architectural option for neural networks in NLP are *recursive* and *recurrent* designs. These models are particularly suited for processing tree-structured and sequence-structured input data, which makes them a good fit for natural language where both types of structures are present in texts and grammatical analyses thereof. A key idea of such networks is to have a component of a fixed size which is repeatedly applied to different parts of the input data, this way allowing to handle texts of varying length. The input to the network at a given step then includes directly (recursive networks) or indirectly (recurrent networks) the preliminary output of the model for the sequence processed so far. Goodfellow et al. (2016, pp. 363–408) gave a detailed introduction to both forms of networks. Socher et al. (2012), Hashimoto et al. (2013), and Ebrahimi and Dou (2015) proposed recursive models for RE; Y. Xu et al. (2015) and J. Li et al. (2015) recurrent ones.

Embeddings A key element of recent neural models for NLP problems are *embeddings* of texts at various granularity levels (words, sentences, paragraphs, documents), i.e., real-valued vectors which represent texts via latent, automatically learned features. These embeddings are created as an intermediate result of neural networks working with input text. However, these embeddings can also be trained independently of the actual NLP task to which they are later applied, due to the general character of these representations. This is exemplified by multi-task approaches like the one of Collobert et al. (2011), who trained a neural network for the tasks of part-of-speech tagging, chunking, named-entity recognition, and semantic role labeling, where the intermediate parts of the model were shared across the tasks.

Apart from methods for producing generic embeddings of sentences (Kiros et al., 2015) and paragraphs (Le and Mikolov, 2014), word embeddings have attracted much attention. Mikolov et al. reported a particularly efficient way of training word embeddings, applicable to any large text corpus without the need for any manual labeling or sophisticated linguistic preprocessing (Mikolov et al., 2013a; b). Follow-up work improved the quality of embeddings via syntactic parsing of texts (Levy and Goldberg, 2014a) and by putting additional focus on the representation of rare words (Sergienya and Schütze, 2015). An interesting question is to what extent linguistic regularities between words are reflected in the word-embedding space and whether algebraic operations on word representations produce meaningful results. Unfortunately, studies on this question did not result in fully conclusive results (Mikolov et al., 2013c; Levy et al., 2015). It should be noted that embeddings of words are by now also available from non-neural approaches (Pennington et al., 2014; Lebret and Collobert, 2014), with comparable quality (Lebret et al., 2013;

Levy and Goldberg, 2014c). Neural networks have influenced IE not only via new neural models, but also via the provision of word embeddings. For extraction approaches that already work with feature vectors, the utilization of word embeddings is straightforward. Here, embeddings can serve as drop-in replacements for hand-crafted features on the lexical level. An example is the work of T. H. Nguyen and Grishman (2014). The authors extended the RE classifier of A. Sun et al. (2011) with embeddings and observed that this allowed for easy domain adaption of models. Further approaches making use of word embeddings for RE purposes were presented by, e.g., Yu et al. (2015), Gormley et al. (2015), and Hashimoto et al. (2015).

2.7 Evaluation

This section reviews means for the evaluation of IE problems. We start by presenting four common evaluation strategies.

1) Extrinsic Evaluation from the Perspective of Users Ultimately, the usefulness of an IE system is determined by the value its output has to users of such a system. Of interest are questions like: how reliable is the extracted information; how much information is extracted; how new is this extracted information with respect to what a user already knows; how much time does it safe the user; etc. There have been evaluation approaches for IE which aimed at directly measuring this usefulness to end-users. An example is the browsing-cost metric, assessing how many non-useful facts, i.e., redundant ones or incorrect ones, a user has to look at before finding a certain amount of correct information (Ji et al., 2009; Q. Li et al., 2011).

2) Evaluation with Manually Created Ground-Truth Data A more common way of assessing the quality of an IE approach is to compare, for a given piece of text, the output of an extraction system against a ground-truth of human-annotated information. In fully supervised training schemes, given that enough annotated data is available, the bulk of data is used for model training, while some is reserved for development and testing purposes. Strategies like *cross validation* (Mohri et al., 2012, pp. 5–6) allow to ascertain system quality in scenarios with less available labeled data. Annotated data for IE tasks often comes in the shape of mention-level annotation. For relation extraction this usually means that the annotation explicitly states where the participating entities for a relation are mentioned in a given text, and where potential lexical clues (*triggers*) for the mention are located.

3) Evaluation with Distantly Labeled Data If there is a lack of labeled data, evaluation options exist which trade accuracy of the evaluation with a reduced need for manually created labels. One such option is a form of automatic annotation similar to the idea of distant supervision. In the case of RE, for a given set of documents, all the relations in a factual knowledge base with arguments being mentioned in the documents would be part of the ground truth. This weak way of producing a gold standard introduces at least two types of noise. The first one is that it overestimates the amount of fact mentions in texts due to the DS assumption, which can easily be violated by tuples of entities participating together in more than one relation. The second flaw of this evaluation type is the use of an invalid closed-world assumption, namely that all information referred to in text is listed in the database. Documents published after the latest curation of the structured knowledge are, however, likely to contain new factual information that would not be reflected in the database. Furthermore, entities and their properties might be missing in the database simply because they are of low general relevance, while a document might be targeted to an audience with a particular field of interest. Potential errors in the recognition and linking of entity mentions to database entries further add to the evaluation inaccuracy.

4) Inspection of a Sample of the System Output Another evaluation possibility is to manually verify a sample of a system's intermediate or final output. This can give insights into the quality of system internals and can add an intrinsic aspect to the evaluation of an approach. For example, a RE system with explicit path representations in the shape of patterns might be intrinsically evaluated by sampling a number of these patterns and inspecting them to check whether they capture reasonable relational semantics. The downside of sampling and verifying system output is that pipeline errors need to be factored in, i.e., RE quality cannot be assessed independently from upstream components.

2.7.1 Common Metrics

Common metrics in the context of the evaluation of RE and EE are *precision*, *recall*, and *F measure*, which were proposed first for the assessment of information-retrieval systems. For a set of documents D from which relations and events are to be extracted, the existence of an annotation of D is assumed. This annotation lists the actual mentioned instances G of relations and events. Given the set of recognized instances O by a system whose extraction quality is to be assessed, the metrics are defined as shown in Equations 2.1–2.3:

Equation 2.1
$$precision = \frac{|G \cap O|}{|O|}$$

Equation 2.2
$$recall = \frac{|G \cap O|}{|G|}$$

Equation 2.3
$$F_\beta\ measure = \frac{\left(\beta^2 + 1\right) \cdot precision \cdot recall}{\beta^2 \cdot precision + recall}$$

While precision states which fraction of extracted mentions is correct, recall grasps how many of the ground-truth facts have been recognized by the system (Chinchor and Dungca, 1995; Grishman and Sundheim, 1995; 1996). F measure is then a way to report a single value for the performance of a system, which incorporates both the extraction accuracy and the coverage (Chinchor, 1998b). β is a parameter to bias the F-measure metric towards one of the two inputs; it is typically set to 1 and then called *F1 measure/score*. Further metrics from information retrieval are used in parts of literature, examples include mean-average precision and mean-reciprocal rank (both taking into account an ordering of extracted facts with respect to system confidence), and derived metrics like the browsing cost (already mentioned above). Baeza-Yates and Ribeiro-Neto (2011, pp. 134–158) provides a detailed introduction into typical retrieval metrics.

From the wide range of employed metrics for the evaluation of (within-document) co-reference resolution of event mentions, the BLANC metric (BiLateral Assessment of Noun-phrase Coreference; Recasens and Hovy, 2011) is a recently proposed example. It balances the impact of positive and negative event-mention links in a document, which is a crucial step since in many cases negative links and consequently singleton event mentions are more common than their counterparts. As Recasens and Hovy (2011) pointed out, the informativeness of other metrics like MUC (Vilain et al., 1995), B^3 (Bagga and Baldwin, 1998), and the naive positive-link metric suffers from such an imbalance. These clustering metrics are of particular importance for the approach to inter-sentential extraction of relations and events, which is discussed later in Chapter 7.

2.7.2 Shared Tasks and Established Datasets

Community-wide evaluation programs conducted in the past two and a half decades made concerted effort to develop meaningful testbeds for competing IE approaches. In the following, these efforts are briefly summarized. For other introductions to the history of IE research see, e.g., Nastase et al. (2013), Thomas (2015), and Akbik (2016).

MUC Early datasets for IE evaluation were produced in context of the Message Understanding Conference (MUC) series (MUC, 1991; 1992; 1993; 1995; 1998), a venue intended to start and support research on text analysis with military applications (Grishman and Sundheim, 1995; 1996). MUC datasets follow an evaluation scheme where the ground truth is given as document-specific *answer keys*, i.e., for each document the correct and actually mentioned fact instances are listed, but no mention-level information pin-pointing the fact reference is given. This is an example of a more application-oriented evaluation style, which is contrary to later evaluation programs that focus on the intrinsic evaluation of systems. The central task of the MUC competitions was called *scenario template filling*, which involved the extraction of information for pre-specified types of events. Various aspects of the recognized events needed to be distinguished from one another and were to be sorted into *slots*, e.g., the type of event, its participants and arguments, the location and time, etc.

Over the course of MUC cycles, the templates to fill became increasingly complex, with MUC 5 even introducing nested template structure. Early iterations of the competition provided the participants with military messages, later datasets then shifted the focus to terrorist events in Central/South America, and international joint ventures (texts covering business activities of organizations/companies) and electronic circuit fabrication (advances in processing techniques). Approaches developed for earlier task iterations (before MUC 6) raised concerns that the achieved progress would be limited to the specific tasks and that no actual progress was made on the fundamental building blocks to automatic semantic understanding (Grishman and Sundheim, 1995; 1996). This is why several sub-tasks were added to MUC, which were considered to be elementary to generic language understanding systems, i.e., NER and CR. The final iteration of the MUC competition featured texts from which instances of airplane crashes were to be extracted (Chinchor, 1998c). Furthermore, this final instantiation introduced the notion of binary RE into the shared task, by asking participants to automatically recognize mentions of the relations *employee-of*, *product-of*, and *location-of*. Many of the MUC datasets are publicly available, e.g., the ones from MUC 6 and MUC 7 via the Linguistic Data Consortium (Chinchor and Sundheim, 1996; 2003; Chinchor, 2001).

ACE Automatic Content Extraction (ACE) was a follow-up research program, which also featured organized evaluations to foster progress on IE technologies. It took place in the 2000s (Doddington et al., 2004). The objective of ACE was in line with the goals of MUC and the separation of IE into sub-tasks with independent evaluations was continued. These sub-tasks dealt with the recognition of entities and the resolution of

their references, as well as with the recognition of relations between them. For the first few evaluations, only the extraction of facts with exactly two participants was done (called *relation detection and characterization* in ACE terminology), later however, a similar task as scenario-template filling was included, which allowed to connect more than just pairs of arguments (called *event detection and recognition*). A conceptual difference between MUC and ACE is that the latter puts emphasis on real-world entities and their relations, rather than on their representation in texts, i.e., compared to MUC, the task got slightly more abstract and co-reference resolution became more important.

Out of the datasets created during ACE's ten-year existence, the most influential publicly available ones (A. Mitchell et al., 2004; 2005; Walker et al., 2006) correspond to the 2003–2005 iterations of the shared task, due to the covering of fact extraction from English language texts (NIST, 2003; 2004; 2005), which belongs to the most popular facets of IE research. A major difference of these datasets, compared to the ones from MUC, is the considerably bigger set of targeted relation and event types. For example, the dataset from the 2005 iteration of the program contains annotation for six high-level relations covering physical relationships as well as business and personal ones, further divided into 18 fine-grained relations (Linguistic Data Consortium, 2005c). Regarding events, there are eight high-level ones and 33 sub-types, including events about the personal life of people, as well as judicial events, business ones, and more (Linguistic Data Consortium, 2005b).

Successors of MUC and ACE The ACE program ended in the late 2000s and was superseded by the Text Analysis Conference, a series of annual workshops hosting various shared tasks in NLP. Starting with the 2009 cycle, the Knowledge Base Population (KBP) track provided a testing ground for various problems related to the extraction of relational information from text with the purpose of building factual KBs. Of particular relevance to this thesis are two sub-tasks:

- *Slot filling*: For pre-defined entities, values for particular properties need to be extracted from texts.
- The *event* sub-task: This gathers various problems around the recognition and linking (resolution) of events in texts.

While slot filling is very similar to what is called relation detection in the ACE datasets (recognition of fact mentions with two arguments), the event task closely resembles the event recognition problem in the same series of datasets, meaning these TAC KBP tracks can be seen as a successor to the ACE program. The types (and subtypes) of events

covered in the event tracks of TAC KBP closely resembled the ones specified in the guidelines for ACE (Aguilar et al., 2014; Song et al., 2015).

Datasets for semantic relation classification tasks were also provided as part of the SemEval workshop series. The 2007 iteration (Girju et al., 2007) contained a binary classification task, where a sentence with marked argument candidates was given, and a system was to decide whether a given semantic relation was mentioned between these arguments or not. The later 2010 iteration (Hendrickx et al., 2010) changed focus slightly by asking whether a given sentence with arguments mentioned one out of several possible relations.

2.7.3 Performance of Humans and State-of-the-Art Approaches

The evaluation of automatic approaches for NLP tasks is complicated by the ambiguity of language, which often enough might be a challenge even for human experts. For the case of RE/EE, this means that human annotators often disagree on the question whether or not a fact is mentioned in a document and what the exact properties of this fact mention are. This issue is amplified by the intricacy of providing concise, yet comprehensive, annotation guidelines on what constitutes a mention of a particular fact type. For many datasets, this results in surprisingly low agreement between human annotators. Below, works that underline this discrepancy are recorded:

- Sundheim (1995) reported approximately 83% agreement between annotators for the MUC 6 task on scenario-template filling, estimated by treating one human expert's annotation as gold-standard and evaluating a second human expert's annotation against it.
- Doddington et al. (2004) stated that agreement for the early relation-annotation process in the ACE program varied between 35 and 52 "overall value score". This agreement was determined as part of the standard ACE multi-step annotation process, which involves several annotators independently annotating the same texts, followed by an adjudication step. According to Doddington et al., annotators had difficulties in differentiating relation types and in consistently distinguishing between implicit and explicit mentions of relations.
- Hendrickx et al. (2010) mentioned that inter-annotator agreements for the SemEval 2010 task on relation extraction (task 8) were between 60% and 95%, depending on the semantic relation.

- Min and Grishman (2012) discussed the TAC shared task on slot filling and the associated resources. They argued that the creation of a comprehensive gold-standard annotation for a collection of documents as large as the one being used in this shared task is close to impossible. Consequently, the ground-truth facts in this evaluation are only estimated by pooling the system outputs of participants, followed by the manual assessment of the correctness. Additionally, a time-limited manual annotation round was performed, simulating an automatic extraction process, the results of which then augmented the pool of system outputs. The annotators reached a performance (with respect to the metrics applied to system output) of 81.4% F1 / 68.5% F1 in TAC KBP 2012 / 2013 (Surdeanu, 2013) and 70% F1 in 2014 (Ellis et al., 2014).

- Song et al. (2016) commented on the annotation process for the TAC KBP event track. The datasets were annotated in full (in contrast to the ones for the slot filling task, see above) by several annotators in parallel. For the 2015 cycle, they stated that the agreement on the detection of events was close to 60%, while for the linking of events 67.63% were recorded. According to Song et al., the complexity of the annotation task can partly be attributed for the low agreement. For instance, annotators often did tag the same event mention but assigned different type labels. An example they provided is that the trafficking of organs can be seen both as a transport action (moving an object between locations) and a transaction (transferring the ownership of an object), and that annotators indeed made conflicting decisions on this.

The apparent difficulty of RE, EE, and closely related tasks even for humans sets a low expectancy on the performance levels that automatic approaches achieve. Indeed the top-performance of participants in shared tasks is typically well below levels acceptable for real-world applications. For the two target templates in the MUC 5 task on scenario-template filling, the top performing systems achieved only 52.75% and 49.18% F1 score (MUC, 1993, Appendix B: MUC-5 Test Scores). The follow-up iteration of this task at MUC 6 was won by a system with 56.4% F1 score (Sundheim, 1995). The top score obtained at MUC 7 in the RE sub-task was 75.63% F1; for the scenario-template filling task it was 50.79% F1 (Chinchor, 1998a). Compared to the MUC tasks, the problem at hand in Task 8 of SemEval 2010 was considerably simpler, i.e., participants were only required to classify the semantic relation of given noun pairs. This is reflected in the high performance score of the top-ranked system, which achieved 82.19% in the F1 metric (Hendrickx et al., 2010). The slot-filling formulation of the RE problem, however,

seems much harder, as the best performing system in the 2013 cycle at TAC scored only 37.28% F1 (Surdeanu, 2013). The winner of the follow-up task in 2014 reached 36.77% F1 (Surdeanu and Ji, 2014).

Despite the fact that these numbers are not directly comparable due to the use of different datasets and varying task definitions, they nevertheless convey an intuitive understanding of the quality level at which RE/EE systems operate today. Note that a performance of 75% F1 score corresponds to various trade-off combinations of precision and recall. For a text with 100 ground-truth mentions, it can mean that 75 of these have been found at the cost of 25 false positive extractions. Alternatively, it can also indicate that there were no false positive reports, but only 60 of the true mentions have been identified; among additional possible combinations. Furthermore, it is noteworthy that the top-system performance at the tasks is still way below human performance. For the case of slot filling at TAC KBP 2013, the top system performed at "barely" 50% of human performance (Surdeanu, 2013), and for the 2014 iteration the respective top participant reached approximately 52% of human performance (Surdeanu and Ji, 2014). Because of this gap, IE is generally considered to be a hard problem, with a lot of headroom for performance improvements in shared tasks.[33]

[33] There are more aspects of the evaluation of IE systems that are worthy of discussion, which, however, we do not review in detail for space reasons. One such aspect is the question of how human-level performance is defined. In particular for the task of IE, the background and proficiency of human annotators are rarely reported. Annotators can have a wide range of qualifications, some may be trained linguists and others may be people whose only trait is to have a particular native language. The former is particularly common in the context of shared tasks, while the latter became popular only since the advent of crowd-sourcing (Callison-Burch and Dredze, 2010). Furthermore, relevant qualifications are not limited to language-understanding capabilities but can also involve application-specific analytical skills. This needs to be taken into account when interpreting ML performance and its gap to human performance. For example, the recent development of the AlphaGo system (Silver et al., 2016) was celebrated as a breakthrough in AI, as for the first time a computer program was capable of beating the best human players in the ancient Chinese game Go. Clearly, when playing Go, the metric with which we measure a system's success should be if it can keep up with professional players. But this event also raises the question if linguists really are the right people to annotate IE datasets: If IE technology is to assist a huge part of the population, would it not be better to compare against the performance of someone with an average educational background? Consider the case of business intelligence, an area with many application opportunities for IE. Here, IE systems can save business people valuable time by summarizing articles and automatically searching the web for particular information. Should we not assess new IE methods with respect to these people's reading-comprehension capabilities and the time they save with the help of IE?

2.8 Summary

This chapter started with the introduction of a wide range of classic and more recent approaches for IE and discussed the open problems of high precision extraction as well as extraction of document-level information. Various kinds of KBs that benefit from IE technology or which themselves facilitate improved IE methods were presented. At last, an overview of typical performance at IE shared tasks was given. This final section of the chapter points out how the research presented in the remainder of the thesis relates to the discussed prior work.

IE Systems Are Pipelines As described in Section 2.2, it is an established practice to separate the handling of entity references in texts from the recognition of potential semantic relations between them. Furthermore, several layers of linguistic information are typically assumed to be present. The methods proposed in Chapters 3–7 also implement this idea and apply standard tools for the tokenization of words, the segmentation of texts into sentences, for syntactic parsing, and for the handling of entities. Only after these steps are completed, RE/EE occurs.

Methods for the Collection of Linguistic Patterns Produce Low-Quality Results Section 2.3 listed approaches which automatically extract patterns from large text collections. For many application domains, the fully supervised setting is infeasible due to the high upfront cost for annotation. At the same time, methods with weaker forms of supervision inherently produce noisy extractions. Despite existing means for the automatic filtering of patterns, produced pattern sets generally lack both accuracy and coverage, as indicated by the reported low RE performance in studies (Section 2.7). Another shortcoming of current approaches is that pattern generation from labeled examples is often limited to the shortest-path between relation arguments in syntactic graphs. This can result in the production of patterns which miss out important semantic indicators. Chapter 4 presents techniques that address the quality issues of current IE systems. Based on the RE system discussed in Chapter 3, it is illustrated how patterns can be filtered more accurately with the help of external lexical-semantic knowledge repositories. Furthermore, means are discussed for improved construction of patterns from syntactic parses.

Disambiguation of Open-Domain Patterns Is Challenging In order to address application needs with emergent relations, we turn to open-IE methodology in Chapter 5. Current open-IE methods suffer from a lack of high-quality disambiguation of discovered linguistic patterns. The open character of the pattern-collection process makes it a necessity to process large amounts of data in order to harvest information about the semantic relatedness of individual patterns. Chapter 5 presents a weak-supervision signal which is useful in true open-domain scenarios and proposes a way to exploit this signal for the automatic learning and grouping of a vast number of relation phrases.

Approaches for Discourse-Level RE Are Too Domain-Dependent Section 2.4 explained that much relational information contained in texts occurs in an inter-sentential fashion, hence discourse-level methods for RE/EE are needed. A particularly plausible choice of approaching the extraction of discourse-level mentions is via event-mention linking, as systems for this task naturally integrate into IE-pipeline approaches and can, e.g., benefit from future quality-wise advances in upstream components. A problematic aspect of existing efforts for event linking is the widespread use of features that rely on external knowledge bases with specific target domains and limited general-domain validity. Furthermore, for many domains, such knowledge is not readily available. This establishes a need for the development of domain-independent approaches which base their decision about co-reference resolution for facts solely on the document content. Chapter 7 discusses work in this direction.

There Is a Missing Link Between Linguistic and Factual Knowledge Resources Among the various types of resources to relevance for IE applications (Section 2.5), two types stand out: databases with factual information and repositories with linguistic knowledge. While factual knowledge bases offer very useful information about instances of relations and events, including their mutual interactions, they do not directly offer linguistic expressions that would embed the facts in natural language. Linguistic resources, on the other hand, do provide information about language expressions, but miss the direct links to factual knowledge and often do not cover the semantic domains relevant for fact types. Later in Chapter 6, we present work on the problem of transforming a set of automatically constructed patterns into a widely useful linguistic resource, which also contains links on a fine-grained level to other resources. This new resource links semantic relations from factual knowledge bases to the language expressions useful to refer to instances of these relations.

Neural Networks Allow to Build Scalable Yet Expressive Models For Language Tasks Section 2.6 covered recent developments in the area of DL and NNs. Many architectures have proven their usefulness for IE applications. Simple methods for learning word embeddings have remarkable capabilities for leveraging distributional similarities of words, i.e., the semantics of words is induced from their usage patterns in large corpora. The approach for large-scale pattern clustering in Chapter 5 builds on this idea by leveraging standard feed-forward neural networks. In contrast to these very efficient, yet simple neural models, more complex ones allow to learn features occurring in textual snippets of varying lengths. CNNs are a useful instance of this class, which are particularly good at capturing local (event) information in longer sentences. We employ such networks in Chapter 7 for event linking.

No One Evaluation Paradigm Fits All Settings A wide spectrum of evaluation methodologies exist for IE (Section 2.7). Some analysis methods are based on comparing against a mention-level gold-standard, others resort to distant evaluation schemes where the accuracy of extraction methods is only estimated. While many shared tasks produced gold-standard data like the ACE 2005 corpus, unfortunately the domains covered by these datasets are quite specific. In the following chapters, we employ a combination of the evaluation approaches introduced in Section 2.7. In Chapters 3 & 4, we make use of a specifically created dataset with mention-level annotations as well as of a distant evaluation approach using the factual knowledge base Freebase. In Chapter 5, we resort to a manual evaluation of the system output with human raters. A publicly available corpus is used for evaluation in Chapter 7.

Chapter 3

Distantly Supervised Pattern Discovery from the Web[34]

Contents

[34] Chapter 3 is based on a conference paper (Krause et al., 2012) and a journal article (Krause et al., 2016a). In context of this PhD thesis, the present chapter serves to motivate the original follow-up work in Chapter 4. The underlying methodology of the Web-DARE system is shared with an approach described in a *Diplom* thesis (Krause, 2012); part of the error analysis from that work is reproduced in Section 3.7 as it inspires the improvements to RE presented in Chapter 4. This chapter presents results from joint work with Hong Li, Hans Uszkoreit, and Feiyu Xu.

3.1 Introduction

The next two chapters cover this thesis' work on large-scale pattern discovery with pre-defined relational schemas. While Chapter 3 introduces the Web-DARE system for pattern collection from the web and presents initial filtering means, Chapter 4 reports on two original ways of improving the pattern quality of this base system.

In recent years, distant supervision has become an important training paradigm for data-driven RE (e.g., T. V. T. Nguyen and Moschitti, 2011a; Takamatsu et al., 2012; Angeli et al., 2014; Roller et al., 2015) because of the availability of large knowledge bases (Section 2.5), which in DS are utilized to automatically label mentions of facts in unannotated text corpora as training data. In the Web-DARE system, we employ facts from Freebase (Section 2.5) as seed knowledge for the automatic collection of RE rules from the web. For these rules, we adopt the formalism of the DARE framework (F. Xu et al., 2007, Subsection 2.3.3), which can accommodate relations with more than two arguments and which is expressive enough to incorporate structures from dependency grammar. When applied to parsed sentences, the learned rules can detect relation mentions and associate the arguments with their respective roles. In contrast to statistical-classifier approaches (e.g., Mintz et al., 2009; T. V. T. Nguyen and Moschitti, 2011a), which implicitly encode rules, Web-DARE also delivers the extraction rules themselves as an important linguistic knowledge source. These rules can be utilized for applications such as question answering and textual entailment. Furthermore, we created a particular type of linguistic resource from them that will be presented in Chapter 6.

The experiments we report in this chapter demonstrate that using the web as a source for extraction patterns is particularly useful regarding covering linguistic variation. However, precision is adversely affected by a large number of invalid rules collected alongside proper ones. We address this issue in a first step by implementing a filter based on the mutual exclusiveness of some relations. This technique is a variant of previously proposed methods, called *counter training* (Yangarber, 2003; Etzioni et al., 2005) and *coupled learning* (Carlson et al., 2009). However, our adaptation is better suited for DS as it works directly on rule sets without the need for confidence feedback from extracted instances. While this filter partially remedies the low precision, overall performance does have room for improvement. Therefore, we conducted an analysis of erroneous patterns passing the proposed filter, which motivates the improved pattern generation and filtering we discuss in Chapter 4.

3.2 Targeted Semantic Relations

The underlying approach of Web-DARE is relation-independent and works for various semantic domains. To allow a more vivid description of Web-DARE in the following sections, we already introduce the relations from the three domains on which experiments were conducted at this point. All three domains (*award*, *business*, and *people*) contain n-ary relations with $n \geq 2$. Let t be an entity type and let $\mathcal{E}_t$ be the set containing *all* named entities of type t. Let T be a bag of entity types and let $n = |T|$. Then any of the n-ary target relations is a set $\mathcal{R}$ for some T with

Equation 3.1

$$\mathcal{R} \subseteq \prod_{t \in T} \mathcal{E}_t$$

For example, the *marriage* relation can formally be described as:

Example 3.1

$$\mathcal{R}_{marriage} \subseteq \mathcal{E}_{\text{person}} \times \mathcal{E}_{\text{person}} \times \mathcal{E}_{\text{location}} \times \mathcal{E}_{\text{date}} \times \mathcal{E}_{\text{date}}$$

Often a subset of k ($k \geq 2$) arguments of a relation are *essential arguments*, meaning that conceptually the relation holds between these entities, while the $(n - k)$ other arguments provide supplementary information. If such a subset of essential arguments is defined, the presence of at least these arguments is required in every text mention of this relation. For the *marriage* relation from the domain *people*, we require both persons to be mentioned, whereas date and location of the wedding are considered optional. The divorce date, if applicable, is also considered optional.

Table 3.1 (p. 60) lists the targeted semantic relations in this work, along with their entity-type signature, grouped by solid horizontal lines with respect to their domain. All relations from a domain share the entity type of the first essential argument. If two of the relations share the entity type of another essential argument, we consider these as being of the same *essential type*. Relations that have the same essential type are, e.g., *acquisition*, *foundation*, and *organization relationship*, since their first two arguments are of the same entity type (organization). All relation definitions in Table 3.1 were derived from Freebase.

Relation	Slot 1	Slot 2	Slot 3	Slot 4	Slot 5
award nomination	⊛ prize	⊛ org/per	date	—	—
award honor	⊛ prize	⊛ org/per	date	—	—
country of nationality	⊛ per	⊛ loc	—	—	—
education	⊛ per	⊛ org	degree	area	date
marriage	⊛ per (SPOUSE1)	⊛ per (SPOUSE2)	loc (CEREMONY)	date (FROM)	date (TO)
person alternate name	⊛ per	⊛ per	—	—	—
person birth	⊛ per	(⊛) loc	(⊛) date	—	—
person death	⊛ per	(⊛) loc	(⊛) date	(⊛) cause	—
person parent	⊛ per (PERSON)	⊛ per (PARENTA)	per (PARENTB)	—	—
person religion	⊛ per	⊛ religion	—	—	—
place lived	⊛ per	⊛ loc	date (FROM)	date (TO)	—
sibling relationship	⊛ per	⊛ per	—	—	—
acquisition	⊛ org (BUYER)	⊛ org (ACQUIRED)	org (SELLER)	date	—
business operation	⊛ org	⊛ business space	—	—	—
company end	⊛ org	(⊛) date	(⊛) termination type	—	—
company product relationship	⊛ org	⊛ product	date (FROM)	date (TO)	—
employment tenure	⊛ org	⊛ per	position	date (FROM)	date (TO)
foundation	⊛ org (ORG)	⊛ org/per (FOUNDER)	loc	date	—
headquarters	⊛ org	⊛ loc	—	—	—
organization alternate name	⊛ org	⊛ org	—	—	—
organization leadership	⊛ org	⊛ per	position	date (FROM)	date (TO)
organization membership	⊛ org (ORG)	⊛ loc/org/per (MEMBER)	date (FROM)	date (TO)	—
organization relationship	⊛ org (PARENT)	⊛ org (CHILD)	date (FROM)	date (TO)	—
organization type	⊛ org	⊛ org type	—	—	—
sponsorship	⊛ org (SPONSOR)	⊛ org/per (RECIPIENT)	date (FROM)	date (TO)	—

Table 3.1 – Definition of the 25 target relations from the domains *award*, *people*, and *business*. For each relation, the entity types of its two to five semantic arguments are listed. ⊛ denotes the essential arguments of the relation, i.e., the core part of a relation instance defining its identity. (⊛) marks alternatives for essential arguments. loc/org/per are short for location/organization/person. Labels for argument roles (in SMALLCAPS) are only given in ambiguous cases.

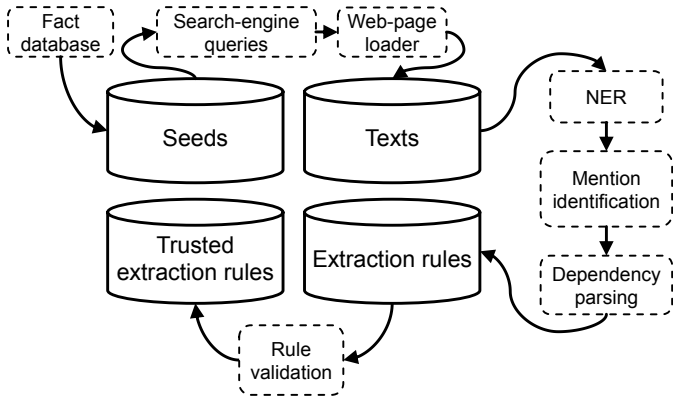


Figure 3.1 – Overview of data flow in Web-DARE.

3.3 Method Description

Figure 3.1 outlines the process of learning relation-extraction patterns in a distantly-supervised, language- and relation-independent way. Given a target relation r, a set of seed instances $\mathcal{I}_r$ of this relation, and a language l, we can create a set of rules $\mathcal{P}_{r,l}$ with the following procedure:

a) Acquire a set of textual mentions $\mathcal{M}_{r,l}$ of instances i for all $i \in \mathcal{I}_r$ from a text corpus.

b) Extract candidate rules $C_{r,l}$ from the dependency trees of elements of $\mathcal{M}_{r,l}$.

c) Validate the rules $p \in C_{r,l}$, yielding a derived set $\mathcal{P}_{r,l}$ of acceptable dependency-construction based rules.

We discuss each of these steps in more detail in the following.

a) Textual Mention Acquisition and Preprocessing The first step in the processing pipeline is to collect a large number of textual mentions of a given target relation, ideally covering many different linguistic constructions used to express the relation. Following F. Xu et al. (2007) and F. Xu (2007), we collect textual mentions using a set of seed instances $\mathcal{I}_r$ of the target relation r as the input. Every sentence which contains the entity tuples of the seed instances is regarded as a textual mention of the relation. Similar to standard distantly-supervised approaches, this seed instance set can be easily obtained from an existing knowledge base.

The seeds are used as queries for a web search engine to find documents that potentially contain mentions of the seeds. We construct a separate query for each seed by concatenating the full names of all argument fillers. Documents returned by the search engine are downloaded and converted into plain text, using standard methods for HTML-to-text conversion and boilerplate detection. Using off-the-shelf tools, we then perform standard NLP preprocessing of the text documents, including sentence detection, tokenization, NER, lemmatization, POS tagging. We also link entity mentions to seed entities with a simple dictionary-based linking strategy that matches name variations of the seed's entities as provided by the KB. We discard all sentences not mentioning a seed instance, as well as sentences that do not express all essential arguments of the relation. Then, the remaining sentences are processed by a dependency parser outputting the dependency relations of de Marneffe and Manning (2008). We use the output of the NER tagger to generalize the dependency parse by replacing all entity mentions with their respective entity tags.

b) Candidate Rule Collection The next step of the pipeline process is to extract candidate dependency-structures from the parse trees of the source sentences. Typically, shortest path or minimum-spanning-tree algorithms are used to select the sub-graph of the dependency tree that connects all the arguments mentioned in the sentence (see F. Xu et al., 2007). In Chapter 4, we present an alternative, knowledge-driven algorithm which employs a large lexical-semantic repository to guide the extraction of dependency structures. The algorithm expands the structure to include semantically relevant material outside the minimal subtree containing the shortest paths. For the moment, we use the rule-learning component of the DARE system.

As an example of pattern extraction, consider the *marriage* relation from Table 3.1 (p. 60) with the arguments Spouse1, Spouse2, Ceremony, From, and To. Given the 5-ary seed in Figure 3.2 (p. 63), the sentence in Figure 3.3 (p. 63) can be used for rule learning. This sentence is processed by the dependency parser, which outputs the structure in Figure 3.5 (p. 64), where the surface strings are augmented with their respective types via NER. Note also the argument-role labels (in brown color), which are propagated from the roles of the entities in Figure 3.2. From the dependency tree, (Web-)DARE generates the rule in Figure 3.4 (p. 63), which contains five arguments: two married persons plus the wedding location and the starting and end date of the marriage. In addition, projections of this rule are produced. Such projected rules contain a subset of the arguments, they could, for example, only connect the person arguments. This way, a single sentence might result in several generated rules. (*Text continues on page 65.*)

⟨ Elizabeth Taylor , Larry Fortensky , 1991-10-06 , 1996-10-31 , Neverland Ranch ⟩

Elizabeth Taylor	Larry Fortensky	1991-10-06	1996-10-31	Neverland Ranch
SPOUSE1 (person)	SPOUSE2 (person)	FROM (date)	TO (date)	CEREMONY (location)

Figure 3.2 – Seed example of relation *marriage*. Under each entity, the respective argument role (in brown) and named-entity type (in blue) are listed.

person/SPOUSE2 person/SPOUSE1 date/FROM
Fortensky and Taylor were married on October 6, 1991,
person location/CEREMONY
at Michael Jackson's Neverland Ranch and divorced on
date/TO
October 31, 1996.

Figure 3.3 – Sentence with a mention of the seed in Figure 3.2. Marked in blue and brown is the NER annotation (i.e., named-entity types and occurrences of seed arguments) as produced by the employed linguistic preprocessing components.

Rule body ::
- head: (marry, VBN)
- prep:
  - head: (at, IN)
  - pobj: head [0] location
- nsubjpass:
  - head: [1] person
  - conj: head [2] person
- conj:
  - head: (divorce, VBD)
  - prep:
    - head: (on, IN)
    - pobj: head [3] date
- prep:
  - head: (on, IN)
  - pobj: head [4] date
- auxpass: head (be, VBD)

Output :: ⟨ [2] SPOUSE1, [1] SPOUSE2, [0] CEREMONY, [4] FROM, [3] TO ⟩

Figure 3.4 – Extraction rule learned from the dependency graph in Figure 3.5 (p. 64; bold-printed sub-graph). Each element in square brackets represents a (sub-)rule. The first line of a rule states the "head" of the rule, i.e., a node of a dependency graph with certain restrictions. The head of a rule is either specified by a lemma and POS tag (in green), or by a named-entity type (in blue). Magenta-colored numbers in squares denote for which argument of the target relation a matched node (the represented named entity) is extracted. Apart from the head, a rule consists of sub-rules, which are connected to the head with dependency relations (here, e.g., "prep" and "pobj"). Note that this is a slightly more compact representation of rules compared to the ones in the previous chapters (lexical nodes collapsed, roles of entities mentioned only below rule). Inventory of POS tags and dependencies corresponds to Figure 3.5.

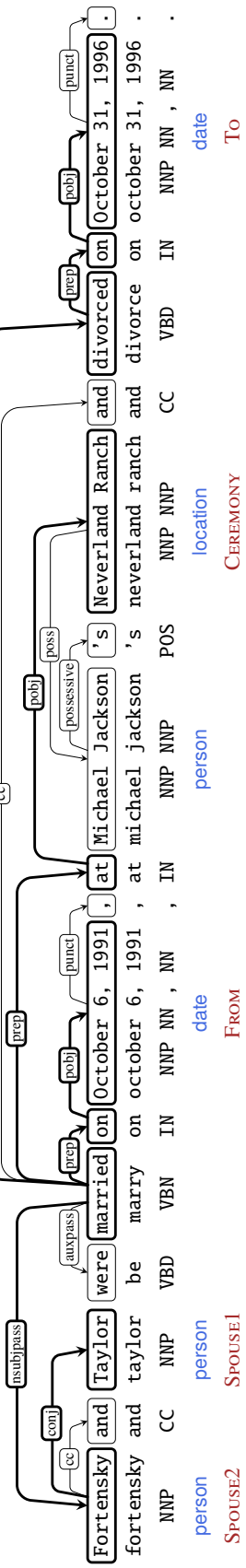


Figure 3.5 – Dependency graph for the sentence in Figure 3.3 as returned by MDParser, with additional mapping of NER results into the graph. Each box represents a token or entity mention of the sentence. Rows below the text state information on lemmas, part-of-speech tags (see Taylor et al., 2003), entity types, and semantic roles (in this order). Dependency-label set follows de Marneffe and Manning (2008).

(*Continues from page 62.*) An important design choice is the utilization of the dependency-relation formalism in the rule model. We assume that any given mention of a target-relation instance can be identified by a somehow characteristic pattern in the sentence's underlying dependency graph. This methodology is intuitively expressive enough for many mentions and it has been shown to perform well in general (F. Xu, 2007; Grishman, 2012). There are, however, limitations. For example, this approach does not cover mentions requiring some kind of semantic understanding (Section 3.7) and it does not find mentions with arguments spread across several sentences (Stevenson, 2006; Swampillai and Stevenson, 2010). The latter (the single-sentence restriction and the associated recall loss) will be discussed in Chapter 7.

c) Rule Filtering Because of the heuristic nature of the distant-supervision assumption (Subsection 2.3.4), inevitable errors in linguistic preprocessing, and potentially even false seed facts or false mentions, a large fraction of the generated (candidate) patterns constitute noise. We implement a method for quality estimation via a pattern filter that analyzes to what extent rules have been learned for more than one relation. Whenever two relations are of the same essential type, they may share a few relation instances. For example, the same two persons might be involved in relations such as marriage and romantic relations. This results in patterns which are collected for several relations at the same time. While some of these patterns can indeed express multiple meanings due to ambiguous language and entailment phenomena, most rules learned for two or more relations are not appropriate for at least one of the relations. Either the rule exhibits a much higher frequency for one of the relations, in which case it can be safely deleted from the other(s), or the rule is wrong for all relations.

We propose a general and parametrizable filtering strategy using information about the applicability of a rule with respect to other relations of the same essential type. If a rule occurs significantly more often in a relation r than in another relation r', this rule most probably belongs to r. Let $f_{p,r}$ be the frequency of candidate rule p in relation r (i.e., the number of sentences for r in which p has been observed) and let C_r be the set of learned candidate rules for r. Then the relative frequency of p in r is defined as:

Equation 3.2

$$rf_{p,r} = f_{p,r} \Big/ \sum_{p' \in C_r} f_{p',r}$$

Next, we define the first component of the filter. Let $\mathcal{R}$ be a set of relations of the same essential type. The candidate rule p is *valid* for the relation $r \in \mathcal{R}$ if the relative frequency

of p in r is higher than its relative frequencies for all other relations in $\mathcal{R}$:

Equation 3.3

$$valid^{r}_{inter}(p) = \begin{cases} true & \text{if } \forall r' \in \mathcal{R} \backslash \{r\} : rf_{p,r} > rf_{p,r'} \\ false & \text{otherwise} \end{cases}$$

The second component is a heuristic which only examines the frequency of a rule with respect to a single relation:

Equation 3.4

$$valid^{r}_{freq}(p) = \begin{cases} true & \text{if } f_{p,r} \geq x, \text{where } x \geq 1 \\ false & \text{otherwise} \end{cases}$$

With this filter component (Equation 3.4), we ensure that in addition to the relative frequency, there is also enough evidence that p belongs to r from an absolute point of view. We merge the two components into the final filter, later referred to as the *combined filter* or the *frequency-overlap (FO) filter*:

Equation 3.5

$$valid^{r}_{FO}(p) = valid^{r}_{freq}(p) \wedge valid^{r}_{inter}(p)$$

Before the filtering of patterns with the combined filter is conducted, all of the rules which do not contain any content words such as verbs, nouns or adjectives are dropped. In addition to the frequency heuristic, we also experimented with other simple features, such as the number of present arguments in a rule or the length of a rule's source sentences. The frequency filter proved to be the most useful.

3.4 Details and Results of Running the Pipeline

This section describes the application of the pattern-discovery pipeline on the 25 target relations listed in Table 3.1 (p. 60). Table 3.2 (p. 67) provides per-relation statistics for the system run, starting with the extraction of relation instances from Freebase (column "#seeds"), during which in total more than 200k items were collected. In the next step, we acquired a corpus of relation-mention examples, for which the instances from Freebase were transformed to search-engine queries and submitted to Bing[35]. We stopped querying Bing either once one million search results per relation had been retrieved or once all seeds from Freebase for this relation had been used. Overall, we observed that for relations of the domain *people* more search results were generated than for the domains *award* and *business*. A possible explanation is that fewer web pages are dealing with

[35] `http://www.bing.com`.

Relation	#seeds	#doc.	#sent.	#rules
award honor	11,013	50,680	16,651	10,522
award nomination	12,969	14,245	2,842	1,297
country of nationality	5,650	94,400	74,286	59,727
education	15,761	61,005	28,723	16,809
marriage	6,294	211,186	147,495	88,456
person alternate name	6,807	42,299	15,334	7,796
person birth	1,808	329,387	39,484	22,377
person death	1,437	241,447	38,775	31,559
person parent	3,447	148,598	58,541	45,093
person religion	8,281	48,902	39,439	37,086
place lived	5,259	89,682	57,840	48,158
sibling relationship	8,246	130,448	45,201	26,250
acquisition	1,768	40,541	30,116	26,986
business operation	12,607	51,718	31,274	15,376
company end	1,689	14,790	7,839	5,743
company product relationship	6,467	27,243	19,007	15,902
employment tenure	10,000	116,161	51,848	43,454
foundation	1,529	131,951	61,524	31,570
headquarters	1,987	79,731	33,255	23,690
organization alternate name	8,011	70,595	29,523	10,419
organization leadership	21,579	138,952	74,029	51,295
organization membership	4,180	50,061	32,646	29,220
organization relationship	70,946	37,475	17,167	12,014
organization type	4,625	3,939	1,391	843
sponsorship	1,513	11,009	5,395	4,599
average	9,355	89,458	38,385	26,650
sum	233,873	2,236,445	959,625	666,241

Table 3.2 – Statistics of pattern-discovery process. *#doc.* refers to the number of web documents in which a relation mention was found; no duplicate detection was performed. *#sent.* states the count of duplicate-free sentences with a relation mention. *#rules* corresponds to the number of unique dependency structures learned from these sentences.

award- and business-related topics than there are for biographic relations. It might also be the case that the Freebase facts from *award* and *business* are less prominent in *current* web pages and more of historical character.[36] This concurs with an, on average, greater absolute number of instances for *people* relations in Freebase, which is not surprising given that Freebase in part is based on Wikipedia for gathering knowledge.[37]

[36] The recency of web pages might be a factor influencing the presence and ranking of results in search engines.

[37] Freebase contains 3M topics and 20M facts for the domain *people*. That is a lot more than for the domain *business* (1M topics and 4M facts). Retrieved from `http://www.freebase.com/` on 2015-03-25.

The search results were subsequently processed by downloading the respective web page and extracting plain text from the HTML source code. This process suffered from various problems, leading to a fraction of "lost" documents up to 40% for some relations (e.g., *person death*). The loss of documents can be traced back to several different reasons such as problems with the access of web pages (timeouts etc.), erroneous plain-text extraction from HTML code, or empty web pages. After the creation of the text corpus, we ran entity-recognition components on it to find occurrences of named entities, in particular, those of the respective document's source seed. For the recognition of coarse-grained types (persons, organizations, locations, etc.), we employed the Stanford Named Entity Recognizer as part of the Stanford CoreNLP package (Finkel et al., 2005; Manning et al., 2014) and supplemented this with a regular-expression-based date recognizer. To identify the seed entities, we followed a simple gazetteer-based approach using the name variations of the seeds' entities as stated in Freebase.

Only a small fraction of the web addresses from Bing result in a plain-text document with a (supposed) relation mention. The actual number of successfully downloaded and useful documents per relation is given in the third column of Table 3.2 (p. 67). Non-English documents account for a significant fraction of the successfully downloaded but still unproductive documents. By far most documents fail since for at least one essential argument of the source seed no entity occurrence was found anywhere in the text: Although they were present in the query, the search engine returned results which did not cover all essential seed arguments. In addition, NER errors are likely to have contributed to this issue. The average number of mention-containing sentences per document is approximately 1.5. This is reasonable given the underlying assumption (Mintz et al., 2009) that any such sentence indeed expresses the target relation. After the identification of sentences that contain mentions, these sentences were processed by a dependency parser (*MaltParser*[38], *MDParser*[39]) outputting Stanford dependency relations, followed by the extraction of the minimum spanning tree containing all the seed's arguments present in any given sentence. These rules are also extracted for all argument subsets where at least two essential arguments occur in the sentence. In this sense, projections of the dependency tree which correspond to the full set of arguments are extracted as well. The final training corpus contains for each relation on average 38k distinct sentences with mentions of seed instances, i.e., a total of around 2.2M sentences. All of these mentions

[38] Release v1.7.2, engmalt-linear model v1.7, `http://www.maltparser.org/`.

[39] See `http://mdparser.sb.dfki.de/`. This parser is particularly fast, while maintaining competitive parsing quality when used in an application, as shown by Volokh and Neumann (2010) for the textual entailment task. The parsing results also contain information about part-of-speech tags and word lemmas.

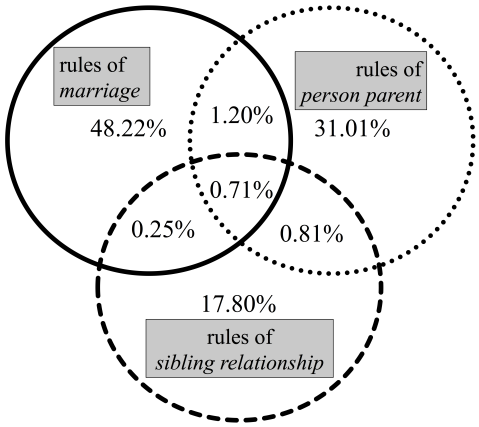


Figure 3.6 – Venn diagram illustrating the overlap of rules for three *people* relations. Rules in an intersection of two or three relations have textual evidence for all of those. 100% corresponds to the union of learned rules for all three relations.

include at least the essential arguments of the corresponding relation. On average, around 26k distinct rules were learned per relation (column "#rules" in Table 3.2, p. 67).

Figure 3.6 shows how the learned rules for three mutually-exclusive relations with the same essential type intersect. A small portion of rules exists at the intersection of two or three relations (ca. three percent); these rules are handled by the overlap component of the FO filter (Equation 3.3, p. 66). All of the remaining rules, which have only been observed for one relation, are addressed by the absolute-frequency filter component (Equation 3.4, p. 66). Even though only a small fraction of rules occurs in more than one relation, their handling has a significant impact on RE performance, as shown in the next section.

3.5 Evaluation of Generated Rules

Since we are particularly interested in recall and coverage of the collected rules, we require some form of mention-level annotation on gold-standard data and cannot rely on purely precision-driven evaluations as presented by, e.g., Mintz et al. (2009), where only the top 100 or 1000 extracted instances of a relation are manually verified. Unfortunately, existing datasets with relation annotation do not sufficiently cover the Freebase relations used here. For example, the popular ACE 2005 corpus (Walker et al., 2006) is too sparse for our evaluation since it only contains 14 mentions with the essential person arguments for the *marriage* relation. The annotation of another popular testbed for RE, the MUC-6 corpus (Grishman and Sundheim, 1996), is document-driven and does not provide direct

Arity	#Rules	Min. Freq.	Avg. Freq.	Med. Freq.	Max. Freq.
2	74,815	1	1.5	1	4,334
3	12,138	1	1.6	1	1,442
4	1,495	1	1.4	1	125
5	8	1	1.0	1	1

Table 3.3 – Distribution of *marriage* rules across arities. "Avg." – Average, "Med." – Median.

links between relation arguments and sentences. Therefore, we decided to prepare a new gold-standard test corpus annotated with relation mentions and their arguments on the sentence level. For the evaluation experiment in this section, we focus on only one relation.[40] We compare the web rules against patterns learned in a bootstrapping fashion with the basic DARE system. In order to investigate the impact of training-corpus size on the coverage in the distant-supervision approach, we also contrast the recall achieved by web rules and rules collected from local corpora of two different sizes.

Marriage Rules from the Web The *marriage* relation has five arguments (SPOUSE1, SPOUSE2, CEREMONY, FROM, and TO), of which at least the two person arguments are present in each learned rule. Table 3.3 presents frequency and arity (number of covered relational arguments) statistics for the *marriage* rules. Although the majority of rules are binary, more than 15% of the rules connect more than two arguments, which demonstrates the importance of handling $n > 2$-ary linguistic constructions to reach high coverage. A large fraction of rules was observed only once in the training corpus (more than 90%), which likely correlates with a high linguistic complexity of the underlying constructions, i.e., the more specific a linguistic construction is, the fewer it occurs in actual language use. Note that this provides additional context for the seemingly low overlap between relations in Figure 3.6 (p. 69), as the rules in the intersection of two or more relations are more likely to be found in application-time text, which makes it particularly important to get their relation assignment right.

Evaluation on Fully Annotated Documents The created gold-standard corpus (called PEOPLE_{Test}) consists of 25,806 sentences from crawled news articles of an online yellow-press magazine[41] with 259 annotated mentions of *marriage*. We compare the web-

[40] However, a variant of this corpus covers two more relations and is introduced in Chapter 4.
[41] `http://www.people.com/`.

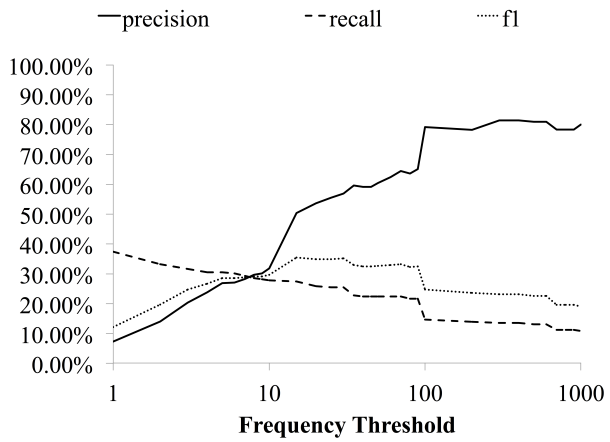


Figure 3.7 – Performance of web rules after filtering. X-axis: frequency thresholds.

based learning to a locally-bound bootstrapping approach which uses the same system components and the same seed relation instances. The learning corpus for bootstrapping (referred to as PEOPLE$_{Training}$) is of the same kind and size as PEOPLE$_{Test}$. Compared to the 88,456 *marriage* rules collected from the web, the bootstrapping system learned only 3,013 candidate rules. The rules from both systems are then applied to PEOPLE$_{Test}$ for evaluation, with the expected result that the web-based system achieves a significantly higher recall compared to the bootstrapping system: 49.42% vs. 30.5%.

Due to the noise in the set of (still unfiltered) web rules, the baseline precision is very low (3.05%), which illustrates the need for pattern filtering. Nevertheless, investigating recall at this stage is important because even an excellent rule filtering would produce below-average results if there are not enough correct rules to separate from wrong ones during the filtering phase. Figure 3.7 depicts the extraction performance after the combined filter $valid^{r}_{FO}$ is applied to the learned *marriage* rules. Precision improves considerably, in particular, it grows with increased frequency. The best F1 score can be obtained by setting the frequency to a threshold of 15, which results in a precision of approximately 50% and a recall of around 28%.

Evaluation with Different Corpus Sizes After the encouraging results on the small-sized PEOPLE$_{Test}$ corpus, we evaluated the rules by applying them to a larger dataset, the New York Times (NYT) subset of the English Gigaword 5 corpus (Parker et al., 2011). Due to the lack of a gold-standard annotation of *marriage* for this corpus, we used two alternative validation methods: (a) manual checking of all mentions detected by our

Corpus	#Docs	#Sentences	#Seeds w/ match	#Generated training sentences	#Rules learned
Web	873,468	81,507,603	5,993	147,495	88,456
LTW	411,032	13,812,110	1,382	2,826	1,508
PEOPLE$_{Training}$	150	17,100	76	204	302
NYT	1,962,178	77,589,138	–	–	–

a) Size of corpora and number of collected rules. Datasets above the dashed line are used for training, the one below for testing. Note that this table states the number of downloaded documents in the *marriage* part of the web corpus, in contrast to the number of such documents with a found mention of the source seed in Table 3.2 (p. 67).

			Mentions in sample		
Source of rules	**Filter applied**	**#Freebase**	**#correct**	**#wrong**	**Precision**
Web	–	1,003	76	1,747	4.17 %
LTW	–	721	47	414	10.20 %
PEOPLE$_{Training}$	–	186	7	65	9.72 %
Web	$valid^r_{inter}$	884	69	869	7,36 %
Web	$valid^r_{FO}$, with $x = 15$	627	52	65	44.44 %
Web	$valid^r_{FO}$, with $x = 30$	599	51	18	73.91 %

b) Extraction results on NYT dataset. "#Freebase" is short for "#Extracted instances confirmed as correct by Freebase."

Table 3.4 – Statistics from DS experiment with different corpus sizes.

rules in a random partition of NYT (100,000 sentences) and (b) automatic matching of extracted instances against the Freebase facts about *marriage*s. Note that before RE was performed, we removed all web-training sentences from NYT to avoid an overlap of training and test data.

The performance of the web rules is compared to rules learned on two local corpora in a distant-supervision fashion. The first corpus is the Los Angeles Times/Washington Post part of the Gigaword corpus (LTW). The second local corpus for rule learning is the one used for bootstrapping in the previous experiment (PEOPLE$_{Training}$). Here, only the rules learned in the first bootstrapping iteration were employed for relation extraction to allow for better comparison. For both local training corpora, the same seed set as for the web learning was used (i.e., 6,294 instances). Table 3.4 (p. 72) states statistics about the corpora and the learned rules (3.4a) and summarizes the extraction results of the different rule sets on NYT (3.4b). The web candidate rules without rule filtering find the highest number of positive *marriage* mentions of Freebase instances in the corpus, namely, 1,003. This experiment confirms the hypothesis that the extraction coverage of the learned rules increases with the size of the training corpus. After rule filtering, the web system improved the precision effectively without significantly hurting recall. Note that different kinds of rule filtering may also be applied to the rules learned from PEOPLE$_{Training}$ and LTW. Since the focus of this chapter is web learning, this work only considers the results for the web system.

3.6 Related Work

During the past 25 years, tremendous amounts of work were invested in the extraction of structured information from text, as delineated in Chapter 2. This section additionally highlights approaches which are particularly related to the Web-DARE system. We describe methods that either exploit a large number of facts as seed examples or process large corpora or internet documents.

Learning Hyponyms and Binary Relations from the Web Hearst (1992) presented an early attempt of gathering instances of the *is-a* relation from texts. Her endeavor was continued by Pantel et al. (2004), who provided a related approach for harvesting hyponyms from large corpora. They compared the results of a lexico-syntactic approach with a clustering-based approach that uses features from dependency parses of sentences. Their studies showed that a higher accuracy of the dependency-parsing approach coincides with a significantly higher runtime. From this observation, they concluded that for the

extraction on large corpora, a linguistically light-weight (shallow) approach is suited best. Consequently, related publications (Kozareva et al., 2008; Hovy et al., 2009; Kozareva and Hovy, 2010a) repeatedly employed lexico-syntactic patterns for learning hyponyms from the web and for the construction of taxonomies. Kozareva and Hovy (2010b) extended the learning of hyponyms to the learning of selectional restrictions for open IE patterns, i.e., the determination of the valid entity types of the relation arguments. This approach was further extended by Kozareva and Hovy (2011), who included the learning of temporal information about events from web texts. In the same vein, Ravichandran and Hovy (2002) presented an algorithm which extracted binary relations from the web using surface-level text patterns. An extension of this algorithm was embedded in the Espresso system by Pantel and Pennacchiotti (2006), who extended it with a ranking component that utilized search-engine queries to estimate the correctness of patterns.

Pasca et al. proposed in their work a bootstrapping-based approach to the extraction of instances of binary relations (Pasca et al., 2006a; b). Starting with only a few hand-crafted seed facts, they extracted about one million facts from a corpus of 100 million web pages by using a pattern-based methodology for fact discovery, relying only on shallow sentence features, namely surface-text pieces and part-of-speech tags. Targeting very large corpora, Pasca et al. state that this kind of light-weight processing is necessary. Jain and Pantel (2010) later extended the system with a graph-based ranking model and re-ran it on a corpus of 500 million web pages. Bunescu and Mooney (2007) proposed an approach in which a support vector machine is trained with text samples automatically retrieved from the web by querying a search engine with manually selected positive and negative instances of target relations. Augenstein et al. implemented an approach to distantly-supervised RE from web texts similar to ours, yet, without explicit patterns and only considering 40 binary relations for which they retrieved approximately one million web pages (Augenstein, 2014; Augenstein et al., 2014; 2016). A particular focus of their work was the improved recognition of entities in web texts, which can increase both precision and recall of extraction systems due to less noise in the weak training labels and at the same time a higher number of these.

The YAGO Ecosystem YAGO is a large ontology about entities and their relations extracted from Wikipedia (Suchanek et al., 2007; 2008; also Section 2.5). SOFIE (Suchanek et al., 2009) and PROSPERA (Nakashole et al., 2010) were two RE systems from the YAGO ecosystem, which implemented a pattern model based on lexico-syntactic expressions and which utilized YAGO in two ways. First, it was used as a source of a limited amount of initial training examples for the learning process and second it

was employed as trusted base knowledge for a reasoning component integrated into SOFIE and PROSPERA. This reasoning model utilized hand-crafted consistency rules to construct Horn clauses from extracted facts, which allowed the RE systems to treat confidence-estimation of extracted facts as a (weighted) MaxSat problem. Recently, YAGO, the extraction systems, and the reasoning components were extended to deal with the time- and space-dependent validity of facts, which led to the system TOB (Q. Zhang et al., 2008) and the ontologies T-YAGO (Wang et al., 2010) and YAGO2 (Hoffart et al., 2010; 2011). An even more recent extension to the YAGO ecosystem was described by Mahdisoltani et al. (2015), who processed online encyclopedias in multiple languages and combined this with the English WordNet to create an even larger ontology with 7 million additional facts compared to original YAGO. Yet another RE system developed in the context of YAGO was LEILA (Suchanek et al., 2006a; b), which used RE patterns based on deep linguistic analysis (i.e., Link Grammar) to extract binary relations.

Large-Scale Coupled IE on the Web The Never Ending Language Learner (NELL; T. M. Mitchell et al., 2015) was a system designed to learn factual knowledge from an immense text corpus over a long period. NELL's background ontology contained several hundred entity types (called categories) and binary relations. These were connected by subsumption links, some of them were additionally marked as mutually exclusive. This *coupling* of relations was helpful when the correctness of newly extracted facts was estimated. While earlier versions of NELL (Betteridge et al., 2009; Carlson et al., 2009) relied mainly on a learner of lexico-syntactic patterns, the system architecture was soon extended with an extractor working on semi-structured parts of web pages (HTML lists and tables, Carlson et al., 2010b) and a classifier for categorizing noun phrases into entity types based on morphological features and an inference-rule learning component (Carlson et al., 2010a; Lao et al., 2011). OntExt by Mohamed et al. (2011) was an extraction system using the output of the traditional-IE system NELL, particularly the collected category instances, to learn *new* target relations, i.e., patterns expressing these new relations and instances belonging to them.

Summary This section presented systems which process the web for RE training, either directly by accessing search engines or indirectly by processing large corpora downloaded from the web in advance. This trend of processing vast amounts of text documents addresses the problem of overfitting on individual corpora and fosters the portability of trained models to other domains. We implemented a similar methodology in Web-DARE by querying a search engine to retrieve locations of potentially useful web pages, which

Error class	Affected rules
Text extraction from HTML	22 %
DS assumption not holding	62 %
Annotation wrong	2 %
NER	8 %
Parsing	36 %
Mapping NER → parse	2 %
Shortest-path problem	24 %

a) Error classes in rule learning.

	Affected fn. in %	#
Total	100.00	131
Annotation error	4.58	6
Linguistic preprocessing error[42]	84.73	111
• NER error	41.22	54
• Parsing error	59.54	78
Total	100.00	125
Matching rule actually learned	50.40	63
No matching rule learned	27.20	34
Semantic understanding required	22.40	28

b) False-negative mentions.

Table 3.5 – Analysis of erroneous rules and false-negative mentions. The error classes leading to the learning of erroneous rules in Web-DARE are estimated on a sample of 50 patterns. Some rules are affected by more than one error source; hence values do not sum up to 100%. False-negative mentions (abbr. as fn.) are part of the gold-standard annotation on PEOPLE$_{test}$.

were subsequently downloaded and processed. Most approaches mentioned in this section rely on surface-level patterns and handle relations with exactly two arguments. In contrast, Web-DARE employs dependency-grammar analysis and can be applied to relations with a higher complexity.

The rule-filtering strategy of Web-DARE adapts the idea of coupling relations from the NELL system. Additionally, the frequency of occurrence of patterns is integrated as a simple filter feature. In Chapter 4, strategies for rule generation and confidence estimation are presented which operate with external semantic knowledge about target relations. The next section presents an error analysis of Web-DARE rules which motivates the need for such more involved strategies.

3.7 Analysis of Extraction Errors

We examined 50 erroneous *marriage* rules and assigned them to different categories; Table 3.5a presents the results. The largest error source were sentences which do not contain any information about the target relation, even though the arguments of an instance of this relation were mentioned. From this, we can conclude that further means are needed for the classification of a sentence as relation-relevant. Another high-impact error source was dependency parsing, as more than one-third of the rules contained parsing

[42] False-negative mentions can be affected by both error types.

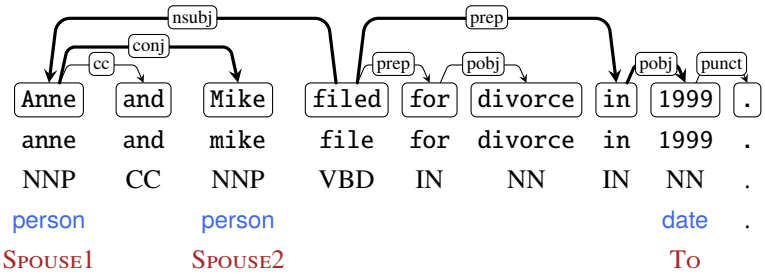


Figure 3.8 – Dependency graph illustrating the shortest-path problem. The shown parse was produced by the MDParser, with additional mapping of NER results into the graph. Each box represents a token or entity mention of the sentence. Rows below the text state information on lemmas, part-of-speech tags (see Taylor et al., 2003), entity types, and semantic roles (in this order). Dependency-label set follows (de Marneffe and Manning, 2008). Bold edges depict shortest paths between all three arguments.

errors, i.e., were learned from an erroneous dependency graph. Problems also arose from another assumption in the rule-learning algorithm, which is that the shortest path between two arguments in a dependency graph always covers the semantics of the target relation. In some instances, this assumption does not hold and important semantic indicators of the relation lay outside of the shortest paths in the parse tree. Figure 3.8 illustrates this issue. The sentence in this figure mentions three arguments of a *marriage* relation instance, the shortest paths connecting the arguments are highlighted by bold printing of the edges. Notice that the key semantic indicator of the target relation (`divorce`) is outside of the part of the parse that would be used to generate a rule. Improving the rule-learning algorithm in a way that it integrated relevant parts of the dependency graph in the context of the relation arguments would resolve this problem.

Section 3.5 states that the learned *marriage* rules covered 49.42% of the gold-standard mentions in PEOPLE$_{test}$. Table 3.5b (p. 76) shows the results of investigating the false negatives in this evaluation (the undetected mentions). Since Web-DARE operates on top of NER and parsing results, its performance heavily degrades when these preprocessing tools produce incorrect output. For 41.22% of false negatives, flawed NER rendered annotated mentions undetectable for extraction rules; example errors include unrecognized person entities and broken co-reference resolution. Parser errors account for 59.54% of false negatives; these are cases where errors on the paths between arguments render extraction impossible. To approximate the system recall in a setting with perfect linguistic preprocessing, we removed the mistakenly annotated mentions and fixed the errors in NER and parsing. We then reassessed whether a matching extraction rule had been learned in the training phase. For about half of the remaining false negatives, an extraction rule had actually been learned, meaning that the recall value stated in Section 3.5 would have been

about 25 percentage points higher if NER and parsing had worked perfectly. In other words, the upper bound for RE performance is defined by the unreliability of linguistic preprocessing rather than by a lack of coverage of the rules. An error class that cannot be attributed to accuracy deficits of linguistic processing are sentences which require a deeper level of semantic understanding. These are sentences where an instance of the *marriage* relation is only indirectly mentioned and cannot be determined unambiguously from the sentence syntax.

3.8 Summary

This chapter described the Web-DARE system for the extraction of *n*-ary relation instances from texts. By using the web as a training corpus, we achieved a recall improvement over DS and bootstrapping approaches that only work with local corpora. Low extraction precision was addressed by the introduction of a rule-filtering scheme which exploits the mutual exclusiveness of certain relations with the same essential type. If the error analysis had shown that recall cannot be further improved within the presented method, this would have meant that even using thousands of relation instances as seed does not suffice for collecting a high-coverage set of linguistic patterns. However, the error analysis indicates that the major impediment for recall is inaccurate NER and parsing. These tasks are beyond the scope of this thesis. Instead, in the following chapter, we focus on two of the other problems listed in Tables 3.5a and 3.5b (p. 76), namely a violated distant-supervision assumption and the shortest-path problem.

Chapter 4

Lexical Semantics for Enhanced Pattern Discovery[43]

Contents

[43] This chapter presents results from joint work with Hong Li, Andrea Moro, Roberto Navigli, Hans Uszkoreit, and Feiyu Xu.

4.1 Introduction

This chapter elaborates on two ways of incorporating lexical-semantic information for improved pattern-based RE: (a) enhanced filtering of patterns and (b) reformed generation of patterns.

Enhanced Filtering of Patterns Web-based systems for pattern discovery inevitably produce a certain amount of low-precision and low-recall extraction rules due to optimistic assumptions about their training procedure. Filtering by lexical features (e.g., POS information, word sequences), syntactic features, or simple manually-defined heuristics (Agichtein, 2006; Mintz et al., 2009; Carlson et al., 2010a) often does not suffice. A major open question is how semantic information about the target relation and beyond the seed data can be exploited. Several existing approaches add secondary semantic features to their systems, which, however, were shown to offer only slight improvements in RE precision (Jiang and Zhai, 2007; G. Zhou and M. Zhang, 2007).

In Section 4.2, we propose a method that automatically identifies relation-relevant parts in lexical-semantic resources, without the need for any task-specific manual annotation. The input of this unsupervised learning method are collections of noisy candidate linguistic patterns together with their sentence mentions. The method acquires relation-relevant word senses by applying WSD (word-sense disambiguation; Navigli, 2009) to the words in the patterns and then extracts the corresponding relation-specific sub-graphs from the lexical-semantic networks WordNet (Fellbaum, 1998) and BabelNet (Navigli and Ponzetto, 2012). Then, the acquired sub-graphs serve as semantic knowledge for identifying incorrect patterns which do not express the target relation. In contrast to frequency-based filters, such a relation-specific filter preserves low-frequency rules that are semantically relevant.

Reformed Pattern Generation Another issue of large-scale pattern discovery is the inadequacy of the shortest-path-based algorithm for pattern extraction from sentences, namely, the pattern-extraction algorithm fails at times to cover important semantic indicators. This has an adverse impact on both precision and recall in the downstream application of these patterns for RE. Section 4.3 explores how the elements of the relation-specific sub-graphs from Section 4.2 can help to improve the performance of pattern extraction and thereby enable the production more useful patterns. The graphs are used to annotate all occurrences of semantically relevant word senses in each parsed sentence. With this information, the pattern-discovery algorithm can extract from an

annotated parse all of the minimal trees containing every present argument entity and one or more semantically related terms. Furthermore, Section 4.3 also describes the creation of a dataset with mention-level RE labels for three relations, which is needed for a direct evaluation of the impact of the new extraction method.

4.2 Relation-Specific Sub-graphs for Pattern Filtering

In this section, we present a method for identifying relation-relevant parts in lexical-semantic networks and provide results for the two particular repositories WordNet and BabelNet (Section 2.5). Core concepts of the resources are sets of synonymous words (so-called synsets) and the relations that hold between these (hypernymy, meronymy, semantically-related form). Current statistical approaches to pattern filtering do not take into account all semantic clues available within patterns. As a consequence, they are not able to identify erroneous patterns which were extracted from sentences that mention arguments in a different context. Consider the following pattern[44]:

Example 4.1 $\textsf{person} \xleftarrow{\text{nsubj}} \texttt{met} \xrightarrow{\text{dobj}} \textsf{person}$

This pattern may be found with the help of a *marriage*-relation instance and may occasionally extract correct relation instances for *marriage*. However, it is not specific to this semantic relation. We tackle this issue by introducing a novel approach to represent the semantics of each relation, thereby excluding semantically irrelevant terms for *marriage* like `meet`. For a given semantic relation r, the input to our method is the set of linguistic patterns $\mathcal{P}_r$ for this relation, plus the set $\mathcal{S}_r$ of sentences from which the patterns originate. Our goal is to build the semantic sub-graph of the resource which corresponds to the target relation.

4.2.1 Algorithm

Algorithm 4.1 (p. 82) depicts the construction of sub-graphs specific to a given semantic relation. The first part of the algorithm computes a frequency distribution over the synsets $\mathcal{V}$ in a lexical-semantic KB G, given the set of sentences used to generate a set of linguistic patterns. For instance, the above pattern (Example 4.1) can be extracted from the sentence in Example 4.2 (p. 84). (*Text continues on page 84.*)

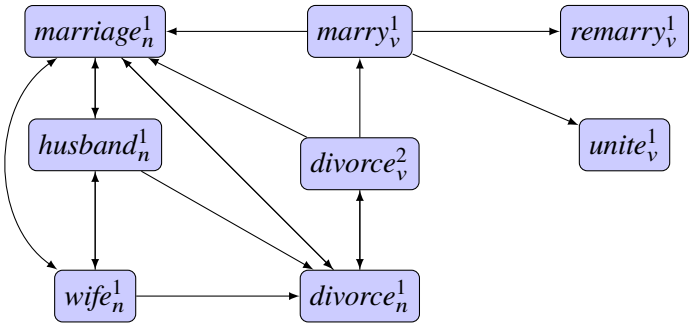
[44] From now on, we adopt a simplified visualization style of extraction rules and their underlying dependency patterns. In contrast to the attribute-value matrix representation, e.g., Figure 1.1 (p. 5), we depict patterns as graphs of lemmas and dependency edges and mostly do not display information on POS tags and argument roles. In accordance with the previous chapters, text/lemmas are typeset in `typewriter font`, and sans-serif corresponds to placeholders matching named entities.

function FINDSUBGRAPH($\mathcal{S}_r, \mathcal{P}_r, G = (\mathcal{V}, \mathcal{E}), n$)
  $\mathcal{V}_r \leftarrow \emptyset$
  $counts : \mathcal{V} \mapsto \mathbb{N}$ ▷ Counter for synsets. Values are initialized to 0.
  for $p \in \mathcal{P}_r$ **do**
    $\mathcal{S}_p \leftarrow \{s \in \mathcal{S}_r \mid p \text{ was generated from } s\}$
    for $s \in \mathcal{S}_p$ **do**
      for each content word $w \in p$ **do**
        $v \leftarrow WSD(w, s)$ ▷ Disambiguate word sense of w in sentential context.
        $\mathcal{V}_r \leftarrow \mathcal{V}_r \cup \{v\}$
        $counts(v) \leftarrow counts(v) + 1$
  for $v \in \mathcal{V}_r$ **do**
    if $counts(v) \leq 1$ **then**
      $\mathcal{V}_r \leftarrow \mathcal{V}_r \setminus \{v\}$
  find $\mathcal{V}_{r,n}$ **s.t.** ▷ Select most frequent synsets.
    - $|\mathcal{V}_{r,n}| = n$
    - $\forall v \in \mathcal{V}_{r,n} \;\; \forall v' \in \mathcal{V}_r \setminus \mathcal{V}_{r,n} : counts(v) \geq counts(v')$
  $\mathcal{V}_{r,n} \leftarrow \mathcal{V}_{r,n} \cup \{v \in \mathcal{V}_r \setminus \mathcal{V}_{r,n} \mid \exists v' \in \mathcal{V}_{r,n} : (v, v') \in \mathcal{E}\}$
  return $G_r := (\mathcal{V}_{r,n}, \{(v_1, v_2) \in \mathcal{E} \mid v_1, v_2 \in \mathcal{V}_{r,n}\})$

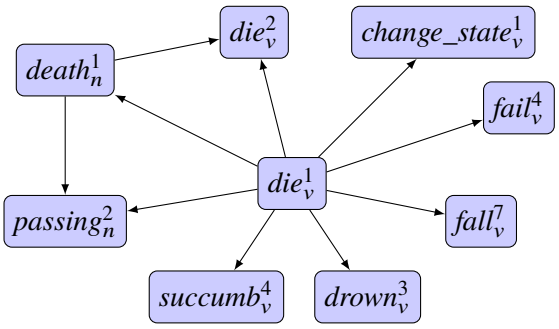
Algorithm 4.1 – Construction algorithm for a relation-specific sub-graph in a lexical-semantic resource $G = (\mathcal{V}, \mathcal{E})$, with $\mathcal{V}$ and $\mathcal{E}$ being the set of nodes (synsets) and edges in this resource. $\mathcal{S}_r$ is the set of sentences from which a set of linguistic patterns $\mathcal{P}_r$ for a semantic relation r were extracted. n is a threshold.

function FILTERPATTERNS($G_r = (\mathcal{V}, \mathcal{E}), \mathcal{P}_r$)
  $\mathcal{P}'_r \leftarrow \emptyset$
  $\mathcal{L} \leftarrow \{lex(v) \mid v \in \mathcal{V}\}$ ▷ Gather lexicalizations of synsets in sub-graph.
  for $p \in \mathcal{P}_r$ **do**
    for each content word $w \in p$ **do**
      if $w \in \mathcal{L}$ **then**
        $\mathcal{P}'_r \leftarrow \mathcal{P}'_r \cup \{p\}$
        continue with next pattern from $\mathcal{P}_r$
  return $\mathcal{P}'_r$;

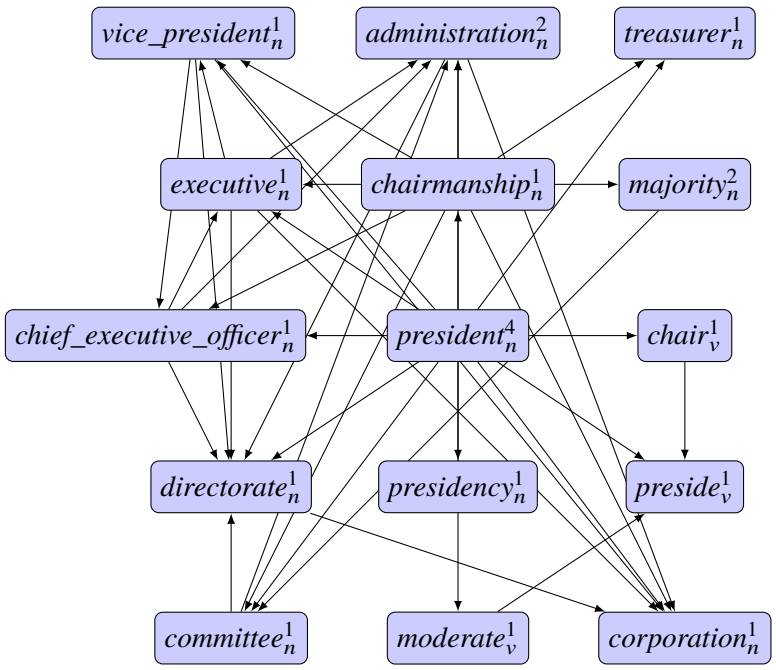
Algorithm 4.2 – Pattern filter based on sub-graphs of lexical-semantic resources. The filter is specific to a relation r. G_r is the semantic graph associated with the relation r, as generated by Algorithm 4.1. $\mathcal{P}_r$ is the set of linguistic patterns learned for r. $lex(\cdot)$ returns for a given synset all the surface forms (lexicalizations) which are listed for this synset in the resource.



a) Sub-graph for *marriage*.



b) Sub-graph for *person death*.



c) Sub-graph for *organization leadership*.

Figure 4.1 – Excerpts from three relation-specific sub-graphs of a lexical-semantic resource.

(*Continues from page 81.*)

Example 4.2 `It was here that the beautiful Etta Place first met Harry Longabaugh.`

The algorithm disambiguates the word senses of all content words in the patterns, with the respective source sentence providing contextual information. In the example sentence, the term `met` would be resolved to the synset[45] $meet^1_v$, given the other words as context: `was`, `here`, `beautiful`, etc.

The next step is to build a core subset of synsets which are most representative for the semantic relation of interest. This set is initially populated with the most frequent synsets in the linguistic patterns. For example, the two most frequent terms for *marriage* in a given corpus could be $marry^1_v$ and $wife^1_n$. The algorithm extends this set with all observed synsets (from the patterns) that have a direct link in the sense repository to one of the most-frequent synsets. In the running example, this means terms like $husband^1_n$, $marriage^1_n$, $divorce^2_v$, and others might be added to the core synsets $marry^1_v$ and $wife^1_n$. Finally, the algorithm returns the desired sub-graph of the given KB. Figures 4.1a–4.1c (p. 83) present excerpts of the sub-graphs we obtain for three relations.

Algorithm 4.2 (p. 82) shows the *semantic filter*, which assesses linguistic patterns based on the semantic sub-graph of a relation. For each pattern associated with the semantic relation, the filter examines if any of the pattern's content words matches a lexicalization of the synsets from the semantic graph. If this is the case, the pattern is retained as a likely correct pattern, otherwise, it is discarded. For example, the filter would recognize the pattern in Example 4.3 below as likely to be correct given the semantic sub-graph shown in Figure 4.1a (p. 83):

Example 4.3 $\text{person} \xleftarrow{\text{nsubj}} \texttt{married} \xrightarrow{\text{dobj}} \text{person}$

In contrast, the pattern from Example 4.1 (p. 81) would be filtered out because none of the potential synsets of `meet` matches any of the core synsets automatically associated with the relation *marriage*.

[45] Where w^i_p denotes the i-th sense of w with POS p.

4.2.2 Experimental Setup

To assess the quality of the semantic-filtering algorithm, more specifically to determine its effects on RE recall and precision, we carried out an extrinsic evaluation. Our goal was to gain insights on the following aspects:

- *Filtering with pattern frequency vs. filtering on the basis of lexical semantics:* We compared the new semantic filter against the FO filter (Chapter 3) to investigate, whether lexical-semantic information allows to identify more noise in learned patterns than existing means do.
- *Impact of the selection among lexical-semantic resources:* We evaluated the effects of running Algorithm 4.1 on two different knowledge bases: WordNet and BabelNet.
- *General applicability of the filtering method:* We applied the method to linguistic patterns from two sources with different kinds of sentence representations, namely dependency grammar and plain token sequences.

Linguistic Patterns The first type of linguistic patterns we used are the ones generated by Web-DARE in the experiment of Section 3.4. Furthermore, we tested the approach with rules of the NELL system (Carlson et al., 2010a; T. M. Mitchell et al., 2015), which was already introduced in Section 3.6. In contrast to Web-DARE, NELL's extraction rules are binary and surface-level oriented. Thus, they convey no syntactic information and solely consist of token sequences. The following are two examples of NELL rules:

Example 4.4 person `who is married to` person.

Example 4.5 person `and husband` person.

Table 4.1 (p. 86) lists the seven target relations of this experiment in column "RELATION", while in column "INPUT" we show the respective number of rules produced by Web-DARE and NELL.[46]

Dataset We employed the Los Angeles Times/Washington Post (henceforth LTW) portion of the English Gigaword v5 corpus (Parker et al., 2011) as a testbed for RE. LTW is comprised of 400K newswire documents from the period 1994–2009. We matched all Web-DARE and NELL rules against the LTW corpus, resulting in approximately 500K detected relation mentions, shown in the column "Extracted Mentions" of Table 4.1. To

[46] NELL rules were taken from iteration 680, `http://rtw.ml.cmu.edu/resources/results/08m/`.

Relation	Input		Evaluation						
	# Rules		# Extracted Mentions		Baseline Precision		# Freebase Mentions		
	WD	N	WD	N	WD	N	WD	N	WD ∪ N
acquisition	26,986	272	17,913	296	14.20%	28.04%	93	1	93
marriage	88,350	547	92,780	2,586	11.60%	8.50%	161	9	168
person birth	22,377	995	63,819	2,607	36.50%	5.60%	77	0	77
person death	31,559	5	84,739	17	18.00%	100.00%	300	0	300
person parent	45,093	956	93,800	358	13.20%	66.20%	91	5	92
place lived	47,689	829	84,389	3,155	47.90%	92.00%	68	38	106
sibling relationship	26,250	432	59,465	211	5.60%	51.18%	48	2	49
sum	288,304	4,036	496,905	9,230	–	–	838	55	885
average	41,186	577	70,986	1,319	21.00%	50.22%	120	20	126

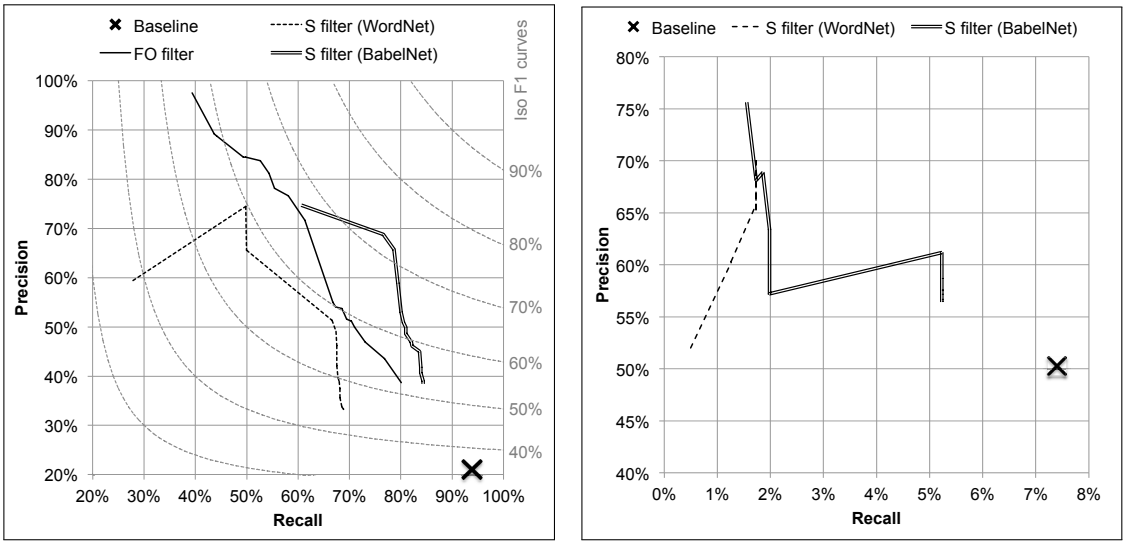
Table 4.1 – Statistics about extraction systems before filtering: (a) the input data for the rule filters, (b) the baseline (pre-filtering) performance for the evaluation. Values are shown for both Web-DARE (WD) and NELL (N) systems. "Freebase Mentions" refers to the number of correctly identified Freebase mentions in a sample of the evaluation corpus.

estimate the precision of RE, we manually checked a random sample of 1K extracted mentions per relation and system; the pre-filtering performance is reported in column "Baseline Precision." To estimate the RE coverage of the rules, we investigated how many mentions of Freebase facts the systems found on LTW; the values are listed in the last three columns of the table, labeled "Freebase Mentions." Only actual mentions were taken into account, i.e., sentences that contain the entities of a Freebase fact and actually refer to the corresponding target relation. Relative recall values stated in this section are to be understood as recall with regard to the set of Freebase-fact mentions found by at least one of the two rule sets (Web-DARE/NELL; last column of Table 4.1).

Miscellaneous WSD was conducted using an off-the-shelf API for knowledge-based disambiguation (Weissenborn et al., 2015a; b). We experimented with different values of n for Algorithm 4.1 (p. 82), ranging from 1 to 15.

4.2.3 Evaluation Results

Figure 4.2a (p. 87) presents the results in terms of precision vs. relative recall when performing RE with the unfiltered Web-DARE rules (our baseline), the statistical approach (FO filter), and the semantic filtering algorithm (S filter). The FO filter can increase precision from the baseline value of 20% up to almost 100%. However, this filtering sacrifices a large portion of the initial recall. The semantic filter trained with BabelNet,



a) Rules from Web-DARE system. **b)** Rules from NELL system.

Figure 4.2 – Performance impact of applying various filters to extraction rules from two systems. "Baseline" corresponds to no filtering. For the semantic filter ("S filter"), the curves resulted from varying n from 1 to 15. "FO filter" is defined in Equation 3.5 on page 66. Results are averaged over seven relations. In the left figure, dashed curves in gray depict points with an equal F1 score.

n (Alg. 4.1)	Precision		Recall		F1 score	
	WN	BN	WN	BN	WN	BN
(Basel.)	21.00		93.83		34.32	
15	33.24	38.50	68.87	84.37	44.84	52.87
14	33.34	39.03	68.87	84.37	44.93	53.37
13	33.71	40.66	68.56	83.89	45.20	54.77
12	35.66	41.80	68.14	83.89	46.82	55.80
11	37.50	45.04	68.14	83.61	48.38	58.55
10	38.89	46.16	68.01	82.20	49.48	59.12
9	39.69	46.91	67.73	82.04	50.05	59.69
8	42.14	48.64	67.54	80.93	51.90	60.76
7	43.81	49.91	67.54	80.93	53.15	61.75
6	46.19	50.89	67.54	80.47	54.86	62.35
5	49.07	52.99	67.40	80.04	56.79	63.76
4	51.34	58.81	66.61	79.44	57.99	67.59
3	65.57	65.76	49.93	78.69	56.69	71.64
2	**74.43**	**68.79**	**49.84**	**76.61**	**59.70**	**72.49**
1	59.43	74.66	27.84	60.73	37.92	66.98

Table 4.2 – Impact of using WordNet (WN) vs. BabelNet (BN) on Web-DARE rule filtering. Results are averaged over seven relations; all values are in %.

in contrast, does not reach a level of precision beyond 75% for the average of the relations targeted in this experiment. However, at the same time, it leads to a more reasonable precision-recall trade-off. For example, the S filter achieves approximately 15 percentage points more recall than the FO filter at a precision level of around 70%. In the recall range covered by the BabelNet filter, its precision remains higher. As illustrated by the chart, training the S filter with WordNet instead of BabelNet leads to inferior performance. Table 4.2 (p. 87) shows the Web-DARE RE performance for different parameter values of Algorithm 4.1. The use of BabelNet consistently leads to a higher F1 score compared to WordNet. For example at $n = 2$, the F1 score is 13 percentage points higher.

Figure 4.2b (p. 87) plots the precision versus relative recall results of the baseline and the semantic-filtering algorithm when applied to NELL's patterns. Again, the RE precision increases. Due to the low number of mentions found in the NELL recall baseline (see Table 4.1, p. 86), the filter application has a high impact on the depicted recall values, and thus the curves show a non-monotonic growth. Nevertheless, as the chart indicates, the proposed filter can also be applied to pattern sets of different RE-rule formalisms. Similarly to Figure 4.2a, Figure 4.2b demonstrates that training the filter on BabelNet leads to superior RE performance compared to the filter variant trained on WordNet.

4.2.4 Result Analysis and Insights

Generality Both Figures 4.2a & 4.2b, as well as Table 4.2, show significant performance improvements after the application of the semantic filter, regardless of the underlying pattern formalism, i.e., dependency-analysis-based or surface-level-based. This indicates that the proposed algorithm could be applied to a variety of application scenarios, as long as the patterns or rules contain content words to which the semantic filter can be applied.

BabelNet vs. WordNet Filtering with sub-graphs of BabelNet led to better RE performance than the alternative scenario where WordNet was utilized. It follows that BabelNet, with its richer inventory of lexical-semantic relations, is better suited for effective pattern filtering. The performance boost by the BabelNet resource can be attributed to the higher coverage of the semantic sub-graphs learned from it. Consider the following example from the *marriage* relation:

Example 4.6

$$\text{person} \xrightarrow{\text{appos}} \texttt{widow} \xrightarrow{\text{prep}} \texttt{of} \xrightarrow{\text{pobj}} \text{person}$$

Example 4.7

$$\text{person} \xleftarrow{\text{poss}} \texttt{widower} \xrightarrow{\text{appos}} \text{person}$$

These rules draw on the concept of deceased spouses, i.e. the term *widow(er)* for detecting the target relation. Since the sub-graph created with BabelNet contains this concept, the rules are identified as being useful for RE and hence they are not filtered out. On the contrary, the filter from WordNet erroneously excludes these rules.

Individual Relations The performance of the semantic filter varies across relations. While it works particularly well for relations like *acquisition* and *person birth/death*, the results are rather discouraging for the relation *place lived*. Investigating the sampled mentions, we found that the unsatisfying results can be attributed to the larger lexical diversity of the *place lived* relation. Often the semantic information is carried by constructions such as "Belfast writer J. Adams", where the lexical anchor "writer" is semantically insignificant to the relation. To get high coverage on such mentions, extraction rules would have to match a certain set of semantically diverse nouns, without matching all nouns ("Belfast visitor Cameron"). The relation seems to require much background knowledge, which may have to include entailment and other inferences. For example, a mention of a person being a senator for some (US) state could, depending on legal requirements, indeed be a mention for *place lived*.

Improvements over FO Filter Finally, we investigated the causes of the superior performance of the new semantic filter compared to the existing FO filter. In addition to being limited to mutually-exclusive relations with compatible entity signatures, the FO filter also has the disadvantage of not excluding erroneous rules which neither belong to the particular target relation nor to any of the compatible relations. In contrast, the new semantic filter works independently for each relation.

The following low-precision rules from Web-DARE illustrate this point. All of them were learned for the *marriage* relation:

Example 4.8 person $\xleftarrow{\text{nsubj}}$ lose $\xrightarrow{\text{prep}}$ to $\xrightarrow{\text{pobj}}$ person

Example 4.9 person $\xleftarrow{\text{nsubj}}$ date $\xrightarrow{\text{dobj}}$ person

Example 4.10 person $\xleftarrow{\text{nsubj}}$ meet $\xrightarrow{\text{dobj}}$ person

These rules express typical relations for married couples and hence get strong statistical support for the *marriage* relation against any other relation. Therefore, the FO filter is not able to correctly identify them as wrong. In contrast, the semantic filter correctly disposes of them.

Another shortcoming of the FO filter is the recurring exclusion of high-quality patterns for which there is only limited support in the training data. When only taking into account the frequency of a pattern, these patterns cannot be distinguished from erroneously learned ones. Our use of an additional lexical-semantic resource provides a filtering mechanism that correctly identifies the appropriate meaning of the target relation. An example is the pattern from Example 4.7 (p. 88). Due to its low frequency, the pattern gets filtered out by the FO filter. At the same time, it expresses a relevant word sense for the considered relation, and thus it gets classified as correct by the semantic filter.

4.2.5 Filter Combinations

A particularly interesting result of the analysis of the semantic filter is that its strengths lie in different areas than the ones of the FO filter. Consequently, we also investigated whether it is possible to jointly apply these two filters in a way that the respective strong points are retained while the weak aspects are eliminated. The individual components that we combine are:

- **S**: The semantic filter of Algorithm 4.2 (p. 82).
- **F**: The absolute-frequency filter component $valid^{r}_{freq}$ from Equation 3.4 (p. 66).
- **O**: The relative-frequency filter component $valid^{r}_{inter}$ from Equation 3.3 (p. 66).

All potential subsets of these components are combined via disjunction and conjunction, meaning that a candidate pattern must pass all or at least one of the filter components. We define the following set of pattern filters, with $\wp$ denoting the power set:

Equation 4.1

$$\left\{ \bigwedge_{c \in C} c \;\middle|\; C \in \wp(\{\mathbf{S}, \mathbf{F}, \mathbf{O}\}) \right\} \cup \left\{ \bigvee_{c \in C} c \;\middle|\; C \in \wp(\{\mathbf{S}, \mathbf{F}, \mathbf{O}\}) \right\}$$

We conduct the experiments on the same dataset for which we reported RE performance in Subsection 4.2.4. Again seven relations are examined, and precision, recall, and F1 score are reported for each of them. As only Web-DARE patterns are used, we deviate from the earlier, "relative" definition of recall and use standard (absolute) recall values.

Results Table 4.3 (p. 91) presents the performance of the pattern filters. Printed in bold are the results of the filters from Subsection 4.2.4, i.e., $\mathbf{F} \wedge \mathbf{O}$ and **S**. We only report results for the best parameter setting per pattern-filter combination and relation. The specific parameter values are depicted in the table, with n corresponding to **S**, and x corresponding to **F**.

	acquisition			*marriage*			*person birth*			*person death*			*person parent*			*place lived*			*sibling relationship*		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
unfiltered	14.20	32.06	19.68	11.60	38.60	17.84	36.50	34.52	35.48	18.00	23.96	20.56	13.20	38.23	19.62	47.90	3.47	6.47	5.60	26.08	9.22
F	32.60	$x \geq 6$ 26.50	29.24	83.50	$x \geq 55$ 27.50	41.37	63.20	$x \geq 4$ 29.50	40.22	46.40	$x \geq 10$ 16.20	24.02	89.30	$x \geq 36$ 31.50	46.57	57.20	$x \geq 2$ 2.80	5.34	100.00	$x \geq 62$ 21.70	35.66
O	17.60	28.60	21.79	24.70	35.00	28.96	67.40	32.70	44.04	27.20	23.40	25.16	41.30	33.60	37.05	34.50	2.60	4.84	24.40	23.30	23.84
S	**61.90**	$n \leq 2$ **31.00**	**41.31**	**63.10**	$n \leq 2$ **32.60**	**42.99**	**65.30**	$n \leq 17$ **29.50**	**40.64**	**79.80**	$n \leq 5$ **22.50**	**35.10**	**56.30**	$n \leq 1$ **37.80**	**45.23**	**63.40**	$n \leq 15$ **2.00**	**3.88**	**26.90**	$n \leq 1$ **25.00**	**25.92**
$\mathbf{F} \wedge \mathbf{O}$	**50.00**	$x \geq 6$ **23.10**	**31.60**	**62.90**	$x \geq 10$ **32.10**	**42.51**	**80.90**	$x \geq 2$ **29.50**	**43.23**	**45.80**	$x \geq 3$ **20.00**	**27.84**	**88.10**	$x \geq 15$ **32.70**	**47.70**	**42.30**	$x \geq 2$ **2.10**	**4.00**	**100.00**	$x \geq 46$ **21.70**	**35.66**
$\mathbf{F} \wedge \mathbf{S}$	84.60	$x \geq 3, n \leq 2$ 27.20	41.16	79.80	$x \geq 8, n \leq 5$ 31.40	45.07	73.90	$x \geq 5, n \leq 17$ 27.30	39.87	84.50	$x \geq 2, n \leq 5$ 20.20	32.61	88.40	$x \geq 17, n \leq 1$ 32.70	47.74	75.40	$x \geq 2, n \leq 15$ 1.90	3.71	100.00	$x \geq 62, n \leq 1$ 21.70	35.66
$\mathbf{O} \wedge \mathbf{S}$	67.00	$n \leq 2$ 28.60	40.09	75.70	$n \leq 5$ 33.50	46.45	87.00	$n \leq 17$ 29.50	44.06	75.20	$n \leq 9$ 22.90	35.11	77.40	$n \leq 1$ 36.10	49.24	50.90	$n \leq 15$ 1.60	3.10	62.30	$n \leq 1$ 24.40	35.07
$\mathbf{F} \wedge \mathbf{O} \wedge \mathbf{S}$	**88.50**	$x \geq 3, n \leq 2$ **24.80**	**38.74**	**97.80**	$x \geq 5, n \leq 5$ **31.60**	**47.77**	**92.90**	$x \geq 2, n \leq 17$ **27.30**	**42.20**	**82.70**	$x \geq 2, n \leq 9$ **20.70**	**33.11**	**95.30**	$x \geq 13, n \leq 1$ **32.70**	**48.69**	**67.70**	$x \geq 2, n \leq 15$ **1.50**	**2.93**	**79.00**	$x \geq 8, n \leq 1$ **23.30**	**35.99**
$\mathbf{F} \vee \mathbf{O}$	17.90	$x \geq 126$ 28.60	22.02	24.70	$x \geq 57$ 35.00	28.96	62.70	$x \geq 12$ 34.50	44.51	28.90	$x \geq 99$ 22.60	25.36	43.80	$x \geq 38$ 34.80	38.78	47.60	$x \geq 2$ 3.30	6.17	24.40	$x \geq 62$ 23.30	23.84
$\mathbf{F} \vee \mathbf{S}$	61.90	$x \geq 126, n \leq 2$ 31.00	41.31	82.40	$x \geq 55, n \leq 1$ 30.40	44.41	59.10	$x \geq 12, n \leq 17$ 33.60	42.84	79.80	$x \geq 217, n \leq 5$ 22.50	35.10	56.70	$x \geq 71, n \leq 1$ 37.80	45.36	57.50	$x \geq 2, n \leq 5$ 3.00	5.70	26.90	$x \geq 57, n \leq 1$ 25.00	25.92
$\mathbf{O} \vee \mathbf{S}$	18.10	$n \leq 2$ 31.70	23.04	24.70	$n \leq 1$ 35.20	29.03	68.70	$n \leq 1$ 32.70	44.31	33.90	$n \leq 1$ 23.70	27.90	39.20	$n \leq 1$ 37.80	38.49	46.60	$n \leq 16$ 3.10	5.81	17.00	$n \leq 1$ 25.00	20.24
$\mathbf{F} \vee \mathbf{O} \vee \mathbf{S}$	18.10	$x \geq 126, n \leq 2$ 31.70	23.04	24.70	$x \geq 110, n \leq 1$ 35.20	29.03	62.80	$x \geq 12, n \leq 1$ 34.50	44.53	33.90	$x \geq 217, n \leq 1$ 23.70	27.90	39.20	$x \geq 71, n \leq 1$ 37.80	38.49	49.00	$x \geq 2, n \leq 7$ 3.40	6.36	17.00	$x \geq 57, n \leq 1$ 25.00	20.24

Table 4.3 – RE performance of Web-DARE patterns for seven relations and several filter combinations. *P*, *R*, *F* are short for precision, recall, and F1 score, respectively.

It can be observed that combining $\mathbf{F} \wedge \mathbf{O}$ and $\mathbf{S}$ into one filter results in improved precision, without a drastically negative impact on recall. For the example of relations *marriage* and *person birth*, the precision is boosted from around 60%/80% to well above 90% at the same recall level. While for individual relations other filter combinations outperform $\mathbf{F} \wedge \mathbf{O} \wedge \mathbf{S}$ (e.g., $\mathbf{O} \wedge \mathbf{S}$ for relation *person parent*), the full combination still shows good performance. The generally low quality of the disjunctive filter aggregation is in part expected. In the end, this kind of combination emphasizes the weaknesses of the individual filters, in contrast to the conjunctive combination which allows a filter component to weed out the patterns erroneously accepted by a concurrent filter.

Remaining Obstacles We further manually investigated the remaining false positive relation mentions for error causes. While for relations like *marriage* and *person parent*, no remaining cases of a violated distant-supervision assumption were identified, relations like *person birth* and *person death* are still affected by this problem, albeit only to a small degree. Consider Examples 4.11 and 4.12 (below) from the latter relation, erroneously extracted as positive mentions.

Example 4.11 In the months after Laci Peterson **disappeared** from her Modesto home on Christmas Eve 2002, the Peterson case garnered more airtime on the big three TV networks' morning news shows than any other story except the war in Iraq.

Example 4.12 Pryor suffered a **heart attack** at his home in the San Fernando Valley early Saturday morning.

While both sentences refer to an event very relevant and indicative, in certain contexts potentially even expressive, for the *person death* relation, on their own they are not directly referring to it. A potential future solution for such probabilistic mentions might be to rather associate patterns and mentions with a confidence score and to embed the use of patterns into a holistic probabilistic framework. In fact, we are taking steps into this direction with the work on sar-graphs, described in Chapter 6. Apart from the wrong-relation issue, the results from Section 3.7 have been confirmed in that preprocessing errors from entity recognition and parsing again have a severe negative impact on performance.

4.3 Relation-Specific Sub-graphs for Enhanced Pattern Generation

So far, the work in this chapter has addressed the problem of low extraction performance. This section investigates how the relation-specific semantic sub-graphs created by Algorithm 4.1 (p. 82) can help to improve the extraction of patterns from parse trees of sentences. At the core of the pattern-generation approach followed in Chapter 3 is the identification of the minimal part of a parse tree that covers all given semantic arguments in a sentence. This approach has the following problems:

- Except for the entities themselves, the minimal subtrees can be semantically empty and may therefore not express any explicit semantic relation between the entities. The symptomatic example for relations between two (or more) persons is the semantically lightweight pattern which solely consists of two person placeholders connected via the conjunction "and".
- A shortest path can be semantically incomplete. The sentence fragment "person celebrates ... wedding with person" indeed suggests a *marriage* relation. However, the pattern-generation method from Chapter 3 would produce a pattern covering only "person celebrates with person", which extracts many events of celebration that are not weddings.

The two problems above indicate that a minimal-subtree solution does not provide sufficient semantic conditions for a correct extraction. In various statistical approaches (e.g., Mintz et al., 2009; Chowdhury and Lavelli, 2012), additional features such as words around entities, words between entities, or trigger words are employed to compensate this shortcoming. We propose an extension of the pattern-discovery algorithm which integrates relation-relevant lexical-semantic information. With this additional guidance, the pattern-discovery algorithm can extract from an annotated parse all of the minimal trees that contain argument entities and one or more semantically related terms.

4.3.1 Algorithms for Pattern Generation

In this section, we briefly revisit the pattern-extraction component from the Web-DARE system. We refer to this baseline strategy as *Shortest-Path Learner*, abbreviated as SPL.

Pattern Discovery in Web-DARE SPL regards a sentence as a candidate of a relation mention if it contains the essential entities of a relation instance. The pattern-extraction

function ExtractPattern($r(a_1, \ldots, a_n), s$)

augment s with morphologic and syntactic information
- ▷ Create dependency-parse $d_s = (\mathcal{V}, \mathcal{E})$ for s.
- ▷ Attach lemmatization information to nodes $\mathcal{V}$ of d_s.

process s & d_s with entity recognition
- ▷ Detect mentions of the arguments $a_1, \ldots, a_n$ in s.
- ▷ Replace the corresponding nodes in $\mathcal{V}$ with placeholders for entity type and role label.

find all sub-graphs $\mathcal{P}_s$ of D_s such that $\forall p \in \mathcal{P}_s, p = (\mathcal{V}_p, \mathcal{E}_p)$:
- (a) $\mathcal{V}_p$ contains two or more of the argument mentions $a_1, \ldots, a_n$,
- (b) p is the minimal subtree of d_s containing shortest paths connecting the nodes defined by (a),
- (c) $\mathcal{V}_p$ contains a content word (i.e., nouns, verbs, adjectives, adverbs).

return $\mathcal{P}_s$

Algorithm 4.3 – Spl pattern-learning algorithm. $r(a_1, \ldots, a_n)$ is an instance of the n-ary target relation r. s is a sentence with mentions of a subset of $\{a_1, \ldots, a_n\}$. The output is a set of graphs $\mathcal{P}_s$, which can be used for pattern-based relation extraction.

algorithm of Spl is outlined in Algorithm 4.3. Given a mention of a target-relation instance, Spl learns one or more RE rules, which are all sub-graphs of the sentence's dependency parse that satisfy the criteria listed in items (a-c) of the algorithm. As an example, consider the *marriage* seed fact in Example 4.13:

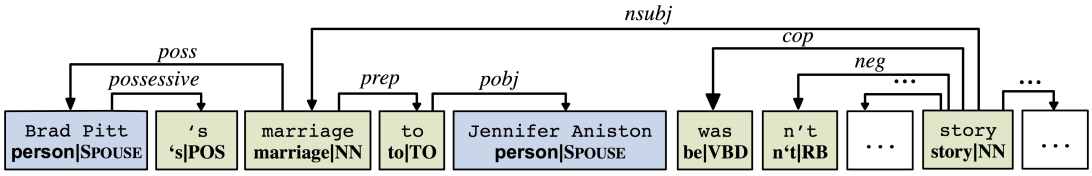
Example 4.13 ⟨ `Brad Pitt` (Spouse), `Jennifer Aniston` (Spouse), – (Ceremony), – (From), – (To) ⟩

Given the sentence in Example 4.14 (below), Spl produces the analysis depicted in Figure 4.3a (p. 95). Here, the entity mentions (in blue) have already been assigned to their semantic roles by exploiting the role mapping from the seed fact.

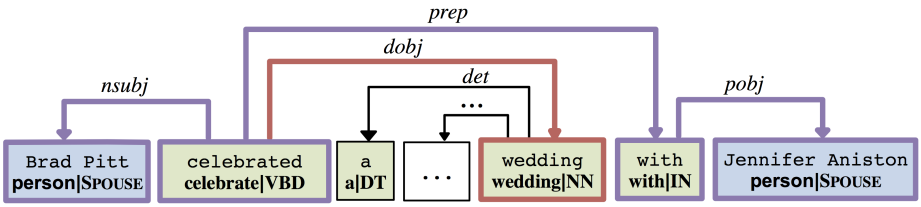
Example 4.14 `In addition, a friend says, Brad Pitt's marriage to Jennifer Aniston wasn't the golden love story it appeared to be.`

Processing this linguistic analysis of the input sentence, Algorithm 4.3 yields the learned rule in Example 4.15 (below), namely, the shortest path that connects the two person names.

Example 4.15 person (Spouse) $\xleftarrow{poss}$ `marriage` $\xrightarrow{prep}$ `to` $\xrightarrow{pobj}$ person (Spouse)



a) Parse of sentence in Example 4.14.



b) Parse of sentence in Example 4.16.

Figure 4.3 – Dependency parses of two sentences. Parts of the parses are left out for brevity. Blue nodes represent detected entity mentions, green nodes correspond directly to tokens of the respective input sentence.

Relation Clues Outside of Minimal Subtrees While the pattern-learning algorithm works reasonably well for many sentences with target-relation mentions, the algorithm fails to extract the gist of the mention if important relation-relevant terms are not contained within the component of the parse that links the arguments. In such cases, the algorithm extracts semantically underspecified rules. Consider the following sentence (Example 4.16) and its linguistic analysis in Figure 4.3b:

Example 4.16 `Brad Pitt celebrated a wonderful wedding with Jennifer Aniston.`

Algorithm 4.3 (p. 94) identifies the sub-graph highlighted in purple as semantically relevant, but misses the path to the verb's object `wedding` (highlighted in red), thus returning a misleading pattern which only captures that two people celebrated something.

A New Paradigm: Enhanced Pattern Learning Algorithm 4.4 (p. 96) presents an extended version of the pattern-extraction approach which relies on the relation-relevant terms in the BabelNet sub-graphs (Section 4.2), described earlier in this chapter. The major difference of the new algorithm with respect to the original one is in items (b) and (c). This change allows the dependency-subtree detection to make a lexical-semantically informed choice. Transferred to the relation *marriage*, this means that $\mathcal{V}_G$ would contain the corresponding synsets for terms like `bride`, `divorce`, `fiance`, `hubbie`, and

function ExtractPatternEnhanced($r(a_1, \ldots, a_n), s, G = (\mathcal{V}_G, \mathcal{E}_G)$)

augment s with morphologic and syntactic information
▷ See Algorithm 4.3.

process s & d_s with entity recognition
▷ See Algorithm 4.3.

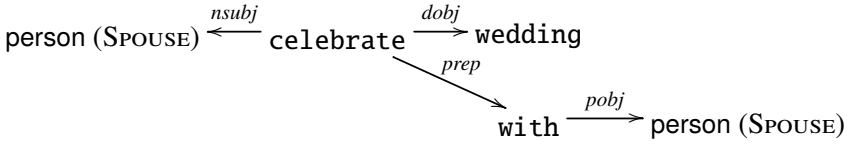
find all sub-graphs $\mathcal{P}_s$ of D_s such that $\forall p \in \mathcal{P}_s, p = (\mathcal{V}_p, \mathcal{E}_p)$:
(a) $\mathcal{V}_p$ contains two or more of the argument mentions $a_1, \ldots, a_n$,
(b) $\mathcal{V}_p$ contains one or more relation-specific semantic terms, i.e., $\mathcal{V}_p \cap lex(\mathcal{V}_G) \neq \emptyset$
(c) p is the minimal subtree of d_s containing shortest paths connecting the nodes defined by (a) & (b)

return $\mathcal{P}_s$

Algorithm 4.4 – Pattern learning with lexical-semantic information. This algorithm is an extension of Algorithm 4.3 (p. 94). $r(a_1, \ldots, a_n)$ is an instance of the n-ary target relation r, while s is a sentence with mentions of a subset of $\{a_1, \ldots, a_n\}$. The additional parameter G is a semantic sub-graph of a lexical-semantic resource, specific to the target relation r and generated by Algorithm 4.1 (p. 82). The output of this algorithm is a set of patterns $\mathcal{P}_s$. $lex(\cdot)$ returns for a given synset all the surface forms (lexicalizations) which are listed for this synset in the resource.

`wedding`, among others. The pattern-extraction process exploits this information during the identification of the shortest paths between arguments by extending the sub-graph until one or more of such terms are included. For the example in Figure 4.3b (p. 95), Algorithm 4.4 identifies the semantic term `wedding` and extracts the following relevant pattern (Example 4.17), which indeed catches the main content of the relation mention:

Example 4.17



4.3.2 Experimental Setup

In the following, we evaluate the impact of the proposed extension on the RE performance of three semantic relations. We compare the performance of patterns learned using Algorithms 4.3 and 4.4, as well as a third pattern set, which represents an alternative way to incorporate lexical semantics into pattern learning:

- **Spl:** Patterns from Algorithm 4.3 (p. 94).
- **Spl+S filter:** Patterns from Algorithm 4.3 after a subsequent pattern-filtering step. Only patterns containing the semantic terms in the lexical semantic sub-graphs are

	training data			learned patterns			matched patterns		
	#seeds	#docs	#synsets	SPL	SPL+ S filter	NEWPL	SPL	SPL+ S filter	NEWPL
marriage	5,993	211,186	54	88,456	33,822	79,178	498	112	166
person parent	3,379	148,598	126	45,093	29,592	76,765	357	159	272
sibling rel.	7,630	130,448	56	26,250	13,004	38,412	204	70	132

Table 4.4 – Statistics about training data and RE rules. "Matched patterns" refers to the amount of patterns which matched at least one sentence in the evaluation corpus.

kept. This combination employs the semantic filter from Algorithm 4.2 (p. 82).

- **NEWPL:** Patterns from Algorithm 4.4 (p. 96).

To generate training examples for the pattern-learning step, we re-ran the DS approach from Chapter 3 for three target relations. The first part of Table 4.4 lists details about the training data for the individual relations. In this table, *synsets* refers to the nodes of the respective relation-specific sub-graph (i.e., "# synsets" = $|\mathcal{V}_G|$, for $\mathcal{V}_G$ from Algorithm 4.4).

Table 4.4 also lists statistics about the number of RE rules per pattern set and relation. The number of patterns that is generated by this new approach is similar to the number of patterns generated by the original approach. Yet, compared to SPL+S filter, the amount of rules is twice as high. Since all the rules in one set differ lexically and/or syntactically, an ideal evaluation of the rules would require an enormous annotated corpus in order to validate a larger fraction of the patterns. As corpora of such size are expensive to create, the experiments for this task were carried out using an extension of the PEOPLE$_{Test}$ corpus used in Chapter 3, called the CELEBRITY corpus. The right half of Table 4.4 states the number of rules which matched a sentence in this corpus. In the following, we briefly describe the annotation process of CELEBRITY.

Evaluation Corpus CELEBRITY consists of newspaper articles which are annotated with gold mentions of three kinship relations: *marriage*, *person parent*, and *sibling relationship*, whose argument signatures are given on page 60 in Table 3.1. The provided annotation specifies the marked facts down to the token level and thereby allows the detailed analysis of language phenomena. The corpus was annotated using the markup tool Recon (H. Li et al., 2012) by two human experts, with additional conflict resolution performed by a third expert. The annotation effort was focused on relations mentioned within individual sentences.

We selected the 150 longest documents from a collection of PEOPLE-magazine articles from the years 2001–2008, the same article basis as used to build PEOPLE$_{Test}$. After the duplicate removal, 142 documents with 364,400 words remained. To speed up the annotation process, we preprocessed the corpus with entity recognizers for persons, organizations, locations, and dates. Approximately 30k mentions of concepts were detected. In some cases, an entity being part of a relation mention is referenced to by several concept mentions within a sentence. To ensure that experts consistently annotate such cases, we define the "nearest arguments" principle, stating that the mutually nearest concept mentions should serve as the arguments of relation mentions. In total, 971 sentences were identified as relation-bearing by at least one of the annotators.

After the initial annotation, we calculated the agreement between the two experts both on the sentence level and on the level of relation mentions. The sentence-level agreement is evaluated by reducing the complexity of the annotation to the binary choice whether a given sentence of the corpus contains a mention of a given target relation. This abstraction allows to employ standard inter-annotator agreement metrics. Across the three relations, the two initial annotators reached a Pearson correlation coefficient of 0.8554 and Cohen's κ of 0.8910. To measure the agreement of the actual relation-mention annotation, we used the *agr* metric described by Wiebe et al. (2005):

Equation 4.2
$$agr(A||B) = \frac{\text{\# of relation mentions annotated by raters } A \text{ and } B}{\text{\# of relation mentions annotated by rater } A}$$

The agreement in terms of this metric is $agr(A||B) = 0.8792$ and $agr(B||A) = 0.8486$. The high agreement values on both levels indicate that relations between persons are commonly expressed in a clear and relatively objective way.

In order to unify the annotation of the two raters, we performed a combination of automatic and manual conflict-resolution. At first, we merged the concept mentions in the sentences with marked relation mentions. The major fraction of the concept mentions were marked by both annotators in exactly the same way. Conflict cases included mentions only annotated by one of the experts or disagreement in the exact extent of a mention. The unification of concept mentions resulted in approximately 4k concept mentions In addition, approximately 400 relation mentions with disagreement in the arguments were adjudicated by a third human expert. Using all 875 conflict-free relation mentions annotated both by A and B and all of the correct mentions as judged by the third rater and after removing projections of mentions, we eventually obtained 1,220 relation mentions. The final annotation of CELEBRITY features 142 documents with 25k sentences, out of which 967 contain one of the 1,220 relations mentions.

		SPL	SPL+S filter	NEWPL
Precision	*marriage*	16.49%	40.00%	38.70%
	parent-child	17.89%	36.80%	33.30%
	siblings	4.89%	13.40%	27.70%
	macro-avg.	13.09%	**30.07%**	**33.23%**
Recall	*marriage*	50.96%	43.80%	48.40%
	parent-child	40.76%	35.50%	49.50%
	siblings	18.36%	17.80%	62.20%
	macro-avg.	36.69%	**32.37%**	**53.37%**
F1 score	*marriage*	24.91%	41.81%	43.00%
	parent-child	24.86%	36.13%	39.81%
	siblings	7.72%	15.28%	38.33%
	macro-avg.	19.30%	**31.17%**	**40.96%**

Table 4.5 – Relation-extraction performance on CELEBRITY corpus.

4.3.3 Evaluation Results

Table 4.5 lists statistics about the RE performance of the three pattern sets (from SPL, SPL+S filter, and NEWPL) on CELEBRITY. The new method improved both precision and recall significantly for each target relation. The average precision improvement in comparison to the baseline system is 20.4%, while the improvement of recall is 16.68%.

While applying lexical semantics to rule filtering does help to improve precision (SPL+S filter vs. SPL; confirming the results from Section 4.2), it inevitably leads to a recall drop due to a reduced number of rules available for extraction. The new algorithm NEWPL is naturally able to achieve the same precision improvement because it restricts the possible pattern set during pattern learning by utilizing the same lexical-semantic information as SPL+S filter. However, NEWPL is also capable of lifting recall to a higher level since it creates patterns from relation mentions with semantic indicators outside the shortest paths between arguments. We discuss examples in the next paragraphs. The results show that, in comparison to both baseline approaches, the extended pattern-generation algorithm is capable of increasing recall without hurting precision.

Result Analysis Finally, to conclude this section, we analyze differences in the pattern sets that lead to the increased recall of NEWPL compared to SPL and SPL+S filter. We also present examples of cases where mistakes in the learning process led to the extraction of erroneous patterns.

Some target-relation mentions link the persons participating in the relation only by

a conjunction, shifting relation triggers to the context of the argument mentions. In many cases, with the help of the new approach, we can identify a trigger word as being semantically relevant, and thus we can incorporate it in the extracted pattern. This is illustrated by the *marriage* patterns in Examples 4.18 and 4.20, matching the sentences in Examples 4.19 and 4.21, respectively:

Example 4.18 wedding $\xrightarrow{nn}$ person (SPOUSE) $\xrightarrow{conj}$ person (SPOUSE)

Example 4.19 The good feelings were on display the evening of Scott and Laci's wedding.

Example 4.20 marry $\xrightarrow{nsubj}$ person (SPOUSE) $\xrightarrow{conj}$ person (SPOUSE)

Example 4.21 Two years after Aniston and Pitt married, ...

A similar example pattern from the same relation is Example 4.22, which again contains a semantic key term outside of the shortest path between the relation arguments. In addition, corresponding rules were learned for the other two relations, i.e., for *parent-child* patterns like Example 4.23 and for *siblings* ones like Example 4.24:

Example 4.22 ex-husband $\xleftarrow{nn}$ person (SPOUSE) $\xrightarrow{poss}$ person (SPOUSE)

Example 4.23 person (PARENT) $\xleftarrow{poss}$ person (CHILD) $\xrightarrow{nn}$ daughter|son|child

Example 4.24 person (SIBLING) $\xleftarrow{poss}$ person (SIBLING) $\xrightarrow{nn}$ brother|sister

We also found that sound patterns exclusively learned by NEWPL sometimes produce incorrect extractions due to erroneous syntax analysis in application sentences. For example, the rule in Example 4.25 mistakenly matches the sentence in Example 4.26 because of an incorrect dependency analysis:

Example 4.25 person (SPOUSE) $\xleftarrow{conj}$ person (SPOUSE) $\xrightarrow{rcmod}$ marry

Example 4.26 ... between Amber and Scott, who had told her he was not married.

Finally, another issue resulting in false-positive extractions can be attributed to the fact that the semantic sub-graph for a relation may contain terms of unclear significance to the relation. For example, the following patterns, Examples 4.27 and 4.28, were learned

for the relation *marriage*.

Example 4.27 person (SPOUSE) $\xrightarrow{conj}$ person (SPOUSE) $\xrightarrow{nn}$ `partner|girlfriend`

Example 4.28 `relationship` $\xrightarrow{prep}$ `with` $\xrightarrow{pobj}$ person (SPOUSE) $\xrightarrow{conj}$ person (SPOUSE)

The semantic terms in them may in some cases indeed indicate an embedded mention of this relation, but will usually not be of great utility to distinguish actual relation mentions from negative ones. These examples suggest that further work might need to be invested into the creation of stricter versions of the relation-specific semantic graphs.

4.4 Related Work

In this chapter, we pointed out that minimal subtrees or shortest paths connecting entities often do not provide sufficient semantic context for extracting target relations. This is not an entirely new insight. In fact, domain- or relation-relevant terms were integrated into RE rules before. Early IE systems used event trigger words to locate relevant sentences or instances (e.g., Grishman and Sundheim, 1996; Appelt, 1999; Grishman et al., 2005). Other systems automatically learned lexical-syntactic patterns that included words between and around the entities of the instance (e.g., Ravichandran and Hovy, 2002; Agichtein, 2006). Words in the textual context of the entities of a relation mention were commonly employed as features in statistical approaches, in addition to dependency patterns, particularly in DS approaches (Mintz et al., 2009; Jean-Louis et al., 2013; Min et al., 2013). However, in these works, the words were selected by their textual distance to the entities and not on the basis of their semantic domain relevance.

Two closely related approaches to pattern discovery with semantic indicators are Grishman et al. (2005) and H. Xu et al. (2009). Grishman et al. presented a supervised pattern-discovery approach to EE. Utilizing a training corpus annotated with both event arguments and event anchors, paths were learned between the event trigger and the individual arguments. The drawback of this method was the need for manual labeling of training data with event triggers. In the second related approach, proposed by H. Xu et al., dependency patterns were learned for the detection of binary relations. The patterns had to contain at least three nodes: The two semantic arguments and a key word which indicated the semantic relation. This approach was mostly suitable for learning from small, manually annotated corpora as the relation-relevant keywords were only acquired from manual annotation.

Other works dealt with the acquisition of relevant terms for semantic relations. Q.

Nguyen et al. (2010) analyzed the distribution of *trigger words* for semantic relations in annotated data in order to filter extraction patterns. In the same vein, F. Xu et al. (2002) collected relevant terms with a TFIDF-based strategy. Further approaches incorporated lexical knowledge from WordNet. G. Zhou et al. (2005) presented a feature-based relation extractor which utilized semi-automatically-built trigger-word lists from WordNet. Culotta and Sorensen (2004) used WordNet hypernyms for increased extraction coverage. Stevenson and Greenwood (2005) defined a similarity function for learned linguistic patterns that was built on WordNet information. They started with some positive patterns, i.e., patterns that definitely expressed the target relation. When processing the input documents, their algorithm tried to identify patterns with a similar *meaning* to those already known by exploiting information from WordNet. Their approach was based on the assumption that useful patterns have similar meanings to the already accepted patterns, which can be measured with cosine similarity. A drawback of these methods is that none of them explicitly determines and outputs which parts of the lexical-semantic resource contain the terms that are relevant to a given semantic relation.

Apart from semantic indicators, many other ways for the quality assessment of extracted patterns and instances are implemented in literature. Some approaches used the confidence value of the extracted instances or the seed examples as feedback for estimating the confidence of rules (Agichtein, 2006; Brin, 1998; Yangarber et al., 2000). In most cases, however, the confidence values relied on redundancy. Many approaches used negative examples for filtering (Mintz et al., 2009; F. Xu, Uszkoreit, and Krause, et al., 2010; Yangarber, 2003). Also, lexical features such as word sequences or part-of-speech information were often utilized for further filtering (Banko and Etzioni, 2008; Banko et al., 2007; Mintz et al., 2009; Yates et al., 2007). While some of the approaches listed above do achieve high extraction precision, this comes at the cost of a drastic recall loss.

To obtain high precision while at the same time preserving recall, the use of semantic approaches can be highly beneficial. One of the first attempts was presented by S. Miller et al. (2000), where the authors proposed a method for adding semantic features to labeled training data for a syntactic parser. This, again, has the drawback of requiring huge volumes of manually annotated data, which, even today, is hard to obtain for some particular domains. Other approaches added semantic features to feature-based RE systems that learned relation-specific extractors (Kambhatla, 2004; G. Zhou and M. Zhang, 2007). However, none of these took full advantage of syntactic and semantic analysis, and thus they achieved only small improvements (Jiang and Zhai, 2007). Another trend in this research strand was the utilization of tree kernel-based approaches, which can efficiently represent high-dimensional feature spaces (T. V. T. Nguyen and Moschitti,

2011b; G. Zhou et al., 2010). However, supervision was still required, and semantic analysis was only marginally employed.

4.5 Summary

After the successful utilization of parse-based patterns for large-scale RE in Chapter 3, Chapter 4 dealt with two problems not sufficiently addressed by the Web-DARE system. The first issue concerns the frequent violation of the DS assumption, which leads to erroneous patterns that are not in all cases detected by Web-DARE's original filter. The second issue is related to deficient pattern abstraction from sentences with true relation mentions. Both problems are approached by identifying, for a given target relation, relevant terms in a lexical-semantic repository. These terms then help to reliably extract high-quality patterns from examples. As a side result of the comparison between the FO-Filter and the new semantic filter, we observed that the two methods exhibit different shortcomings. This gave rise to the hope that a combination of both filters may further improve RE performance, which we subsequently confirmed through additional experiments.

Finally, we demonstrated in this chapter that exploiting advanced comprehensive semantic knowledge resources can significantly improve extraction performance in closed extraction settings. However, this target-relation specific setting is of limited use for explorative applications, where not only predefined fact types are of interest, but rather the full range of knowledge and information embedded into texts is to be harvested. We will proceed in the next chapter with the design of a method that ensures high-extraction quality in such open settings.

Chapter 5

Distributed Representations for Patterns[47]

Contents

[47] This chapter presents results from joint work with Enrique Alfonseca, Katja Filippova, and Daniele Pighin.

5.1 Introduction

The two previous chapters discussed RE methods for settings with fixed relation schemas, defined a-priori. There are, however, less restricted application settings where a predefined relation taxonomy is not readily available and cannot be easily created. This is the case in *open*-domain scenarios (Yates et al., 2007, Subsection 2.3.5) and applications with an *exploratory* character (Akbik, 2016). Here, the relations may vary from one document to the next, they may even be completely unknown with respect to background KBs, and often a relation taxonomy needs to be built up from scratch during the extraction phase. A further motivation for a different kind of extraction methodology is the vision of *machine reading*, which calls for automatic, unsupervised, and thorough understanding of text (see the proceedings of the 2007 AAAI Spring Symposium on this topic, AAAI, 2007, in particular, Etzioni et al., 2007; also Poon et al., 2010). In systems for machine reading, the targeted relations cover a greater semantic range than in classic KB domains, namely they include common-sense knowledge (Akbik and Michael, 2014) and fine-grained types of semantic relations and events. Examples of the latter may include separate categories for celebrating a wedding vs. being married and also a distinction between committing a terrorist attack vs. placing a briefcase with a bomb in a public area – note that both examples would be seen as one respective relation at the granularity level of Web-DARE. In this chapter, we report on the design of a system from the area of open IE, which can cope with these demands.

Open-IE systems (Banko et al., 2007) extract textual relational patterns between entities automatically (Fader et al., 2011; Mausam et al., 2012) and subsequently organize them into paraphrase clusters, each of which can be interpreted as a fine-grained relation type. These pattern clusters were found to be useful for RE (Moro and Navigli, 2012; Grycner and Weikum, 2014) and many other tasks like question answering (Lin and Pantel, 2001; Fader et al., 2013). A particular sub-problem of open IE is that of automatically extracting and paraphrasing *event patterns*: those that describe changes in the state or attribute values of one or several entities. An existing approach to learning paraphrases of event patterns is to build on the following idea for a weak supervision signal:

> *News articles that were published on the same day and that mention the same entities should contain good paraphrase candidates.*

Two state-of-the-art event-paraphrasing systems that are based on this assumption are NewsSpike (Congle Zhang and Weld, 2013) and Heady (Alfonseca et al., 2013; Pighin et al., 2014). These two systems have specific weak and strong points and follow different design principles.

1) **Scope of generalization.** In NewsSpike, the paraphrase clusters are learned separately for each publication day and entity set, and the system cannot generalize across events of the same type involving different entities occurring on the same or different days. Consider the following example:

   - Assume the event verbs `has married` and `wed` appeared in news about two persons *A* and *B* marrying.
   - Further assume that the verbs `has married` and `tied the knot with` occurred in news involving two different persons *C* and *D*.
   - Then, NewsSpike would not be able to infer that `wed` and `tied the knot with` are also paraphrases, unless a post-processing is done.

   Heady overcomes this limitation thanks to a global model that learns event representations across different days and sets of entities. However, the global nature of the learning problem can incur other drawbacks. First, training a global model is more costly and harder to parallelize. Second, relatively frequent patterns that erroneously co-occur with other patterns may have an adverse impact on the final models, potentially resulting in noisier clusters. Lastly, low-frequency patterns are likely to be discarded as noise in the final model. Overall, Heady is better at capturing paraphrases from the head of the pattern distribution and is likely to ignore most of the long tail where useful paraphrases can still be found.

2) **Simplifying assumptions.** We already mentioned that the two systems share a common underlying assumption, i.e., that good paraphrase candidates can be found by looking at news published on the same day and mentioning the same entities. On top of this, NewsSpike also assumes that better paraphrases are reported around time-wise spiky entities and that there is one event mention per discourse, in addition, verb tenses may not differ. These restrictions are not enforced by Heady, where the common assumption is indeed even relaxed across days and entity sets.

3) **Annotated data.** NewsSpike requires manually annotated data to train the parameters of a supervised model that combines the different heuristics, whereas Heady does not need annotated data.

This chapter presents the Idest system, a new method for learning paraphrases of event patterns that is designed to combine the advantages of these two systems and to compensate for their respective weaknesses. It is based on a new neural-network architecture that, like Heady, only relies on the weak supervision signal that comes from news published on the

same day, requiring no additional heuristics or training data. Unlike NewsSpike, it can generalize across different sets of extracted patterns, and each event pattern is mapped into a low-dimensional embedding space. This allows us to define a neighborhood around a pattern to find the ones that are close in meaning.

Idest produces a robust global model that can also capture meaningful representations for rare patterns, and thus it overcomes one of Heady's main limitations. Our evaluation of the potential trade-off between local and global paraphrase models shows that comparably good results to NewsSpike can be attained without relying on supervised training. At the same time, the ability of Idest to produce a global model allows it to benefit from a much larger news corpus.

5.2 Related Work

In this section, we (briefly) review work on the core topics of this chapter, namely open IE (Subsection 2.3.5) and neural methods for automatic text processing (Section 2.6). We further introduce important terminology and explain the two systems NewsSpike and Heady in more depth.

Relational Open IE In an early attempt to move away from domain-specific, supervised IE systems, Riloff (1996) strived for an automatic solution to find relational patterns on the web and in other unstructured resources in an open-domain setting. This idea was further explored in more recent years by Brin (1998), Agichtein and Gravano (2000), Ravichandran and Hovy (2002), and Sekine (2006), among others. Banko et al. (2007) introduced open IE and the TextRunner system, which extracted binary patterns using a few selection rules applied on the dependency tree. More recent systems such as ReVerb (Fader et al., 2011) and Ollie (Mausam et al., 2012) also defined linguistically-motivated heuristics to find text fragments or dependency structures that can be used as relational patterns. In this chapter, we experiment with different pattern-extraction methods, all of them applicable to a-priori unknown relations. A description of the methods will follow in Section 5.4.

A natural extension of the previous work is to automatically identify which of the extracted patterns share the same meaning. This can be achieved by producing either a hard or a soft clustering. Lin and Pantel (2001) used the mutual information between the patterns and their observed slot fillers. Resolver (Yates and Etzioni, 2007) introduced a probabilistic model called the Extracted Shared Property where the probability that two instances or patterns are paraphrases was based on how many properties or instances they

shared. USP (Poon and Domingos, 2009) produced a clustering by greedily merging the extracted relations. Yao et al. (2012) and Alfonseca et al. (2012) employed topic models to learn a probabilistic model that could capture the ambiguity of polysemous patterns as well. More recent work also organized patterns in clusters or taxonomies using distributional methods on the pattern contexts or extracted entities (Moro and Navigli, 2012; Nakashole et al., 2012b), or implicitly clustered relational text-patterns via the learning of latent feature vectors for entity tuples and relations, in a setting similar to knowledge-base completion (Riedel et al., 2013).

A shared difficulty for systems that cluster patterns based on the arguments they select is that it is tough for them to distinguish between identity and entailment. If one pattern entails another, both are likely to be observed in the corpus involving the same entity sets. The two patterns in Examples 5.1 and 5.2 illustrate this problem. Both patterns can be observed involving the same pairs of entities, but carry a different meaning. As discussed below, relying on the temporal dimension (given by the publication date of the input documents) is one way to overcome this problem.

Example 5.1 [48] person $\xleftarrow{\text{nsubj}}$ married $\xrightarrow{\text{dobj}}$ person

Example 5.2 person $\xleftarrow{\text{nsubj}}$ dated $\xrightarrow{\text{dobj}}$ person

Event Patterns and Open IE Although some earlier work uses the temporal dimension of text as filters to improve the precision of relational pattern clusters, NewsSpike and Heady fully rely on it as their main supervision signal. In order to compare the two approaches, we start by defining some terms:

- An *event pattern* encodes an expression that describes an event. It can be a linear surface-pattern or a lexico-syntactic pattern, and it can include entity-type restrictions on the arguments. For instance, the pattern in Example 5.1 represents a binary lexico-syntactic pattern that corresponds to a wedding event between two people.

- An *extraction* is a pattern instance obtained from an input sentence, which involves specific entities. For example, the sub-graph represented with solid dependency edges in Figure 5.1 (p. 109) is an extraction corresponding to the pattern in Example 5.1.

[48] Patterns presented in this chapter preserve the original verb inflections. This allows covering more subtleties in the expression of events in patterns.

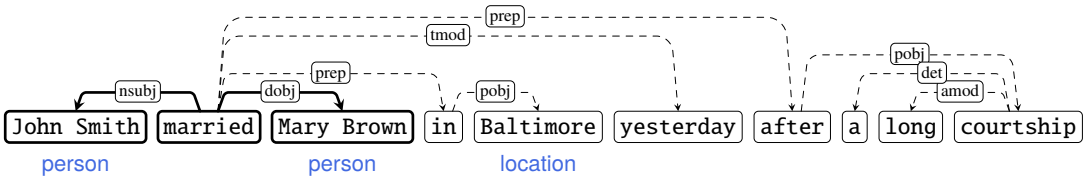


Figure 5.1 – Example sentence and corresponding extraction. The extraction is highlighted with solid edges.

- An *extracted-event-candidate set* (EEC; Congle Zhang and Weld, 2013, used a similar terminology) is the set of extractions obtained from news articles published on the same day and involving the same set of entities.
- Two extractions are *co-occurrent* if there is at least one EEC that contains both of them.

NewsSpike produces extractions from the input documents using ReVerb (Fader et al., 2011). The EECs are generated from the titles and all the sentences of the first paragraph of the documents published on the same day. From each EEC, potentially one paraphrase cluster may be generated. The model is a factor graph that captures several additional heuristics. Integer linear-programming is then used to find the maximum-a-posteriori solution for each set of patterns. Model parameters are trained using a labeled corpus that contains 500 of these sets.

Heady only considers titles and first sentences for pattern extraction and trains a two-layer noisy-or Bayesian network, in which the hidden nodes represent possible event types and the observed nodes represent textual patterns. A maximum-likelihood model is the one in which highly co-occurring patterns are generated by the same latent events. The output is a global soft clustering, in which two patterns may also be clustered together even if they never co-occur in any EEC, as long as there is a chain of co-occurring patterns generated by the same hidden node. Heady was evaluated using three different extraction methods: a heuristic-based pattern extractor, a sentence compression algorithm, and a memory-based method (Section 5.4). While this model produces a soft clustering of patterns, Heady was evaluated only on a headline-generation task and not intrinsically with respect to the quality of the clustering itself.

Neural Networks and Distributed Representations Another related field aims to learn continuous vector-representations for various abstraction levels of natural language. In particular, the creation of so-called word embeddings has attracted a lot of attention

in the past years, often by implementing neural-network language models. Prominent examples include the works by Bengio et al. (2003) and Mikolov et al. (2013b), with the skip-gram model of the latter providing a basis for the vector representations learned in our approach. Also closely related to Idest are approaches which employ neural networks capable of handling word sequences of variable length. For example, Le and Mikolov (2014) extended the architectures of Mikolov et al. (2013b) with artificial paragraph tokens, which accumulated the meaning of words appearing in the respective paragraphs.

In contrast to these shallow methods, other approaches employed deep multi-layer networks for the processing of sentences. Examples include Kalchbrenner et al. (2014), who employed convolutional neural networks for analyzing the sentiment of sentences, and Socher et al. (2013), who presented a special kind of recursive neural network utilizing tensors to model sentence semantics in a compositional way, guided by parse trees. A frequent issue with deeper methods is the high computational complexity coming with the large number of parameters in a multi-layer neural network or in the value propagation in unfolded recursive neural networks. To circumvent this problem, our model is inspired by Mikolov's simpler skip-gram model, as described in Section 5.3.

5.3 Proposed Model

Similar to Heady and NewsSpike, our model is based on the underlying assumption that if sentences from two news articles were published on the same day and, at the same time, mention the same entity set, then they are good paraphrase candidates. The main novelty is the way we train the paraphrase model from the source data. We propose a new neural-network architecture which can learn meaningful distributed representations of patterns.

Skip-gram Neural Network The original skip-gram architecture (Mikolov et al., 2013b) is a feed-forward neural network which is trained with distributional input examples following the assumption that each word should be able to predict to some extent the other words in its context. A skip-gram architecture consists of:

1. An input layer, usually represented as a so-called one-of-V (or one-hot-spot) layer. This layer type has as many input nodes as there are items in the vocabulary ($\hookrightarrow V$). Each training example activates exactly one input node corresponding to the current word x_i in a sequence of words $x_1, x_2, \ldots, x_i, \ldots$ (a text), all other input nodes are set to zero.

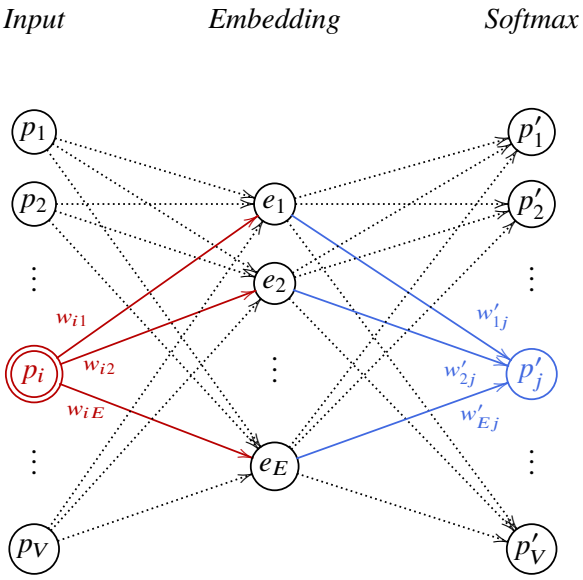


Figure 5.2 – Visualization of the model used for training. V is the total number of unique patterns, which are used both in the one-of-V input and output. E is the dimensionality of the embedding space.

2. A first hidden layer, the embedding or projection layer, that will learn a distributed representation for each possible input word.

3. Zero or more additional hidden layers.

4. An output layer, expected to predict the words in the context K words to the left and right of x_i: $x_{i-K}, \ldots, x_{i-1}, x_{i+1}, \ldots, x_{i+K}$.

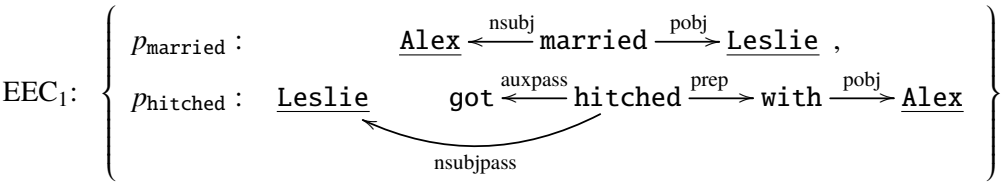
In practice, when training with this architecture, the network converges towards representing words that appear in similar contexts with vectors that are close to each other, as close vectors will produce a similar distribution of output labels in the network.

Idest Neural Network Figure 5.2 shows the network architecture we use for training our paraphrase model in Idest, which is inspired by the skip-gram network. In our case, the input vocabulary is the set of V unique event patterns extracted from texts, and our supervision signal is the co-occurrence of event patterns in EECs. Both input and output layers are set to be a one-of-V layer and will have only one active node (value is 1) in each example. For each pair of patterns that belongs to the same EECs, we will have these patterns predict each other respectively (in two separate training examples).

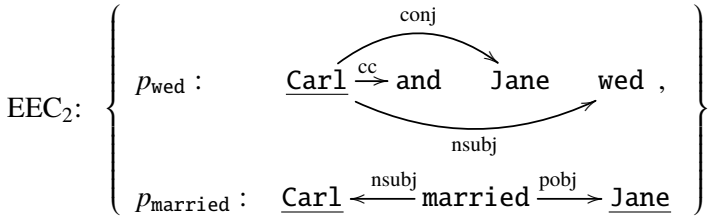
After training, if two patterns p_i and p_j have a large overlap in the set of entities they co-occur with, they should be mapped onto similar internal representations. Note

that the actual entities are only used for EEC construction, but they do not play a role in the training itself. Thus they allow the network to generalize over specific entity instantiations. For example, given the EECs in Examples 5.3 and 5.4, Idest could learn an embedding space in which $p_{\texttt{hitched}}$ and $p_{\texttt{wed}}$ are relatively close, even though the two patterns do not co-occur in the same EEC. This is possible because both patterns have been trained to predict the same pattern $p_{\texttt{married}}$.

Example 5.3



Example 5.4



Reported representations of word embeddings typically use between 50 and 600 dimensions (Mikolov et al., 2013b; Levy and Goldberg, 2014b). For our pattern embeddings, we have opted for an embedding layer size of 200 nodes. We also experimented with larger sizes and with adding more intermediate hidden layers. While the added cost regarding training time was substantial, we did not observe a significant difference in the results.

5.4 Experimental Settings

Pattern Extraction Methods Used In previous works on open IE, different methods for pattern extraction from sentences were used:

- *Heuristic-based*: A number of hand-written rules or regular expressions based on POS tags or dependency trees are used to select the most likely pattern from the source sentence (Fader et al., 2011; Mausam et al., 2012; Alfonseca et al., 2013).
- *Sentence compression*: Takes as input the original sentence and the entities of interest and produces a shorter version of the sentence that still includes the entities (Pighin et al., 2014).

```
function ComputeClusters(P, θ)
    Result = {}
    SortByFrequency(P)
    while |P| > 0 do
        p = Pop(P)                          ▷ Take highest-frequency pattern
        C_p = {p}                           ▷ Initialize cluster around p
        N = Neighbors(p, P, θ)              ▷ N ⊂ P,  ∀n∈N: Sim(n, p) > θ
        for all n ∈ N do
            C_p = C_p ∪ {n}
            Remove(P, n)                    ▷ Remember n has been used
        Result = Result ∪ {C_p}
    return Result
```

Algorithm 5.1 – Cluster generation for extracted patterns, based on their distributed representations. $\mathcal{P}$ is the set of extracted patterns, and θ is the similarity threshold to include two patterns in the same cluster. Sim() returns the cosine similarity of (the representation of) two patterns. SortByFrequency() arranges patterns by descending frequency.

- *Memory-based*: Tries to find the shortest reduction of the sentence that still includes the entities, with the constraint that its lexico-syntactic structure has been seen previously as a full sentence in a high-quality corpus (Pighin et al., 2014).

The pattern extraction method we described as part of the Web-DARE system (Chapter 3) belongs to the first category. The enhanced variant presented in Section 4.3 is not applicable here, as it requires an a-priori relation definition.

It is important to note that the final purpose of the system influences the decision of which extraction method to choose. Pighin et al. (2014) used the event models to generate headlines. They found that using the memory-based method resulted in more grammatical headlines at the cost of coverage. If the purpose of the patterns is the extraction for knowledge-base population, then the importance of having well-formed and complete sentences as patterns becomes less obvious, and higher coverage methods become more attractive. For these reasons, in this chapter we focus on the first two approaches, heuristic-based and sentence compression, as they are well-established and can produce high-coverage output. More specifically, we use ReVerb extractions and a statistical compression model trained on pairs of sentences and their compressed counterparts, implemented after Filippova and Altun (2013). The patterns based on the former formalism are linear surface-patterns similar in style to Examples 4.4 and 4.5 (p. 85); the compression-derived patterns build on dependencies and correspond to what is shown in Examples 5.3 and 5.4 (p. 112).

Generating Clusters from the Embedding Vectors In its native form, Idest does not produce a clustering like NewsSpike and Heady. Hence, to be able to compare against these extraction systems, we used Algorithm 5.1 (p. 113) to build paraphrase clusters from the pattern embeddings. Given a similarity threshold on the cosine similarity of embedding vectors, we start by sorting the patterns by extraction frequency and proceed in order along the sorted vector. Each visited frequent pattern starts a new cluster, and all neighboring patterns are added with respect to the threshold. Used patterns are removed from the original set to ensure that a pattern is not added to two clusters at the same time.

5.5 Evaluation Results

This section opens with a quantitative look at the clusterings obtained with the different methods. Here, we aim to understand the implications of design decisions with respect to the distribution of event clusters and their internal diversity. Then, in Subsection 5.5.2, we complement these figures with the results of a manual quality evaluation. Note that both Heady and NewsSpike are not publicly available systems, which prevents a straightforward evaluation. In order to achieve meaningful results, we used the native dataset of the respective system, referred to in the following as $\textsc{Dataset}_{\text{NewsSpike}}$ and $\textsc{Dataset}_{\text{Heady}}$. We not only used the texts from these datasets, but also the provided results of standard preprocessing components, i.e., sentence splitting and tokenization, entity recognition and linking, as well as chunking or parsing. We then applied Idest on top of the respective system's pattern-extraction component, i.e., *ReVerb* for NewsSpike and *Compression* for Heady. In summary, we evaluated clusterings produced with the following setups, with the first and third one being generated by the respective third-party system on its native dataset and the remaining three setups featuring the proposed system:

1. System NewsSpike with ReVerb extractions from $\textsc{Dataset}_{\text{NewsSpike}}$.
2. System Idest with ReVerb extractions from $\textsc{Dataset}_{\text{NewsSpike}}$.
3. System Heady with Compression extractions from $\textsc{Dataset}_{\text{Heady}}$.
4. System Idest with Compression extractions from $\textsc{Dataset}_{\text{Heady}}$.
5. System Idest with Compression extractions from $\textsc{Dataset}_{\text{NewsSpike}}$.

Note that the fifth item is a compromise between the three systems, since it runs Idest on the Heady-style extractions from the data associated with NewsSpike.

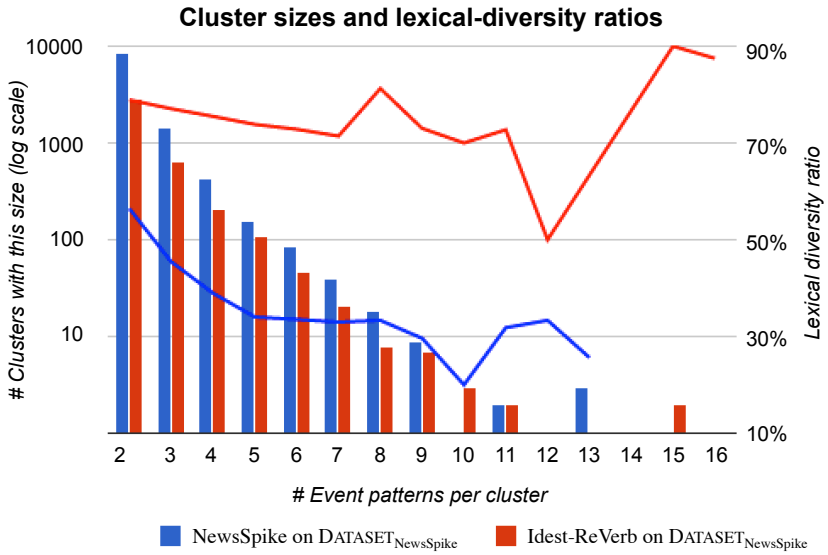


Figure 5.3 – Quantitative comparison of NewsSpike and Idest clusters (1 of 2). This chart depicts the number of clusters with a particular size and the lexical-diversity ratio of the clusters generated from NewsSpike and Idest with the ReVerb extractions as input. Columns correspond to primary y-axis, lines go with secondary y-axis.

5.5.1 Quantitative Analysis

NewsSpike vs. Idest-ReVerb, both on DATASET$_{\text{NewsSpike}}$ In order to evaluate the performance of the factor-graph-based method and the neural-network method on exactly the same EECs, this paragraph compares the clustering models that were output by NewsSpike and Idest when using the same set of extractions. As input we used the dataset DATASET$_{\text{NewsSpike}}$, released by Congle Zhang and Weld (2013)[49], which contains 546,713 news articles, from which 2.6 million ReVerb extractions were reportedly produced. 84,023 of these are grouped into the 23,078 distributed EECs, based on mentions of the same entities on the same day.

Figure 5.3 shows a comparative analysis of the two sets of clusters. We observe two results in this chart. First, Idest generates fewer clusters for every cluster size than NewsSpike. Here, cluster size is defined as the number of paraphrases in a cluster. This result means that for a given random pattern, NewsSpike is likely to provide more paraphrases than Idest. We have also computed a lexical-diversity ratio, defined as the percentage of root-verb lemmas in the patterns of a cluster that are unique. This metric captures whether a cluster mainly contains the same verb with different inflections or modifiers, or whether it contains different predicates. The second observation is that

[49] `http://www.cs.washington.edu/node/9473`, last access: 2017-04-24.

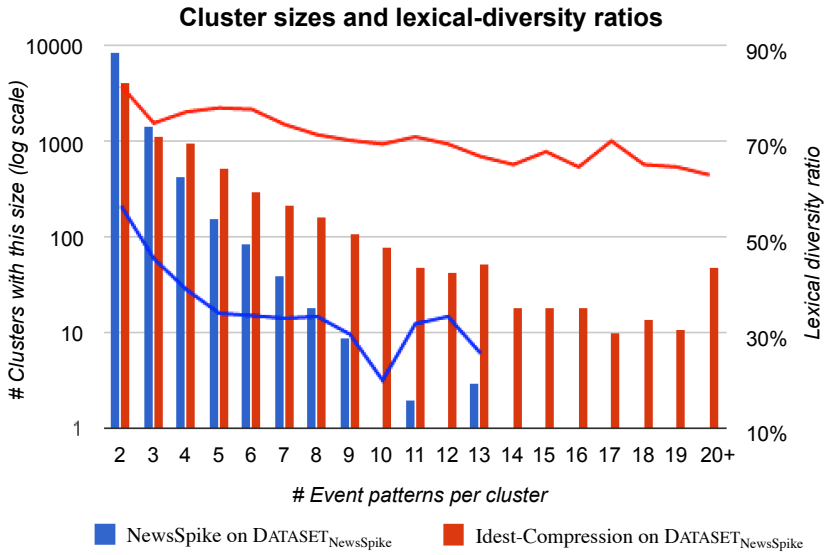


Figure 5.4 – Quantitative comparison of NewsSpike and Idest clusters (2 of 2). This chart depicts the number of clusters with a particular size and the lexical-diversity ratio of the clusters generated from NewsSpike and Idest with compression-based pattern extraction. Columns correspond to primary y-axis, lines go with secondary y-axis.

Idest generates clusters with a greater lexical diversity. These results make intuitive sense, as a global model should be able to produce more aggregated clusters by merging patterns originating from different EECs, which eventually results in fewer clusters with a respective higher lexical diversity. A higher lexical diversity may be a signal of richer paraphrases or noisier clusters. The manual evaluation in Subsection 5.5.2 addresses this question by comparing the quality of the clusterings.

NewsSpike vs. Idest-Compression, both on $\text{DATASET}_{\text{NewsSpike}}$ Figure 5.4 compares NewsSpike's clusters against Idest clusters obtained using sentence compression instead of ReVerb for extracting patterns. Both systems were trained on the same set of input news. Using sentence compression, the total number of extracted patterns was 321,130, organized in 41,740 EECs. We can observe that Idest produced larger clusters than NewsSpike. For cluster sizes larger or equal to 4, this configuration of Idest produced more clusters than NewsSpike. At the same time, lexical diversity remained consistently at much higher levels, well over 60%.

Idest-Compression on $\text{DATASET}_{\text{NewsSpike}}$ vs. Idest-Compression on $\text{DATASET}_{\text{Heady}}$ Next, we evaluated the impact of the size of training data by producing a clustering from embedding vectors trained from a much larger dataset. We used the Heady

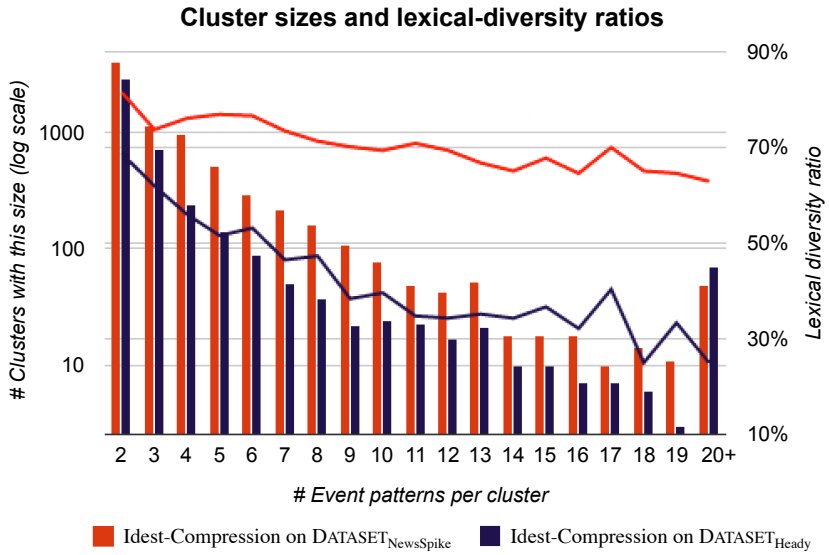


Figure 5.5 – Quantitative comparison of clusters from Idest on two datasets. This chart depicts the number of clusters with a particular size and the lexical-diversity ratio of the clusters generated from Idest with compression-based pattern extraction, using only the 500,000 NewsSpike articles, or the large dataset. Columns correspond to primary y-axis, lines go with secondary y-axis.

crawl of news collected between 2008 and 2014. Using sentence compression, hundreds of millions of extractions were produced; which is at least two orders of magnitude higher than were generated from $\text{DATASET}_{\text{NewsSpike}}$. In order to keep the dataset at a reasonable size and aiming at producing a model of comparable size to the other approaches, we applied a filtering step in which we removed all the event patterns that were not extracted at least five times from the dataset. After this filtering, 28,014,423 extractions remained, grouped in 8,340,162 non-singleton EECs.

Figure 5.5 compares the resulting clusterings. In the setting with more data, clusters were generally smaller and showed less lexical variability. We believe that this is due to the removal of the long tail of low-frequency and noisy patterns. Indeed, while high lexical variability is desirable, it can also be a sign of noisy, unrelated patterns in the clusters. The cohesiveness of the clusters, which we evaluate in Subsection 5.5.2, must also be considered to tell apart constructive and destructive lexical variability.

Heady on $\text{DATASET}_{\text{Heady}}$ Heady produces a soft-clustering from a generative model, and expects the maximum number of clusters to be provided beforehand. The model then tries to approximate this number. In our experiments, 5,496 clusters were finally generated. One weak point of Heady that has already been mentioned above is that

low-frequency patterns do not have sufficient evidence and noisy-or Bayesian networks tend to discard them. In our experiments, only 4.3% of the unique extracted patterns actually ended up in the final model.

5.5.2 Qualitative Analysis

The clusters obtained with different systems and datasets were evaluated by five expert raters with respect to three metrics, according to the following rating workflow:

1. The rater is shown a cluster and is asked to annotate which patterns are meaningless or unreadable[50]. This provides us with a *readability* score, which measures both the quality of the extraction algorithm and the ability of the method to filter out noise.

2. Next, the rater examines whether there is a majority theme in the cluster, defined as having at least half of the readable patterns refer to the same real-world event. If this is not the case, the cluster is annotated as noise. We call this metric *cohesiveness*.

3. If a cluster is cohesive, the rater is finally asked to indicate which patterns express the main theme and which ones are unrelated to it. The third metric, *relatedness*, is defined as the percentage of patterns that are related to the main cluster theme. All the patterns in a non-cohesive cluster are automatically marked as unrelated.

The inter-annotator agreement on the three metrics, measured as the intraclass correlation (ICC), was strong (Cicchetti, 1994; Hallgren, 2012). More precisely, the observed ICC scores with 0.95 confidence intervals were 0.71 [0.70, 0.72] for cohesiveness, 0.71 [0.70, 0.73] for relatedness and 0.66 [0.64, 0.67] for readability. For the evaluation, from each model we selected enough clusters to achieve an overall size (number of distinct event patterns) comparable to NewsSpike's. For Heady and Idest, the stopping condition in Algorithm 5.1 (p. 113) was modified accordingly.

Table 5.1 (p. 119) shows the annotation outcome. The column block on the left lists the system, the employed pattern-extraction algorithm, the dataset used to train the model, and the average size of the resulting clusters. The column block on the right reports the values of the quality metrics. As expected, using a global model that can merge patterns from different EECs into single clusters and using the whole news dataset both led to larger clusters. At the same time, we observe that using ReVerb extractions generally

[50] In the data released by NewsSpike, ReVerb patterns are lemmatized, but the original inflected sentences are also provided. We have restored the original inflection of all the words to make those patterns more readable for the raters.

System	Extraction	Data	Size	Coh(%)	Rel(%)	Read(%)
Heady	Compression	$\text{DATASET}_{\text{Heady}}$	12.66^{bcd}	$34.40^{!}$	$27.70^{!}$	60.70
NewsSpike	ReVerb	$\text{DATASET}_{\text{NewsSpike}}$	$3.40^{!}$	56.20^{ac}	66.42^{acd}	56.66
Idest	ReVerb	$\text{DATASET}_{\text{NewsSpike}}$	3.62^{b}	40.00	47.10^{a}	65.16^{b}
Idest	Compression	$\text{DATASET}_{\text{NewsSpike}}$	5.54^{bc}	50.31^{ac}	46.58^{a}	66.04^{b}
Idest	Compression	$\text{DATASET}_{\text{Heady}}$	44.09^{*}	87.93^{*}	68.28^{acd}	80.13^{*}

Table 5.1 – Qualitative evaluation results of NewsSpike, Heady, and Idest. Results were averaged over all the clusters produced by each configuration listed.

Quality metrics:

- *Size*: average cluster size;
- *Coh*: cohesiveness;
- *Rel*: relatedness;
- *Read*: readability.

Statistical significance, 0.95 confidence intervals, bootstrap resampling:

- ^{a}: better than Heady;
- ^{b}: better than NewsSpike;
- ^{c}: better than Idest-ReVerb on $\text{DATASET}_{\text{NewsSpike}}$;
- ^{d}: better than Idest-Compression on $\text{DATASET}_{\text{NewsSpike}}$;
- *: better than all others;
- $^{!}$: worse than all others.

led to smaller clusters. This is probably due to the fact that ReVerb produced fewer extractions than the sentence compressor does on the same input documents.

On ReVerb extractions, NewsSpike outperformed Idest in terms of cohesiveness and relatedness, but NewsSpike's low cluster size and lexical diversity make it difficult to prefer any of the two models only with respect to the quality of the clusters. On the other hand, the patterns retained by Idest-ReVerb on $\text{DATASET}_{\text{NewsSpike}}$ were generally more readable (65.16 vs. 56.66). On the same original news data, using Idest with sentence compression produced comparable results to Idest-ReVerb on $\text{DATASET}_{\text{NewsSpike}}$, cohesiveness being the only metric that improved significantly.

More generally, in terms of readability all the models that rely on global optimization (i.e., all but NewsSpike) showed better readability than NewsSpike. This supports the intuition that global models are more effective in filtering out noisy extractions. Also, the more data was available to Idest, the better the quality across all metrics. Idest model using all data, i.e., Idest-Compression on $\text{DATASET}_{\text{Heady}}$, was significantly better (with 0.95 confidence) than all other configurations in terms of cluster size, cohesiveness and pattern readability. Pattern relatedness was higher, though not significantly better, than

NewsSpike, whose clusters were on average more than ten times smaller.

We did not evaluate NewsSpike on the whole news dataset. Being a local model, extending the dataset to cover six years of news would only lead to many more EECs. It would not affect the reported metrics as each final cluster would still be generated from one single EEC and it cannot benefit from the larger dataset as the global models do. It is interesting to see that, even though they were trained on the same data, Idest outperformed Heady significantly across all metrics, sometimes by a very large margin. Given the improvements on cluster quality, it would be interesting to evaluate Idest performance on the headline-generation task for which Heady was initially designed, but we leave this as future work.

5.5.3 Cluster Example and Typical Errors

Figure 5.6 (p. 121) shows an excerpt[51] from one of Idest's generated clusters from compressions on $\text{DATASET}_{\text{Heady}}$, in the way it was presented to human raters, i.e., with instantiated entity placeholders. The event reported on in this cluster is that two celebrities are dating. Despite this inherently noisy domain, Idest was capable of identifying paraphrases which exhibit great lexical and syntactic variety.

This cluster also helps to exemplify some of the system's typical errors, via the erroneous patterns marked with a leading asterisk. One such type of error is the false clustering of patterns together with their negated forms or antonymic patterns. For example, pattern (d) negates pattern (a) via adding a `not` and pattern (i) is an antonym of pattern (m). This error is in part caused by the doubtful truth value of some of the source documents, namely yellow-press news articles full of rumors and speculations. It is quite possible that two contradictory statements are published about the same two people on the same day in two different news sites. Another error type is the use of patterns extracted from questions, i.e., pattern (r). The impact of these two error types might be reduced by further work on the engineering level, meaning more fine-tuning with respect to the same-timestamp constraint or additional pre-clustering constraints on the level of news articles could help. Additionally, pre-filtering heuristics based on sentence types would allow to identify interrogative patterns.

[51] Patterns with only slight syntactic or lexical differences have been removed for brevity.

(a) Charlene_Choi is dating William_Chan
(b) Wilmer_Valderrama and Demi_Lovato are dating
(c) Taylor_Swift and Conor_Kennedy are an item
(d) * Zach_Braff is not dating Taylor_Swift
(e) Victor_Garber is in a relationship with Rainer_Andreesen
(f) Taylor_Lautner and Marie_Avgeropoulous are a couple
(g) Sofia_Hayat is rumoured to be dating Rohit_Sharma
(h) Taylor_Swift is seeing Conor_Kennedy
(i) * Taylor_Lautner and Selena_Gomez are just friends
(j) Vidya_Balan is romancing Naseeruddin_Shah
(k) Serena_Williams has been spotted with Rapper_Common
(l) Zayn_Malik may be dating Perrie_Edwards
(m) Natalie_Portman and Sean_Penn are more than friends
(n) Tom_Daley has opened up about his relationship with Dustin_Lance_Black
(o) Tracy_Young and Kim_Zolciak are together
(p) Zac_Efron has been linked to Halston_Sage
(q) Zayn_Malik is in love with Perrie_Edwards
(r) * are Lindsay_Vonn and Tiger_Woods really dating
(s) Rza opened up about his relationship with Kanye_West
(t) Michelle_Rodriguez is in a relationship with model Cara_Delevingne
(u) Kate_Upton's hooking up with Sean_"Diddy"_Combs
(v) Zac_Efron and Vanessa_Hudgens are hooking up
(w) Taylor_Swift and Harry_Styles went public
(x) Soulja_Boy has professed his love for Kim_Kardashian
(y) Young_Jeezy confirmed his relationship with Keyshia_Cole
(z) Zayn_Malik is said to be dating Stephanie_Davis

Figure 5.6 – Example cluster from Idest containing patterns relevant for *dating*. The patterns have been instantiated in the entity placeholders for readability.

While in general Idest does very well distinguish between actual paraphrases and event patterns which are just topically related, such misclassification does occasionally occur.[52] Further errors in Idest clusters can be traced back to erroneous preprocessing components, e.g., when the pattern extraction algorithm fails to extract all relation-indicating information from the source sentence.

5.6 Summary

This chapter proposed Idest, a new approach based on neural networks to map event patterns into an embedding space. We showed that it can be used to construct high-quality pattern clusters based on neighborhood in the embedding space. The two open-IE systems NewsSpike and Heady inspired the design of Idest. On a small dataset, Idest produces comparable results to NewsSpike, but its main strength lies in its ability to generalize extractions into a single global model. It scales to hundreds of millions of news articles, leading to larger clusters of event patterns with significantly better coherence and readability. When compared to Heady, Idest outperforms it significantly on all the metrics explored.

Finally, while Chapters 3 and 4 presented methods for efficient pattern-based information extraction from the web when fixed relation schemata are available, this chapter introduced a method for less restricted settings where the type of information to be extracted is not known a-priori. In the following chapters, we address the question how large sets of generated patterns can be composited into a more generally useful resource, with potential applications other than IE.

[52] The cluster shown in Figure 5.6 is not affected by this issue.

Chapter 6

Sar-graphs: A Linked Language Resource[53]

Contents

[53] This chapter presents results from joint work with Aleksandra Gabryszak, Leonhard Hennig, Hong Li, Andrea Moro, Hans Uszkoreit, Dirk Weissenborn, and Feiyu Xu.

6.1 Introduction

In recent years, large-scale knowledge resources have been extensively used in academic research, community-driven projects, and industrial development (Section 2.5). We can distinguish at least two types of knowledge resources in language technology: (a) those that store factual information about entities and (b) those that represent general linguistic knowledge. In this chapter, we propose a third type that connects the former two with one another. Instances of this kind are *graphs of semantically-associated relations* (sar-graphs), whose purpose is to link semantic relations from factual KBs with their linguistic representations in human language.

Prominent examples of the factual resource kind include Freebase (Bollacker et al., 2008), YAGO (Suchanek et al., 2007), DBpedia (Lehmann et al., 2015), WikiData (Vrandečić and Krötzsch, 2014), and Google's Knowledge Graph (Singhal, 2012) and Knowledge Vault (Dong et al., 2014). They often exhibit an implicit or explicit graph structure, i.e., they are vast networks which store entities and their semantic types, properties, and relations. Because of this structure, some KBs are also referred to as knowledge graphs. Non-proprietary resources are commonly released in a format that fosters the linking to other repositories, usually in accordance with linked-data principles (Bizer et al., 2009a).

A parallel development is the emergence of large-scale resources with a focus on language. ConceptNet (Speer and Havasi, 2013), BabelNet (Navigli and Ponzetto, 2012), UWN (Melo and Weikum, 2009), and UBY (Gurevych et al., 2012) are all examples of this resource kind. In comparison to (world-)knowledge graphs, the representations and semantic models of linguistic resources are more diverse: ConceptNet makes use of natural-language representations for modeling common-sense information. BabelNet integrates entity information from Wikipedia with word senses from WordNet (Fellbaum, 1998), as well as with other resources such as Wikidata and Wiktionary. UWN automatically builds a multilingual wordnet from various resources, similar to UBY, which integrates multiple resources via linking on the word-sense level. Resources of this second resource type are, too, published as linked data, then sometimes called linguistic linked-open-data (Chiarcos et al., 2013).

While the linking of resources increases coverage of the respective domains, in general, it does not explicitly connect semantic relations from knowledge graphs with their linguistic representations. Sar-graphs fill this gap. A sar-graph can be thought of as a bridge between language and the information encoded in a knowledge graph, a bridge that characterizes how a language can express instances of one or several relations.

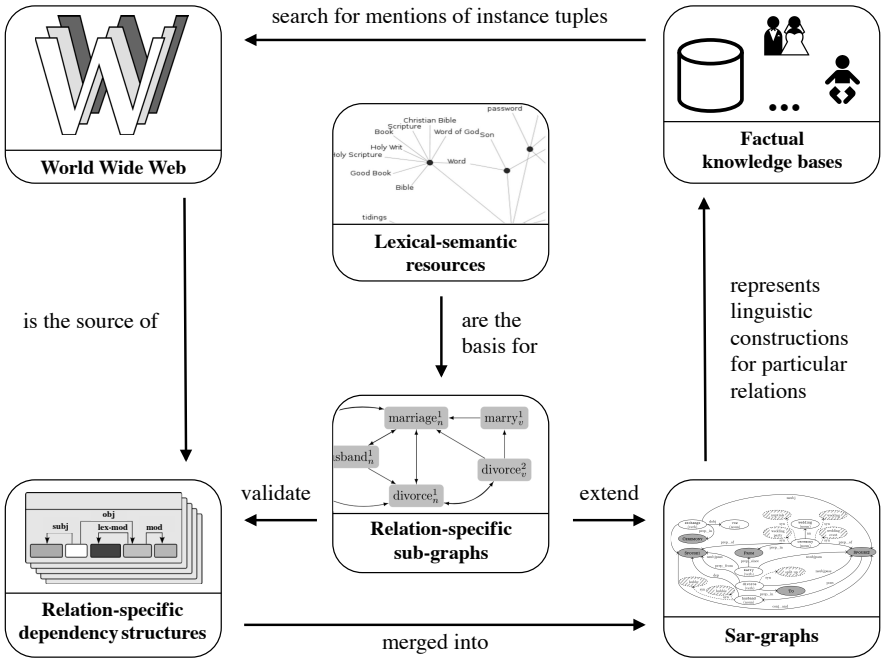


Figure 6.1 – Relation of sar-graphs to other knowledge resources.

RE is one of the central technologies contributing to the automatic creation of fact databases, and, at the same time, it also benefits from the growing number of available factual resources by using them for automatic training and improvement of extraction systems. Consequently, we build on our own work from Chapter 3 for the purpose of collecting linguistic constructions for the sar-graphs. The distantly-supervised nature of this construction methodology requires means for the automatic confidence estimation for the extracted linguistic structures. For this, we rely on the filtering methods proposed in Chapters 3 and 4, which we complement with a large manual verification effort for linguistic structures. Combining manual and automatic quality assessment allows us to construct sar-graphs at varying degrees of coverage of linguistic expressions, by trading precision for recall.

A prerequisite of the filter from Chapter 4 is the disambiguation of content words with respect to lexical-semantic repositories. This step is also an important aspect of the construction of sar-graphs, because it generalizes from the lexical level (content words) to word senses. In addition to making sar-graphs more adjustable to the granularity needs of potential applications, this positions sar-graphs as a link hub between a number of formerly independent resources. Figure 6.1 depicts how sar-graphs fit into the resource landscape: Sar-graphs represent linguistic constructions for semantic relations from factual KBs and

incorporate linguistic structures extracted from mentions of knowledge-graph elements in free texts. Furthermore, they anchor this information in lexical-semantic resources.

In our experiments, we continued our earlier work and created sar-graphs for 25 relations, which underlines the feasibility of the proposed method. We believe that sar-graphs will prove to be a valuable resource for numerous applications other than RE,[54] such as adaptation of parsers to special recognition tasks, text summarization, language generation, query analysis, and even interpretation of telegraphic style in highly elliptical texts as found in SMS, Twitter, headlines, or brief spoken queries. In addition to coupling sar-graphs to lexical-semantic resources and to knowledge graphs, we designed a two-fold linking strategy to FrameNet. FrameNet (Baker et al., 1998) is a linguistic resource that goes beyond the level of lexical items and that provides fine-grained semantic relations of predicates and their arguments. Only a few works considered the linking of FrameNet to other resources (Scheffczyk et al., 2006; Bonial et al., 2013; Aguilar et al., 2014), and none of them linked FrameNet to knowledge-graph relations and, at the same time, extended it with linguistic patterns, which is what we propose to do.

6.2 Idea and Definition

Sar-graphs are directed multi-graphs that contain linguistic knowledge at the syntactic and lexical-semantic level. More formally, a sar-graph is a tuple

Equation 6.1
$$G_{r,l} = (V, E, s, t, f, A_f, \Sigma_f),$$

where V is the set of vertices, E is the set of edges, $s : E \mapsto V$ maps edges to their start vertex, and $t : E \mapsto V$ maps edges to their target vertex. Both vertices and edges are labeled via the labeling function f, which associates them with sets of features (i.e., attribute-value pairs):

Equation 6.2
$$f : V \cup E \mapsto \mathcal{P}(A_f \times \Sigma_f)$$

where $\mathcal{P}(\cdot)$ constructs a powerset, A_f is the set of attributes (i.e., attribute names) which vertices and edges may have, and Σ_f is the value alphabet of the features (i.e., the set of possible attribute values for all attributes).

The information in one instance of a graph is specific to a given language l and target relation r. The function of sar-graphs is to represent the linguistic constructions a language l provides for reporting instances of r or for just referring to such instances. A vertex

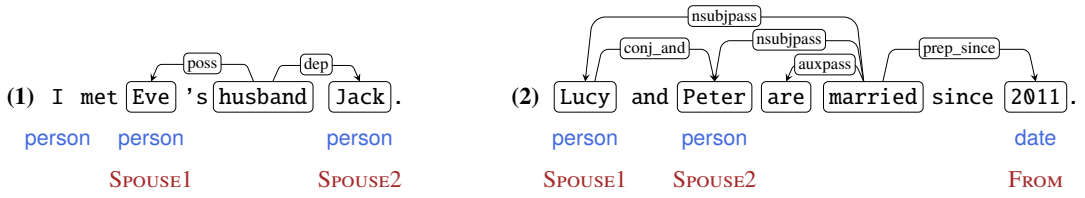
[54] Due to the integration of linguistic patterns, RE is an obvious application of sar-graphs.

preimage(f)	A_f	**Example for** Σ_f
V from lexical tokens	word form, word lemma, word class, word sense	`married`, to `marry`, verb, bn:00085614v
V from entity mentions	entity type, semantic role	person, SPOUSE2
E from syntactic parsing	dependency labels	nsubjpass
E from resource linking	lexical-semantic relation	*synonym*
$V \cup E$	frequency in training set	2
$V \cup E$	identifiers for sentences & dependency structures	[sent:16, sent:21], [pat:16#1, pat:21#2]

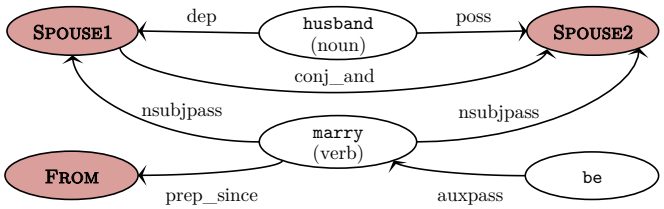
Table 6.1 – Names and example values for attributes of sar-graph elements.

$v \in V$ corresponds to a word in such a construction. The features assigned to a vertex via the labeling function f provide information about lexico-syntactic aspects (word form and lemma, also word class), lexical semantics (word sense), and semantic properties (global entity identifier, entity type, semantic role in the target relation). Additionally, they provide statistical and meta information (e.g., frequency). The linguistic constructions are modeled as sub-trees of dependency-graph representations of sentences, congruent with the kind of information that is stored in the linguistic patterns of Chapters 3–5. Consequently, edges $e \in E$ are labeled with dependency tags via f, in addition to frequency information. Table 6.1 lists possible attributes for vertices and edges.

What separates the collected linguistic patterns from their representation in sar-graphs is that individual dependency structures may or may not be present in a sar-graph as disjunct trees, i.e., we merge constructions and parts thereof. The joint representation of common paths of linguistic expressions allows for a quick identification of dominant phrases and the calculation of frequency distributions for sub-trees and their combinations. This merging step is not destructive, the information about the linguistic structures found in original sentences is still available. We believe that for language expressions, an exhaustive, permanent merging does not make sense, as it results in a loss of language variety which we aim to capture. The merging process is implemented with a conservative default strategy, which cautiously connects dependency constructions at their argument positions, followed by a customizable second step, which further superimposes nodes and paths in a non-destructive manner. We describe this two-step process in Section 6.3.



a) Two sentences with relation mentions.



b) A sar-graph.

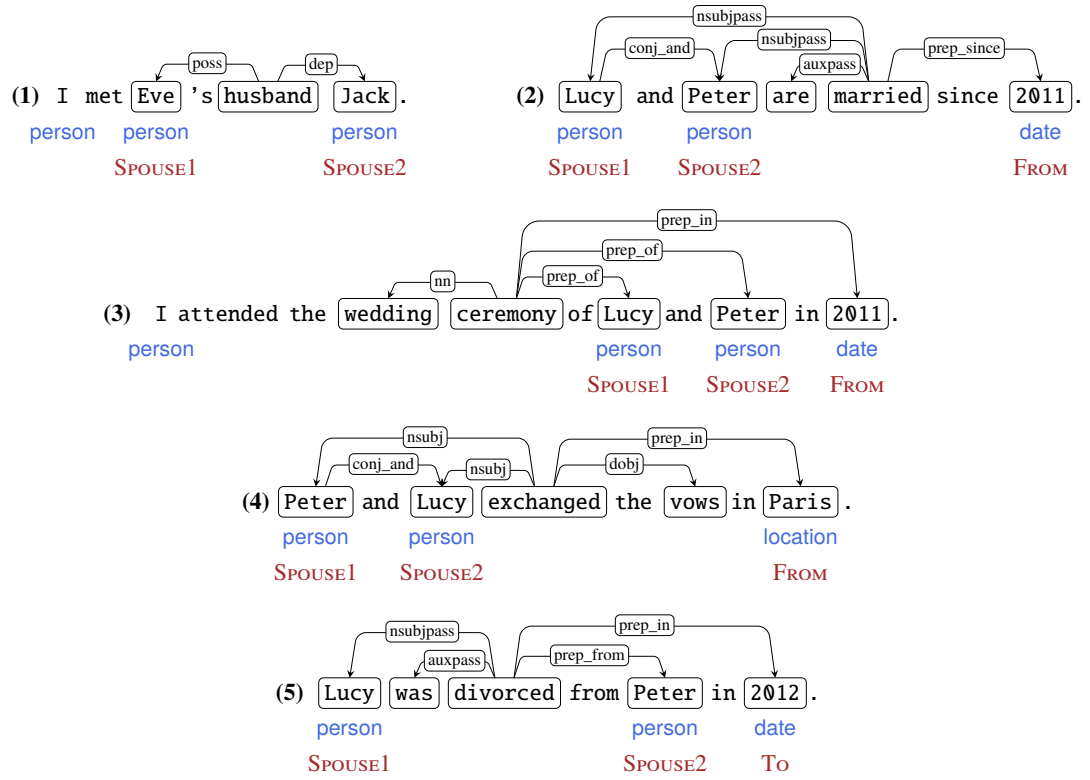
Figure 6.2 – Example of a sar-graph, generated from two English sentences. Target relation is *marriage*. The sar-graph connects the dependency structures via their shared SPOUSE arguments. The dependency labels are a compressed variant of the Stanford label set (de Marneffe and Manning, 2008).

From Individual Constructions to Sar-graphs If a given language l only provided a single construction to express an instance of r, then the dependency structure of this construction would form the entire sar-graph. Yet, if the language offered alternatives to this construction, i.e., paraphrases, their dependency structures would also be added to the sar-graph. The individual constructions superimpose one another based on shared properties and labels of vertices and edges. Specifically, we merge

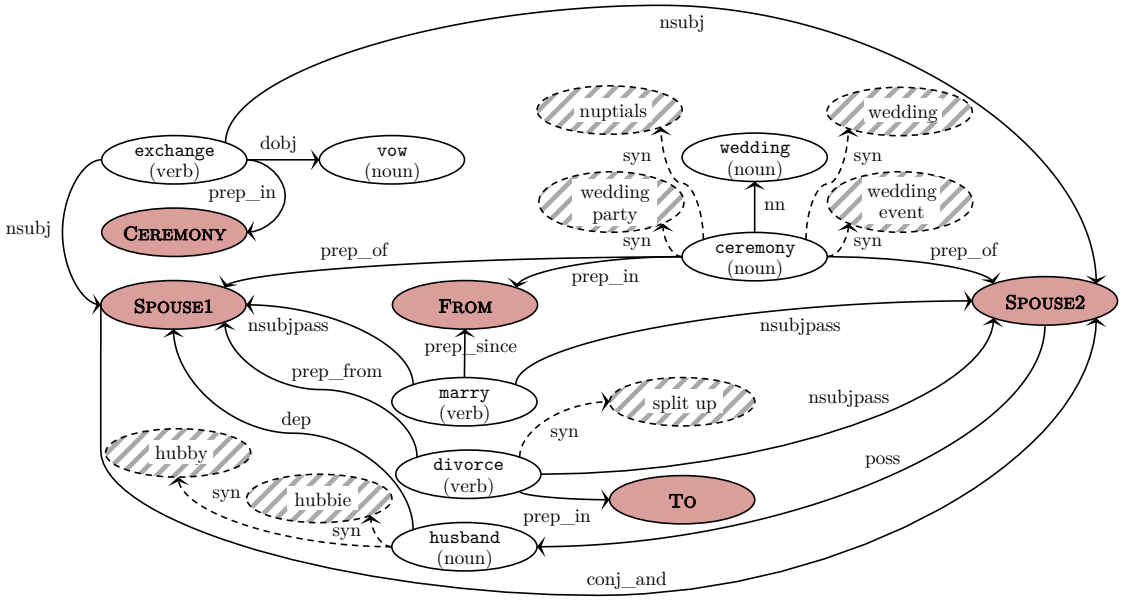
- vertices without a semantic role based on their word lemma or entity type,
- vertices with argument roles with respect to their semantic role in the target relation, and
- edges on the basis of dependency labels.

Our data-driven approach to the creation of sar-graphs integrates both constructions that include all relation arguments as well as those that mention only a subset thereof. As long as constructions indicate an instance of the target relation, they might be relevant for applications, even if they are not true paraphrases of the full expressions.

A sar-graph for the two sentences in Figure 6.2a, both mentioning projections of the target relation, is presented in Figure 6.2b. The first sentence connects the spouses with a possessive construction, while the second sentence uses a conjunction; in addition, the



a) Five sentences with relation mentions.



b) A sar-graph.

Figure 6.3 – More complex example of a sar-graph, constructed from five *marriage* relation mentions. This graph also includes lexical-semantic information (dashed vertices and edges) obtained by linking content words to a lexical-semantic resource.

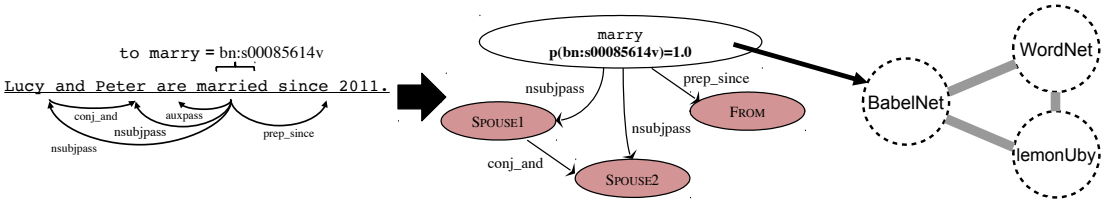


Figure 6.4 – A minimal sar-graph disambiguation example, consisting of a single pattern, where the lexical vertex *marry* is disambiguated and linked to BabelNet, UBY, and WordNet.

second sentence provides the marriage date. We integrate the extracted structures from these sentences by merging their shared arguments, namely, SPOUSE1 and SPOUSE2. As a result, the graph in Figure 6.2b (p. 128) contains a path from SPOUSE1 to SPOUSE2 via the node `husband` for sentence (1) and an edge *conj_and* from SPOUSE1 to SPOUSE2 for sentence (2). Note that the graph contains two types of vertices: argument nodes labeled with their role, and lexical-semantic nodes labeled with their lemma and POS tag.

Figure 6.3 (p. 129) illustrates the structure and content of a more complex sar-graph. We extend the previous example with three new sentences, which provide alternative linguistic constructions as well as the arguments CEREMONY and TO. The graph additionally includes the paraphrases `exchange vows`, `wedding ceremony of`, and `was divorced from`. Note that both sentences (2) and (4) utilize a *conj_and* to connect the SPOUSES. The sar-graph includes this information as a single edge, but we can encode the frequency information as an edge attribute.

Word-level Linking The lexico-syntactic and semantic information specified in sar-graphs is augmented with lexical-semantic knowledge by disambiguating all content words in the context of their original sentences. This results in a labeling of vertices with sense identifiers and additional synonymous surface forms, and also implicitly adds lexical-semantic links among words already contained in the sar-graph. In Figure 6.3b, additional surface forms are illustrated by dashed vertices and edges. For example, for the vertex representing the lemma `husband`, the colloquial synonyms `hubby` and `hubbie` are listed. Technically, each content-word vertex is associated with a distribution over synset assignments, corresponding to local disambiguation decisions in the source sentences of the constructions. In the experiment of Section 6.4, we disambiguate against the sense entries in BabelNet. Since BabelNet provides links for its content to other lexical-semantic resources, namely, WordNet and UBY, we also implicitly integrate sar-graphs with these resources. Figure 6.4 illustrates this multifaceted word-level linking.

Less Explicit Relation Mentions A key property of sar-graphs is that they store linguistic structures with varying degrees of explicitness with respect to the factual relations. We include constructions that refer to only a part or aspect of the relation if they would normally be seen as sufficient evidence of an instance (Example 6.1) even if there could be contexts in which this implication is canceled (Example 6.2). Other constructions in sar-graphs refer to relations that entail the target concept without being part of it (Examples 6.3 and 6.4). Finally, there are constructions that refer to semantically connected relations which by themselves might not be used for a safe detection of instances of r, but that could be employed for recall-optimized applications or for a probabilistic detection-process that combined several pieces of evidence (Example 6.5). Example 6.6 shows an instance of exclusively probabilistic entailments that are caused by social conventions or behavioral preferences. All of these kinds of constructions are intentionally included in sar-graphs.

Example 6.1 `Joan and Edward exchanged rings in 2011.`

Example 6.2 `Joan and Edward exchanged rings during the rehearsal of the ceremony.`

Example 6.3 `Joan and Edward celebrated their 12th wedding anniversary.`

Example 6.4 `Joan and Edward got divorced in 2011.`

Example 6.5 `I met her last October at Joan's bachelorette (engagement) party.`

Example 6.6 `Two years before Joan and Paul had their first child, they bought a larger home.`

6.3 Construction

In this section, we describe in more detail how sar-graphs are created. Figure 6.5 (p. 132) outlines this process. The first part of the method builds on the approach from Chapter 3 to collect a set of linguistic structures for a particular language and relation. Then, WSD is applied to the gathered structures, which produces a mapping of content-word vertices to word senses. Furthermore, the structures are associated with confidence values obtained via the filters from Chapters 3 and 4. In addition, we employ manual verification of linguistic structures. The final phase of the method integrates the constructions into

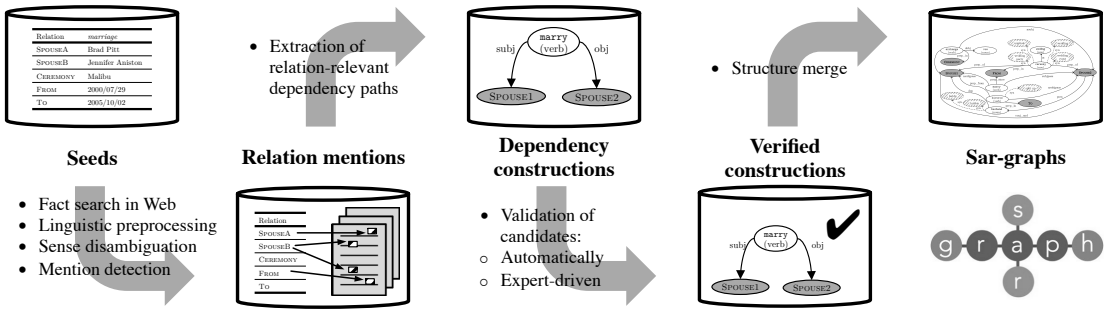


Figure 6.5 – Outline of sar-graph construction. Arrows correspond to processing steps, while boxes show intermediate results.

a sar-graph. More formally, the creation of a sar-graph $G_{r,l}$ (for target relation r and language l) involves the following four steps:

a) Given a set of seed instances $\mathcal{I}_r$ of r, acquire a set of textual mentions $\mathcal{M}_{r,l}$ of instances i for all $i \in \mathcal{I}_r$ from a text corpus.

b) Extract candidate dependency constructions $\mathcal{D}_{r,l}$ from the dependency trees of elements of $\mathcal{M}_{r,l}$.

c) Disambiguate word senses in the candidate structures $d \in \mathcal{D}_{r,l}$ and assess the quality of the structures, either automatically or via human-expert-driven validation, yielding a derived set $\mathcal{D}'_{r,l}$ of acceptable dependency constructions.

d) Merge elements of $d \in \mathcal{D}'_{r,l}$ to create the sar-graph $G_{r,l}$.

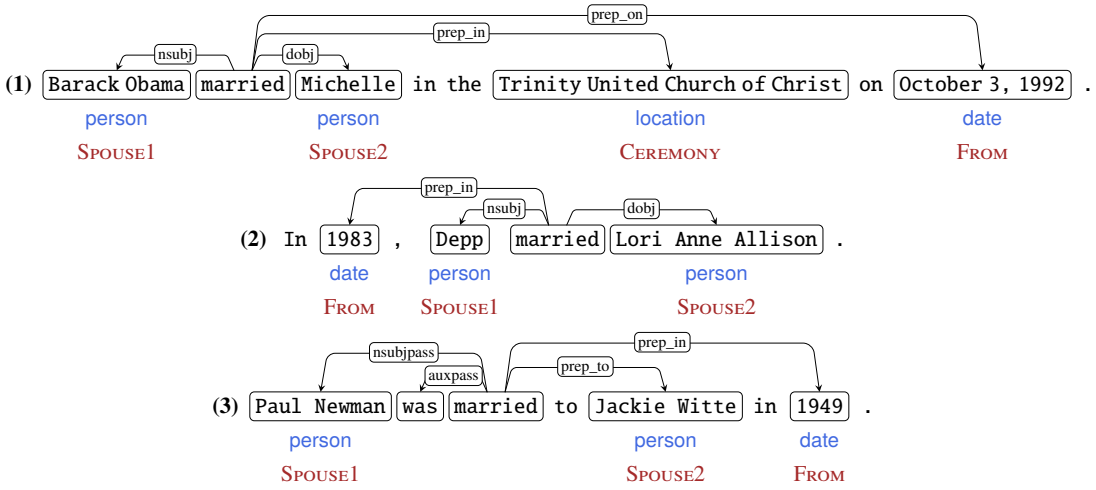
Note that steps (a)–(c) overlap with the method description in Section 3.3. In the following we discuss step (d) in more detail.

Merging of Dependency Structures A sar-graph is created by the superimposition of a set of validated dependency constructions, i.e., the output of step (c). We follow a technically straightforward approach which gradually merges dependency constructions into larger graphs, based on the equality of properties of the graph elements. Initially, this process creates a graph by only merging argument nodes, while retaining the independence of the remainder of the structures. Algorithm 6.1 (p. 133) presents two pseudocode functions that outline this step. Function **createSarGraph** adds structures to the graph by iterating over the contained edges. Whenever a node is to be added to the graph, two aspects have to be considered: (*Text continues on page 136.*)

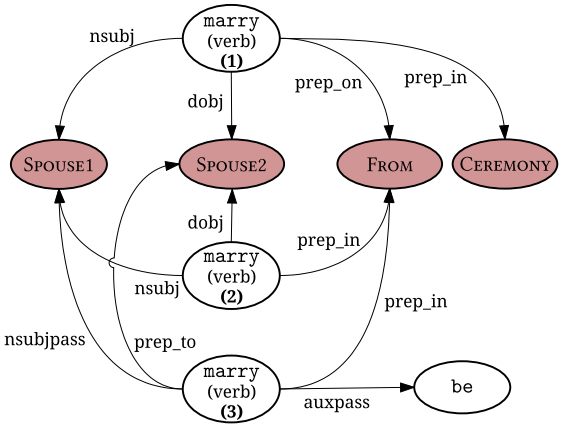
```
FUNCTION NAME:  createSarGraph
INPUT:          DependencyConstruction[] dcs
OUTPUT:         a SarGraph
    // Initialize graph.
 1  SarGraph sg ← (V=∅, E=∅, s=∅, t=∅, f=∅, A_f, Σ_f)
 2  for each dc ∈ dcs :
        // Each dependency construction is a weakly connected,
        // directed simple graph.
 3      for each edge e in dc from n_1 to n_2 :
 4          e' ← new edge
 5          sg.E ← sg.E ∪ { e' }
 6          update function sg.s: Set sg.s(e') to result of addNode(sg, n_1)
 7          update function sg.t: Set sg.t(e') to result of addNode(sg, n_2)
 6          update function sg.f: Set sg.f(e') to attributes of e
 7  return sg
```

```
FUNCTION NAME:  addNode
INPUT:          SarGraph sg, Node n
OUTPUT:         a Node
 1  if n ∈ sg.V then :
 2      return n
 3  elseif ∃n' ∈ sg.V | n, n' are derived from entity mentions
                       ∧ n, n' share entity type and argument role information then :
 4      update function sg.f: Extend sg.f(n') with attributes of n
 5      return n'
 6  else :
 7      sg.V ← sg.V ∪ { n }
 8      update function sg.f: Set sg.f(n) to attributes of n
 9      return n
10  endif
```

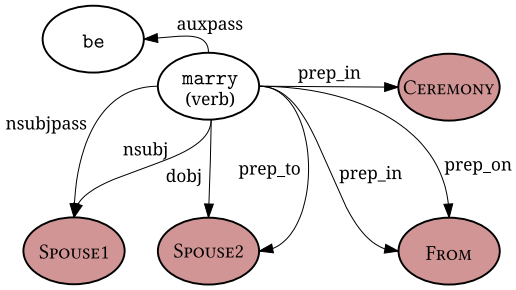
Algorithm 6.1 – Creation of a sar-graph from a set of dependency constructions. f, A_f, Σ_f are defined in Section 6.2. Nodes and edges of dependency constructions have the same attributes as sar-graph elements; see Table 6.1 (p. 127) for a list.



a) Example sentences and dependency structures. Some dependency relations are left out for brevity. This figure uses a collapsed version of the Stanford dependencies (de Marneffe and Manning, 2008).



b) A sar-graph retaining the independence of original structures.



c) A more condensed representation of linguistic phrases.

Figure 6.6 – Two different sar-graph views (b and c) created from the same three sentences (a(1) – a(3)) . Not all properties of nodes and edges are displayed.

```
FUNCTION NAME:  createCondensedSarGraphView
INPUT:          SarGraph sg, Function exampleNodeCompressor,
                Function exampleEdgeCompressor
OUTPUT:         a SarGraph
    // Initialize view on sar-graph.
 1  SarGraph sgView ← (V=∅, E=∅, s=∅, t=∅, f=∅, A_f, Σ_f)
 2  for each edge e ∈ sg.E :
 3      exampleEdgeCompressor(sg, sgView, e,
                              exampleNodeCompressor(sg, sgView, sg.s(e)),
                              exampleNodeCompressor(sg, sgView, sg.t(e)))
 4  return sgView
```

```
FUNCTION NAME:  exampleNodeCompressor
INPUT:          SarGraph sg, SarGraph sgView, Node n
OUTPUT:         a Node
 1  if n ∈ sgView.V then :
 2      return n
 3  elseif n is derived from a lexical token then :
        // Generalize part-of-speech tag of n.
 4      update function sg.f: Replace ("word class", p) ∈ sg.f(n) with ("word class", upcast(p))
 5      if ∃n' ∈ sgView.V | n' is derived from a lexical token
                          ∧ n, n' share word form, word lemma, and word class then :
 6          update function sgView.f: Merge sg.f(n) into sgView.f(n')
 7          return n'
 8      endif
 9  endif
    // Neither is n contained in sgView, nor is there an equivalent node.
10  sgView.V ← sgView.V ∪ { n }
11  update function sgView.f: Set sgView.f(n) to sg.f(n)
12  return n
```

```
FUNCTION NAME:  exampleEdgeCompressor
INPUT:          SarGraph sg, SarGraph sgView, Edge e, Node n_1, Node n_2
OUTPUT:         none
 1  if ∃e' ∈ sgView.E | e' originates from syntactic parsing
                      ∧ sgView.s(e') = n_1 ∧ sgView.t(e') = n_2
                      ∧ e, e' have the same dependency label
 2      update function sgView.f: Merge sg.f(e) into sgView.f(e')
 3  else :
 4      e' ← new edge
 5      sgView.E ← sgView.E ∪ { e' }
 6      update function sgView.s: Set sgView.s(e') to n_1
 7      update function sgView.t: Set sgView.t(e') to n_2
 8      update function sgView.f: Set sgView.f(e') to sg.f(e)
 9  endif
```

Algorithm 6.2 – Construction of a condensed view of a sar-graph, tailored for applications. f, A_f, Σ_f are defined in Section 6.2. In this example, the call **createCondensedSarGraphView**(*sg*, **exampleNodeCompressor**, **exampleEdgeCompressor**) generates a sar-graph suited for manual explorative analysis of linguistic phrases. The produced graph uses a coarse-grained inventory of part-of-speech tags. The function **upcast**() generalizes a given tag, e.g., it maps verb classes (*verb in past tense*, *verb in 3rd person singular present*, ...) to a single base verb class.

(*Continues from page 132.*) First, we verify that the node is processed for the first time by checking whether it is already contained in the graph. Second, we check whether there is a matching argument node present. If the latter holds true, the history information of the currently handled node[55] is merged with the information of the existing node. Finally, if the node is indeed processed for the first time and the node does not correspond to an argument, the node is added to the sar-graph.

To deal with task-specific needs for the granularity of information in a sar-graph, applications can view sar-graphs at varying detail levels. For the task of RE, the coverage of the original patterns is already very high, and merging paths would trade off higher recall with lower precision. Thus, the employed view does not impose any additional merging requirements and is identical to the initially constructed sar-graph. Figure 6.6b (p. 134) illustrates this strategy with a sar-graph constructed from the three example sentences shown in Figure 6.6a (p. 134). The resulting sar-graph resembles the union of the original set of dependency structures, i.e., each path through the graph has a frequency of one.

For analysis purposes, e.g., for carrying out an exploratory analysis of the linguistic expressions used to express a particular target relation, a more condensed representation is advantageous. The pseudocode in Algorithm 6.2 (p. 135) shows the general workflow of the generation of sar-graph views in function **createCondensedSarGraphView**. Functions **exampleNodeCompressor** and **exampleEdgeCompressor** provide a custom implementation for the merging of nodes and edges. Two nodes are combined if they contain the same lexical information, likewise, edges between equal nodes are combined if the dependency labels attached to these edges are the same. In an application where a great number of linguistic expressions will be inspected, we can assume that a user is only interested in a coarse-grained distinction of word classes, which is why **exampleNodeCompressor** generalizes the part-of-speech tags of all lexical nodes. In order to cope with applications which require a different balance of detail vs. generalization of the various sar-graph elements, one only has to provide matching implementations of functions **exampleNodeCompressor** and **exampleEdgeCompressor**. For example, dependency structures could be generalized by merging all vertices that belong to the same synset in a lexical-semantic resource, ignoring differences on the lexical level.

This strategy for building a view on a sar-graph merges all nodes and vertices that are equal according to the above definition. Structures that fully or partially overlap (even with just a single edge or node) are merged. Note that this could mean that in the resulting sar-graph, some of the paths connecting argument nodes are linguistically

[55] This includes identifiers of source sentences and dependency structure as well as statistical information.

invalid. However, since the frequency of dependency paths in the sar-graph corresponds to the number of dependency structures from which they originate, paths with less evidence are spotted easily. Figure 6.6c (p. 134) shows an example sar-graph created with this strategy.

6.4 Experiments and Evaluation

In this section, we report the conducted experiment that compiled sar-graphs for 25 semantic relations. Furthermore, we briefly discuss results from a manual evaluation effort and elaborate on some aspects of the usefulness of sar-graphs.

6.4.1 Created Sar-graphs

To create sar-graphs for the 25 semantic relations from Table 3.1 (p. 60), we proceeded as follows. First, we enriched the Web-DARE patterns with confidence values[56] and with information about word senses (Weissenborn et al., 2015a; b; words were mapped to BabelNet). Then, we fed the patterns to the merging procedure presented in Section 6.3. We employed the more extensive merging view that is implemented by combining Algorithms 6.1 (p. 133) and 6.2 (p. 135). Table 6.2 (p. 138) provides statistics on the created sar-graphs (columns two to four). Naturally, the relations with many available dependency patterns (*marriage*, *organization leadership*) produced larger sar-graphs with more vertices and edges. However, even the smallest pattern set (*organization type*) resulted in a sar-graph of significant size. The sar-graphs for all relations contain in total almost 300k vertices and more than 1.5M edges.

While in Chapters 3 and 4, we showed that automatic means of estimating the quality of dependency structures can, to a large degree, detect erroneous structures, it is close to impossible to completely eliminate noise in an automatic fashion. Therefore, we conducted a manual curation of a sample of structures in order to create sar-graphs for analysis purposes and for potential applications in need of very high precision. Due to the fact that the number of possible structures expressing a given relation is potentially unbounded, a complete manual evaluation would be too resource-intensive. Therefore, we limited this expert-driven quality control to a subset of structures, as chosen by the following process: For each relation, we experimentally determined threshold values for the automatic quality metrics in order to exclude low-confidence and low-frequency structures. Then, we sampled a small set of sentences for each structure and conducted

[56] This is the output of the FO filter, Equation 3.5 (p. 66) and the semantic filter, Algorithm 4.2 (p. 82).

Relation	Full set			Curated subset		
	#struct.	#nodes	#edges	#struct.	#nodes	#edges
award honor	10,522	4,349	18,101	510	303	876
award nomination	1,297	983	3,173	392	369	1,091
country of nationality	59,727	24,554	159,857	560	424	1,265
education	16,809	8,216	39,266	270	233	631
marriage	88,456	24,169	169,774	451	193	584
person alternate name	7,796	6,588	22,917	542	717	1,960
person birth	22,377	10,709	46,432	151	124	319
person death	31,559	14,658	73,069	306	159	425
person parent	45,093	15,156	85,528	387	157	589
person religion	37,086	19,221	113,651	142	196	420
place lived	48,158	20,641	120,239	329	445	1,065
sibling relationship	26,250	13,985	68,132	140	103	260
acquisition	26,986	11,235	64,711	224	268	676
business operation	15,376	10,657	47,116	264	416	876
company end	5,743	4,964	17,413	465	714	1,909
company product relationship	15,902	10,358	47,266	257	421	929
employment tenure	43,454	15,151	92,810	226	131	374
foundation	31,570	13,124	72,320	397	231	708
headquarters	23,690	11,420	54,715	273	220	570
organization alternate name	10,419	8,410	32,137	280	283	720
organization leadership	51,295	17,864	115,296	547	213	717
organization membership	29,220	13,326	76,532	291	262	718
organization relationship	12,014	7,030	32,247	303	317	862
organization type	843	1,445	3,474	264	566	1,168
sponsorship	4,599	4,030	13,813	336	523	1,298
average	26,650	11,690	63,600	332	320	840
sum	666,241	292,243	1,589,989	8,307	7,988	21,010

Table 6.2 – Sar-graphs statistics for the 25 relations from Table 3.1 (p. 60). *#struct.* refers to the number of dependency structures used as input to the merging phase (Section 6.3), i.e., Web-DARE patterns. *# nodes/#edges* corresponds to the number of respective elements in the sar-graphs. *Full set* designates the setting where all Web-DARE patterns regardless of their quality assessment from the filters in Chapters 3 and 4 were processed. The source sentences of the patterns in the *curated subset* were manually verified as to whether the DS assumption holds.

a pass over the data with human annotators who judged whether the sentences were semantically appropriate for the target relation (DS assumption). We discarded all structures whose sentences did not express the target relation. The curated subset of structures was then created from the remaining dependency structures. For each structure and relation, the final dataset comprised all source sentences and not just the ones sampled for the judgments (i.e., patterns retain their frequency information). The right-most three columns of Table 6.2 (p. 138) depict statistics for the sar-graphs in this small, curated dataset.[57]

We publicly released[58] the curated sar-graph data to support research in the areas of RE, question answering, paraphrase generation, and others. Accompanying the data, we provided a Java-based API which simplified the loading, processing, and storing of sar-graphs and also allowed to visualize the individual dependency structures we exploited for the generation of sar-graphs.

6.4.2 Analyzing the Curated Sar-graphs

In Section 6.2, we stated that sar-graphs should not be limited to phrases that explicitly refer to a relation. Rather, they should also include more implicit phrases. To achieve a better understanding of the kind of expressions that are automatically retrieved from the web, we further analyzed the phrases in the curated subset.[59] We conducted the analysis for all relation types using the PatternJudge tool (Hennig, H. Li, and Krause, et al., 2015). The relation *sibling relationship* served as an initial test bed for establishing a set of shared evaluation principles among annotators; there were three annotators in total. The annotators defined four categories, named after the expressivity of the patterns: *explicit*, *specific*, *implicit*, and *other*. Table 6.3 (p. 140) reports the distribution of these categories.

[57] Note that the large gap in the size of both kinds of sar-graphs does not allow to draw conclusions as to the quality of structures in the full set. Sar-graphs are intended to store vast amounts of language expressions in actual use. All of the source sentences of sar-graphs do mention a relation instance. Hence, they are potentially expressing the relations of interest. Premature use of imperfect filters to reduce this mass of data would inevitably mean to lose much of the language variety that we collected from examples in a tedious process. We decide to keep all of the expressions available and expect improved filters to handle them in the future. For now, however, we associated the structures with confidence scores so that present-day applications can directly access the expressions that were determined as correct by current filters.

[58] The data is available for download at `http://sargraph.dfki.de` (last access: 2017-05-09).

[59] Note that this means the investigated patterns have been generated from sentences with verified DS assumption. Furthermore note that this analysis of phrases does not aim at detecting causes for erroneous patterns, like the analysis in Section 3.7, but rather aims at better understanding the linguistic variety in the patterns.

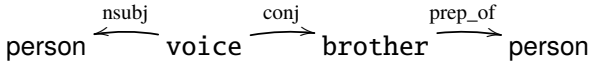
Relation	Explicit	Specific		Implicit		Other
		(source)	(preproc.)	(slightly)	(very)	
award honor	52.3	12.9	21.9	0.9	1.7	10.2
award nomination	17.0	19.5	47.7	3.1	10.6	2.1
country of nationality	14.6	36.5	13.8	17.8	15.1	2.0
education	45.9	22.6	16.2	3.0	10.6	1.8
marriage	38.7	26.3	7.9	6.2	5.6	15.3
person alternate name	7.3	18.9	39.3	6.2	25.6	2.8
person birth	40.4	17.2	15.5	6.0	19.7	1.3
person death	64.0	8.7	12.9	0.9	12.0	1.5
person parent	46.9	17.6	12.4	0.5	17.0	5.6
person religion	42.9	21.1	6.5	10.3	18.7	0.6
place lived	20.7	5.5	16.9	33.2	23.3	0.5
sibling relationship	38.7	22.7	4.7	1.9	1.6	30.5
acquisition	21.5	14.7	3.7	0.4	55.0	4.7
business operation	34.0	7.6	21.3	3.2	30.6	3.4
company end	9.7	7.1	18.0	21.0	39.5	4.7
company product relationship	29.2	22.3	9.3	7.5	25.1	6.6
employment tenure	64.5	5.9	11.8	1.5	10.0	6.5
foundation	48.8	13.6	20.0	3.6	11.9	2.1
headquarters	33.2	20.7	14.3	4.8	22.2	4.8
organization alternate name	20.3	5.8	50.8	8.1	12.1	2.9
organization leadership	63.8	1.6	17.9	0.5	10.2	5.9
organization membership	53.9	8.8	17.2	2.3	13.4	4.3
organization relationship	30.9	6.2	27.1	9.9	25.0	0.9
organization type	12.1	15.9	40.5	1.7	29.0	0.6
sponsorship	36.0	12.5	32.5	2.6	13.0	3.4
All relations	35.0	15.1	20.5	6.7	17.6	5.1

Table 6.3 – Distribution of evaluation categories for the curated subset of phrases.

Explicit Patterns We labeled a structure as *explicit* if it is grammatically correct and if it explicitly refers to the relation. Patterns in this category represent the prototypical way of referring to relations and require few or no contextual information to convey the fact that a particular relation holds between a number of entities. It is not a surprise that this category accounts for the biggest part of the patterns in the curated subset, as the patterns in this dataset were sampled from the ones with the best FO filter and semantic filter scores.

Specific Patterns The second category, *specific*, applies to dependency structures mostly found in the long tail of the frequency distribution. *Specific* structures, while both grammatically and semantically correct, are structures that are complex and sometimes include irrelevant parts of the sentence specific to a particular relation instance or mention thereof. In total, correct yet (overly) specific patterns account for approximately one-third of the curated dataset. We introduced two sub-classes of *specific* patterns, the first one, *specific (source)*, contains patterns whose peculiarities can be attributed to somewhat complex source sentences. Example 6.7 shows an example from *sibling relationship*, which includes the head word `voice`; see the source sentence in Example 6.8. Such dependency structures do not generalize well and are hence unlikely to be "productive" for application tasks (e.g., they are unlikely to yield novel relation instances when applied in RE).

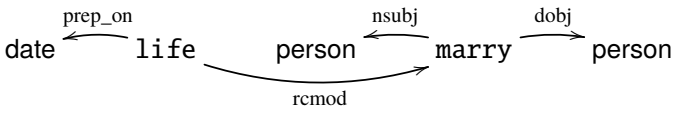
Example 6.7



Example 6.8 [60] `Jansen Panettiere is an American voice and film actor, and the younger brother of actress Hayden Panettiere.`

The patterns in the second sub-class, *specific (preprocessing)*, result from defective linguistic preprocessing. The following examples illustrate how errors in sentence-boundary detection (Examples 6.9 and 6.10) and NER (Example 6.11) can result in overly detailed patterns. In Examples 6.9 and 6.10, the tokens "`Personal life`" (which are most likely a headline or news-category identifier) are erroneously interpreted as part of the sentence with the relation mention and hence are included in the extracted pattern as well:

Example 6.9



[60] In this subsection, we highlight relation arguments in text examples by underlining with a straight line, and indicate lexical pattern-elements with a wavy line.

Example 6.10 Personal life On July 5, 2003, Banks married sportswriter and producer Max Handelman, who had been her boyfriend since she met him on her first day at college, September 6, 1992.

Example 6.11 displays a sentence for the relation *award honor*, where the title of a book was not recognized as an entity by the NER module. Consequently, the words Left Us would be included as plain lexical nodes in any pattern that is produced from this sentence, effectively limiting the coverage of the pattern to the faulty interpretation of this particular text.

Example 6.11 Rahna Reiko Rizzuto is the author of the novel, Why She Left Us, which won an American Book Award in 2000.

Implicit Patterns The next three examples are from category *implicit* and are instances of patterns and sentences where, without contextual information from the texts or common-sense knowledge, the presence of a relation mention can only be inferred. Again, we distinguished two sub-classes, here corresponding to the degree of implicitness, with the more implicit class, *implicit (very)*, being more frequent among the phrases in the curated set. Example 6.12 is an instance of sub-category *implicit (slightly)* from the relation *acquisition*. Judging from the source sentence in this example, we cannot be entirely sure whether or not an acquisition took place, because "felt compelled to" might only express a momentary mindset of the company's leaders that was not followed by action. If it was, it is not clear if "Wyeth" or "a company like Wyeth" (i.e., a similar company) was acquired.

Example 6.12 The looming expiration of Lipitor's patent in 2012 is a big reason Pfizer felt compelled to buy a company like Wyeth.

Examples 6.13 (relation *acquisition*) and 6.14 (relation *award honor*), both p. 143, are taken from the sub-class with even less explicit mentions: *implicit (very)*. Here, the dependency structures may not even include any relation-relevant content words occurring in the sentence. In Example 6.13, the most explicit element expressing an acquisition is the lemma "purchase". However, the pattern extracted from this relation mention does not include this word and instead focuses on someone's job transition, which in this case is caused by a company acquisition.

Example 6.13 Julian joined Old Mutual in August 2000 as Group Finance Director, moving on to become CEO of Skandia following its purchase by Old Mutual in February 2006.

Similarly, in Example 6.14, the target relation *award honor* is indirectly entailed by highlighting that a particular person is hosting an award reception, which correlates with and in some cases might even entail that this person won the award before.

Example 6.14 The 69th Annual Peabody Awards ceremony will be held on May 17 at the Waldorf-Astoria in New York City and will be hosted by Diane Sawyer, the award-winning anchor of ABC's World News.

The fourth coarse category, *other*, pertains to boundary cases and instances of patterns and sentences where annotators were not sure which class to assign.

Observations Table 6.3 (p. 140) shows that, on average, almost one-quarter of patterns belongs to category *implicit*. This is an interesting insight into the degree of explicitness with which information is expressed in language. More specifically, it quantifies the generally accepted belief that deep language understanding requires the processing of context. Furthermore, it is an interesting finding regarding the kind of expressions that are collected with distant-supervision approaches. All current DS methods inherently include this implicitness, and it is yet an unsolved question how explicit mentions can be automatically separated from implicit ones. Future research might address this issue by focusing on the apparent differences between semantic relations.

Indeed, we observe a large variance in the distribution of evaluation categories between the relations listed in Table 6.3. Relations like *marriage* and *acquisition* could be handled more accurately by taking into account their respective properties: *marriage* patterns are twice as often explicit as *acquisition* patterns and are only rarely *implicit*, while the latter are five times as often *implicit*. For *marriage*, *implicit* patterns are mainly due to sentences stating a divorce, which, strictly speaking, is not an explicit mention of the *marriage* relation. However, the former presence of a *marriage* relation can be inferred without any probabilistic entailment and independent of the context.[61] The *acquisition* relation, however, largely depends on contextual information, which determines if a sentence entails an acquisition or not. This is because *acquisition* sentences often state a particular connection between two or more organizations, but make no explicit reference

[61] This is why divorce mentions are generally treated as being part of the *marriage* relation in this thesis.

to the initiation of this connection. For example, it might be mentioned that company x is a subsidiary of company y, but no statement is made about the act of purchase.

Other relations are especially susceptible to preprocessing errors, i.e., many patterns are categorized as *specific (preprocessing)*, e.g., *award honor* and *award nomination*. Sentences in this category often mention the titles of works the prize was awarded for. If those titles are not recognized as entities by the NER tagger, dependency parsing fails and parts of the title can erroneously end up in the extracted dependency structure. For these relations, particular attention is required on the development of high-quality preprocessing components.

6.4.3 Benefits of Sar-graphs

We believe that sar-graphs constitute a resource that can be used in numerous tasks. In the following, we elaborate on two aspects of their utility and significance: their role as a repository of language expressions and their capacity to extract facts and to generate texts.

Repository of Language Expressions Sar-graphs provide access to the diversity of language expressions that humans use to refer to relation instances. Here, we present an explorative comparison of the linguistic expressions in the sar-graphs with a typical representative of a relation-phrase store from literature. To some degree, this is similar to comparing apples and oranges, since sar-graphs go beyond the mere storing of individual patterns. Nevertheless, we conduct this comparison to give some insight into how the language information in sar-graphs relates quality-wise to other works from literature.

As the point of reference, we selected the PATTY system (Nakashole et al., 2012b),[62] which implements a generic approach for the creation of a taxonomy of textual patterns. PATTY starts with the execution of an open-IE approach to the collection of phrases, which is followed by an alignment phase that creates a subsumption hierarchy of patterns. The PATTY authors provide a disambiguation of their patterns to a KB, from which we select four relations with a considerable semantic overlap with the sar-graph relations and for which a reasonable number of patterns is available for both systems.

To get an estimate of the quality of PATTY's patterns, we took a sample of 200 patterns from each relation and, based on the entities with which a pattern co-occurred in Wikipedia, generated instantiations for all associated entity-type signatures. Three annotators were then asked to judge whether the majority of instantiations per pattern

[62] `http://resources.mpi-inf.mpg.de/yago-naga/patty/data/patty-dataset.tar.gz`, last access: 2017-05-08.

	PATTY			Sar-graphs					
				S filter w/ $n = 3$			Curated subset		
	3 ∗ (pattern count/lexical diversity/precision)								
employment tenure	1,246	5 %	44 %	15,656	26 %	39 %	226	32 %	70 %
marriage	3,426	8 %	12 %	48,166	16 %	38 %	451	26 %	66 %
organization relationship	838	6 %	46 %	4,344	45 %	20 %	303	65 %	37 %
person parent	2,327	5 %	29 %	32,771	18 %	31 %	387	24 %	65 %

Table 6.4 – Comparison of relation phrases in PATTY and sar-graphs. The selection of relations was limited by the small relation-level overlap between the resources. The S filter is described in Algorithm 4.2 (p. 82); the sar-graphs in the respective column were created from patterns passing this filter for the stated threshold. The curated subset of sar-graphs is described in Subsection 6.4.2.

expressed the respective target relation. For example, a person `joining` another person does not indicate the mention of an *employment tenure* relation, but a sports person `joining` a sports team does. For the sar-graph patterns, we followed a similar strategy, which resulted in 200 instantiated and string-serialized patterns per relation with entities shown to three human raters. If the annotators could not make sense of a pattern (e.g., because it was overly specific or contained superfluous elements not required for matching a mention of the respective target relation), their guideline was to rate this pattern as incorrect.

Table 6.4 compares PATTY with sar-graph patterns retained after applying the S filter from Algorithm 4.2 (p. 82)[63] and additionally presents statistics for the sar-graphs created from patterns in the curated subset (Subsection 6.4.1). Along with the number of relation phrases (*pattern count*) and the quality assessment from the raters (*precision*), we state a lexical-diversity score which specifies the average amount of distinct, non-function words a relation phrase contains with respect to other expressions in the same set. This score allows for a better interpretation of the absolute pattern numbers, which are subject to the different pattern-representation formalisms.[64]

For the relations in this analysis, sar-graphs provide more linguistic expressions at a higher rate of lexical diversity, i.e., the sar-graph patterns are at least as well suited for applications as the PATTY system with respect to coverage of lexical variations of relation mentions. Furthermore, more than 20% of the sar-graph patterns in this analysis link three or more arguments to one another, in contrast to the PATTY patterns, which are all binary.[65] Note that while PATTY and sar-graphs were created from different

[63] We set $n = 3$ as this typically results in a good precision/recall trade-off.
[64] The metrics precision and lexical diversity resemble the qualitative evaluation from Subsection 5.5.2.
[65] This aspect is not reflected in Table 6.4.

corpora, the size of these corpora is similar, i.e., 2.2 million web documents in the case of sar-graphs, and 1.8 million New York Times articles as well as 3.8 million Wikipedia articles for PATTY.

Both systems produce patterns at a similar level of precision, where for some relations one system trumps the other. Looking for an explanation of the low *marriage* precision of PATTY, we found that on average PATTY patterns contained fewer tokens than the ones in the sar-graphs. Consequently, PATTY patterns are more affected by word ambiguity, in particular when relations share their entity-type signature with many other relations, as *marriage* and *person parent* do.

Extraction and Generation In addition to their roles as an anchor in the linked-data world and as a repository of relation phrases, sar-graphs are also a useful resource both for the detection of fact mentions in text and for the generation of phrases and sentences from database facts. In the context of text generation, sar-graphs can function as business-intelligence tools aiming at producing natural-language reports for recurring review periods. Due to the high range of paraphrases available in sar-graphs, generation could produce stylistic variation as extensively as used in reports written by human authors. An application that combines both aspects, generation and detection, is text summarization, where sar-graphs permit to identify fact-heavy parts of a text (i.e., constructions that express all or most arguments of a relation in one sentence) and also allow these parts of a text to be rephrased in a shorter manner.

Finally, another application for which we applied sar-graphs is related to the area of computer-assisted language learning. We implemented a prototype (Ai and Krause et al., 2015) for the semi-automatic generation of reading-comprehension exercises for second-language learners. A language teacher, who in this scenario has to prepare such exercises for an upcoming class, is presented with news texts retrieved from the web, along with candidate multiple-choice questions and answers, relating to certain facts mentioned in the text. The teacher then picks the most useful question-answer pairs.

In this scenario, sar-graphs are utilized both for the fact-finding phase (i.e., for the detection of true-answer candidates), and for the generation of paraphrases for true facts as well as for the question asking about the facts. During the evaluation of this prototype, we found that for average-length news texts, several correct and potentially useful question-answer pairs were generated for each input text. This shows that sar-graphs are indeed useful in actual application settings.

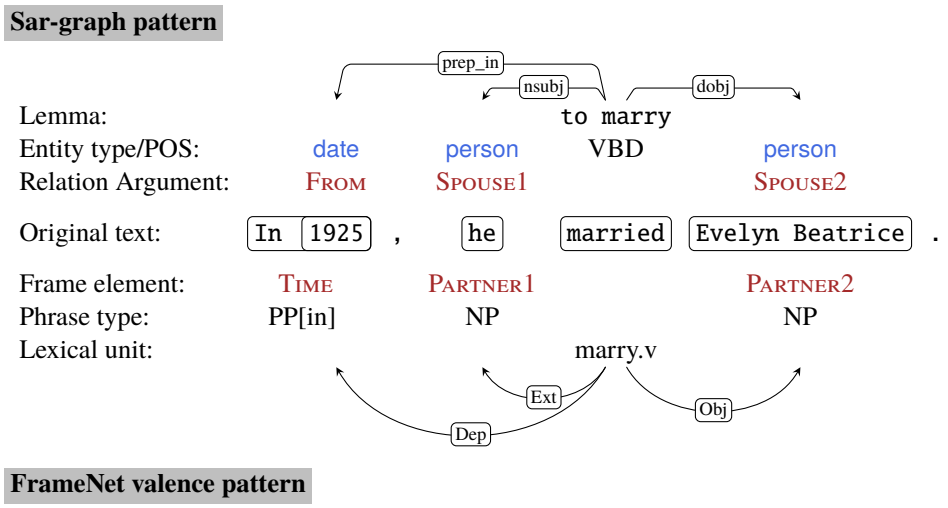


Figure 6.7 – Comparison of pattern representation in sar-graphs (top; for relation *marriage*) and FrameNet (bottom; for frame *forming_relationships*). Both formalisms connect semantic arguments (From, Spouse1, Spouse2, Time, Partner1, Partner2) and lexical items (`to marry`, marry.v) via grammatical relations (prep_in, subj, dobj, Dep, Ext, Obj). Abbreviations: Ext - External Argument, Obj - Object, Dep - Dependent, NP - Noun phrase, PP[in] - Prepositional phrase.

6.5 Linking to FrameNet

The sar-graphs presented in the previous subsections are linked (a) on the relation level to the KB Freebase and (b) on the lexical-semantic level to BabelNet. In this section, we report an effort to link sar-graphs to the resource FrameNet on the levels of both relations and phrases.

The Resource FrameNet is a lexical resource for English that documents semantic and syntactic combinatorial possibilities of words and their senses. The resource consists of schematic representations of situations (called frames). For instance, the frame *win prize* describes an awarding situation with semantic roles (frame elements) like Competitor, Prize, Competition, and so on. A pair of word and frame forms a lexical unit, which is similar to a word sense in a thesaurus. In turn, lexical units are connected to lexical entries, which capture the valence patterns of frames. These patterns provide information about frame elements and their phrase types and grammatical functions in relation to the lexical units. Each pattern is illustrated by a set of annotated sentences. An example of a valence pattern is shown in the bottom of Figure 6.7. For illustration purposes, the figure also displays a dependency structure for the same sentence as it would be represented in a sar-graph.

Comparison of FrameNet to Sar-graphs Sar-graphs resemble frames in many aspects, e.g., both define argument roles for target relations and provide detailed information about linguistic constructions. What differentiates the two resources is that FrameNet contains some very generic frames (e.g., *forming_relationships*) that have no explicit equivalent in a sar-graph relation. Moreover, sar-graphs specify fewer semantic roles than frames typically do and they mainly cover the aspects of a relation that are important from a KBP perspective. For example, the frame *forming_relationships* covers an EXPLANATION (divorce reason, etc.) and an ITERATION counter (for the relationships of a person), in addition to the arguments listed in the corresponding *marriage* sar-graph. Furthermore, FrameNet specifies relations between frames (*inheritance*, *subframe*, *perspective on*, etc.); sar-graphs are not linked in such a way.

Another difference is the relationship between lexical items and their corresponding frames and sar-graph relations. Lexical units in FrameNet imply frames by subsumption, e.g., `to befriend` and `to divorce` are subsumed by *forming_relationships*. In comparison, sar-graphs cluster both expressions that directly refer to instances of the target relation (e.g., `to wed` for *marriage*) and those that only entail them (e.g., `to divorce` for *marriage*).

6.5.1 Phrase-Level Linking

FrameNet 1.5[66] contains 74k valency patterns and more than 170k annotated sentences. We linked them with two variants of the sar-graphs, an automatically filtered version[67] and the curated subset. Instead of directly aligning the valency patterns with the corresponding dependency structures, we applied the Web-DARE pattern-discovery pipeline (Chapter 3) to the FrameNet sentences associated with the valency patterns. Then, the resulting patterns were matched against the patterns in sar-graphs and served as a proxy for the linking. Thus, we could avoid the costly mapping of the two syntax representations.

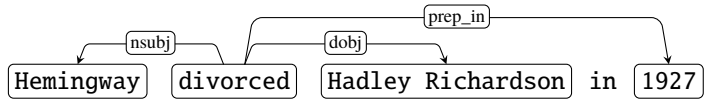
Approach First, we filtered the set of annotated sentences in FrameNet for those that mentioned two or more frame elements. These sentences were then processed by a dependency parser, after which sar-graph-like phrase patterns were extracted. This step resulted in more than 80k *FrameNet patterns*. We determined corresponding patterns from sar-graphs and FrameNet by comparing them via tree-edit distance.[68] We only

[66] Downloaded from `http://framenet.icsi.berkeley.edu` in 2015.

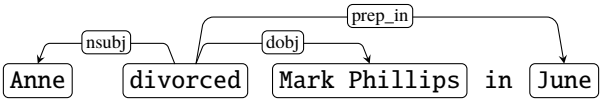
[67] Similar to the methodology depicted in Subsection 6.4.3, we filtered the patterns with Algorithm 4.2 (p. 82) and set $n = 3$.

[68] We used the algorithm by K. Zhang and Shasha (1989) as provided at `http://web.science.mq.edu.au/~swan/howtos/treedistance/package.html`.

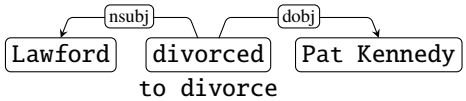
p_{SG}:



p^1_{FN}, ***exact*** **link:**



p^2_{FN}, ***subsumption*** **link:**



p^3_{FN}, ***other*** **link:**

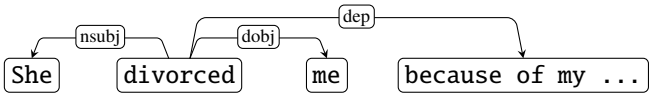


Table 6.5 – Examples of pattern-level links between FrameNet and sar-graphs. p_{SG} is a sar-graph pattern for relation *marriage*; p^1_{FN}, p^2_{FN}, p^3_{FN} are FrameNet patterns from *forming relationships* that were linked to p_{SG}.

Sar-graph variant	Exact	Subsumption	Other
Curated	251	554	4,419
Autom. filtered	2,978	8,329	113,201

Table 6.6 – Distribution of pattern links.

considered the lexical level and the syntax of the patterns, while ignoring differences in the definition of semantic arguments and their names, as these would have been hard to resolve automatically and would have required an ontology integration step (Subsection 6.5.2). The calculated distance between two patterns was normalized by the number of edges in the two patterns, resulting in a normalized distance d. This allowed us to order the links between FrameNet patterns and sar-graph patterns by d and to discard all links with $d >$ a threshold t. Furthermore, we excluded all links where none of the patterns mentioned the source lexical unit of the original valence pattern. We defined three classes of pattern-level links with the following correctness criteria; examples are given in Table 6.5 (p. 149):

- *Exact match*: $d = 0.0$. The link is correct if the patterns are semantically equivalent.
- *Subsumption*: $d > 0.0$ and one of the two patterns is syntactically fully contained in the other. Correctness of the link requires that there is an entailment relation between the patterns.
- *Other*: $d > 0.0$ and neither pattern is included in the other. The link is correct if the meanings of the patterns are related.

Results and Linking Errors We present the distribution of links across the three classes in Table 6.6 (p. 149), for a threshold of 0.5 on the normalized distance d. A large fraction of sar-graph elements were automatically aligned with their FrameNet counterparts. Since the linking step is based solely on the lexical and syntactic features of the patterns, there are two main causes for semantically erroneous links:

- *Semantic ambiguity*: Linked patterns are not synonymous due to polysemy and homonymy. For example, a sar-graph pattern for relation *organization leadership* which contained the lemma `to lead` was erroneously linked to a pattern from frame *cotheme*, which uses this verb to mean *showing someone the way* and not, as in *organization leadership*, *to be in charge of something*.
- *Argument-type mismatch*: Patterns have a different meaning because of the semantic types of their respective arguments. For example, the ORGANIZATION of sar-graph patterns for *organization leadership* was matched to the element DEPICTIVE of frame *leadership*, where the correct mapping would have been to frame element GOVERNED. As an example, consider the difference between `Informatica chairman Sohaib Abbasi` and `deputy chairman Eric Goodman`.

Relation	**Arguments**				
employment tenure	EMPLOYER	EMPLOYEE	POSITION	FROM	TO
↓	↓	↓	↓	↓	↓
Frame	**Frame Elements**				
Being_employed	EMPLOYER	EMPLOYEE	POSITION	TIME	TIME
Employee_scenario	EMPLOYER	EMPLOYEE	POSITION	—	—
Employing	EMPLOYER	EMPLOYEE	POSITION	TIME	TIME
Employment_start	EMPLOYER	EMPLOYEE	POSITION	TIME	—
Employment_continue	EMPLOYER	EMPLOYEE	POSITION	TIME	TIME
Employment_end	EMPLOYER	EMPLOYEE	POSITION	—	TIME
Firing	EMPLOYER	EMPLOYEE	POSITION	—	TIME
Get_a_job	EMPLOYER	EMPLOYEE	POSITION	TIME	—
Hiring	EMPLOYER	EMPLOYEE	POSITION	TIME	—
Leadership	GOVERNED	LEADER	ROLE	TIME	TIME
Appointing	SELECTOR	OFFICIAL	ROLE	TIME	—

Table 6.7 – Linked frames for sar-graph relation *employment tenure*.

6.5.2 Linking Frames and Relations

We now discuss the manual integration of the schemas that the two resources are based on. For each of the sar-graphs, we determined which frames share a similar meaning by reviewing their respective definitions and aligning the arguments of sar-graphs with frame elements. The mapping of frames to relations is a many-to-many mapping, e.g., the relation *employment tenure* is mapped to 22 frames, among them the frame *leadership*. This frame is in turn linked to the sar-graph relations *organization leadership* and *organization membership*. Table 6.7 shows an excerpt of the frame-relation alignment.

We mapped the 25 sar-graph relations to 260 frames, with the number of frames per sar-graph ranging from 1 to 40. Some of the more extreme cases are *sibling relationship*, which is linked only to the frame *kinship*, and relation *acquisition*, which is mapped to frames like *commerce buy*, *commerce sell*, *shopping*, *receiving*, *getting*, and *possession*. Furthermore, the semantic agreement and mutual coverage of an identified pair of frame and relation varies greatly. *Acquisition* has a largely congruent extent with frames *commerce buy*, *commerce sell*, and *shopping*. In contrast, the frame *getting* is more general than *acquisition*, as it requires payment for acquired entities and since it also covers transactions with physical goods. For these uses, *getting* contains very polysemous lexical units like `to get`, which in particular contexts imply commercial interactions, e.g., `Peter got the novel from Amazon`.

Sar-graph variant	FrameNet phrases	# Correct extractions	Recall change	Precision change
Curated	no	42,639	—	—
Curated	yes	67,680	+58.73%	-37.64%
Autom. filtered	no	174,063	—	—
Autom. filtered	yes	184,343	+5.91%	-9.18%

Table 6.8 – Results from extraction experiment on ClueWeb.

Extraction Experiment We evaluated the impact of expanding sar-graphs with FrameNet phrases with an RE experiment. In particular, we were interested in whether the addition would substantially increase the coverage of linguistic expressions. We selected a set of approximately 30 million sentences from the ClueWeb datasets (Callan et al., 2009) with linked mentions of Freebase entities (Gabrilovich et al., 2013).

All patterns of the two sar-graph variants used in Subsection 6.5.1 were matched against the sentences of the corpus to extract facts, as were the FrameNet phrases which were part of a frame linked to a sar-graph relation. We evaluated the detected relation mentions by checking whether they were listed in Freebase. Table 6.8 displays the amount of correct facts that the sar-graphs covered, as well as the amount of them extracted after the addition of FrameNet phrases. For both sar-graph variants, the extraction performance substantially improved after the expansion step.

6.6 Related Work

In the previous sections, we motivated the construction of sar-graphs and outlined a method of building them from an alignment of web text with known facts. Taking into account the implemented construction methodology, it may seem that sar-graphs can be regarded as a mere side-product of pattern discovery for RE. However, sar-graphs are actually an evolution of this; they are a novel linguistic knowledge resource post-processing the results of pattern discovery. Our work is innovative in comparison to traditional pattern-discovery approaches, since we reorganize the collected structures into a coherent, relation-specific linguistic resource, instead of viewing them as sets of independent patterns. The formalism used for the intermediate storing of phrases is based on the work from Chapter 3; we leave a more sophisticated methodology that could return expressions at various granularities (Beedkar and Gemulla, 2015) for the future. Similary, in the future, sar-graphs could be further developed as a platform for merging and fusing available extraction patterns from other sources like NELL or PATTY.

In addition to large-scale linguistic repositories such as BabelNet, there are manually constructed resources with linguistic information on a smaller scale such as VerbNet[69] (Schuler, 2005). These resources already existed before the development of large KBs, and they were constructed with the motivation of modeling the semantics of language at the word or syntactic level, without any explicit link to the real world. In contrast to these resources, sar-graphs are data-driven, constructed automatically, and incorporate statistical information about relations and their arguments. Therefore, sar-graphs complement these manually constructed linguistic resources. Furthermore, since word-sense information is integrated into sar-graphs, a linking to other linguistic resources via word senses is straightforward and, thus, sar-garphs contribute to linguistic linked-open-data.

Finally, projects related to sar-graphs presented ontology formalizations (Lemon by McCrae et al., 2011) and manual and semi-automatic methods to enrich KBs in terms of relation lexicalizations (Walter et al., 2014a; b; Unger et al., 2013; Lao et al., 2012). The main goal of these methods is to produce a large set of phrases for KBP via RE (Cimiano et al., 2014; Gardner et al., 2013; 2014) and to generate natural-language descriptions of KB content (Cimiano et al., 2013). Sar-graphs provide this functionality too and additionally link the discovered patterns to word meanings by employing WSD.

6.7 Summary

In this chapter, we presented the linguistic resource sar-graphs, which aggregates knowledge about the various phrases a language provides for expressing semantic relations. We based our approach on Chapter 3 for automatically accumulating such linguistic knowledge and described a general method for merging it into a single connected graph. Furthermore, we illustrated the described construction methodology by creating and evaluating sar-graphs for 25 relations. We also discussed how sar-graphs are linked to other resources on various granularity levels and reported on an analysis of the explicitness with which phrases refer to a target relation.

An important aspect of sar-graphs is that they are created in a bottom-up fashion. This way, sar-graphs are empirically grounded on the actual ways people communicate about semantic relations. While at least currently, a fully automatic approach cannot produce a noise-free resource due to shortcomings of unsupervised quality assessments, the hybrid approach followed for sar-graphs is still preferable to a fully curated approach. This is because such a curation effort requires to decide a-priori on the application for

[69] VerbNet is a lexicon that maps verbs to predefined classes which define the syntactic and semantic preferences of the verb.

which a resource is created and to tediously elaborate the corner cases of target relations. For example, before starting to work on an intellectually-created ontology for the topic *marriage*, one would need to opt for either the inclusion or the exclusion of expressions like `exchanging the vows` and `tying the knot`. The great advantage of an empirical bottom-up approach is that one is not pressured to make such a-priori ontology-level decisions.

Another important choice we make is the association of sar-graphs to specific languages, which deviates from the approach taken in the creation of some multilingual lexical-semantic resources (e.g., BabelNet). We made this decision because phrases expressing semantic relations can be highly specific to cultural backgrounds. For example, a Greek report on a wedding may refer to `wedding crowns` for bride and groom, while in an English sar-graph for the *marriage* relation, such crowns would not show up. In a Greek wedding the `betrothal` can be a part of the entire ceremony, in other cultures it must have taken place a certain period before the wedding. In some cultures, `exchanging the rings` means `getting married`, while in others there is no such concept.

A potential way of extending the presented sar-graphs is to additionally incorporate relation mentions with fewer than two arguments. Such snippets can contain valuable information that might not be covered by the $n \geq 2$-ary mentions in a document, in particular when they are part of cross-sentence relation mentions. We address this particular type of relational information in the next chapter.

Chapter 7

Document-Level Extraction of Facts[70]

Contents

[70] This chapter presents results from joint work with Hong Li, Hans Uszkoreit, Dirk Weissenborn, and Feiyu Xu.

7.1 Introduction

In Chapters 3–6, we presented IE approaches that primarily rely on linguistic patterns. Inherent to their nature of being a template for the grammatical structure of sentences, patterns do not cover relational information that is expressed on a granularity beyond sentences, namely paragraphs and documents. Many studies investigated the problem of cross-sentence RE (CS-RE)[71] and while the exact amount of beyond-sentence-level relation mentions greatly varies between datasets under investigation (Section 2.4),[72] the significance of this task is generally accepted. In this chapter, we discuss our work that addresses CS-RE by extending the standard IE-pipeline (Section 2.2) with co-reference resolution for relation mentions, in literature often referred to as *event linking*.[73] The design of our event-linking method was guided by an analysis of annotated documents, which we conducted to better understand the difficulties and challenges of handling document-level relations. In addition to examining standard datasets for RE (Section 2.7), we created a new dataset, called COCKRACE, which allows observing further properties of mentions due to its native inter-sentential annotation. Similarly to the evaluation corpus whose creation was described in Section 4.3, COCKRACE contains English texts crawled from the website of the PEOPLE magazine.

The main conclusion we drew from the analysis is that implementing a flexible way for the representation of relation mentions at linking time is crucial, since many different linguistic phenomena and properties of the mentions have an impact on whether or not they are co-referential. NN-based methods that produce latent-feature vectors for natural-language concepts seem ideally suited for this job (Section 2.6). We drew a further motivation for the design of our event-linking model from state-of-the-art systems in literature. Current systems base their decisions on rich semantic features from various KBs, thus restricting themselves to domains where such external sources are available. Consequently, we designed a model which does not rely on such features but instead utilizes sentential features coming from CNNs. Two such networks first process co-reference candidates and their respective context, thereby generating latent-feature representations, which are tuned towards aspects of relations and events that are relevant for a linking decision. These representations are augmented with lexical-level and pairwise features and serve as input to a trainable similarity function that produces a

[71] We treat the following terms as synonymous to CS-RE: inter-sentential RE, discourse-level RE, document-level fact extraction.

[72] One instance of such studies was presented by Swampillai and Stevenson (2010), who reported that 9.4% of relation mentions in ACE texts from 2003 cross the sentence boundary, as do 28.5% of instances in MUC 6 data.

[73] See Section 2.4 for a discussion of linking approaches and another class of CS-RE methods.

co-reference score. Our model achieves state-of-the-art performance on two datasets, one of which is publicly available. We also give account of an error analysis, which points out directions for future research.

7.2 The COCKRACE Corpus

This section focuses on the description of the new dataset COCKRACE, which covers the same text genre and semantic information as the evaluation datasets in Sections 3.5 (corpus PEOPLE_{Test}) and 4.3 (corpus CELEBRITY). The resulting corpus was made publicly available;[74] the name of the corpus is short for *Corpus of co-references and kinship relations in ACE fashion*.

7.2.1 Annotation Elements and Preparation

A number of co-reference annotation guidelines were previously proposed in literature, both for general linguistic phenomena and for entity and event-detection tasks (e.g., Mitkov et al., 2000; Hasler et al., 2006). For the annotation of COCKRACE (Text Analytics Group, 2013), we mainly built on two sets of guidelines: the ACE annotation manuals (Doddington et al., 2004; Linguistic Data Consortium, 2004a; b) and the co-reference guidelines proposed by Komen (2009). The following list summarizes the different aspects of the conducted annotation:

- Mentions of different classes of entities were marked, more specifically, individual persons and person groups, dates, and locations.
- Mentions of the semantic relations *marriage*, *person parent*, *sibling relationship* (defined in Table 3.2, p. 59) were identified. In contrast to the ACE definitions, we did not limit arguments of mentions to single sentences. Furthermore, we did not distinguish between relations and events. We also allowed relations to have more than two arguments.
- Annotators also highlighted semantic terms that were relevant for the target relations and words that triggered mentions of relations. These could occur in various forms (name, nominal, pronoun, verb, adjective) and were typically lexicalizations of kinship-related word senses (e.g., `marriage`, `sister`).
- We linked entity mentions with co-reference relations of two types: an identity link and, inspired by Komen (2009), a variant for weaker forms of co-reference.

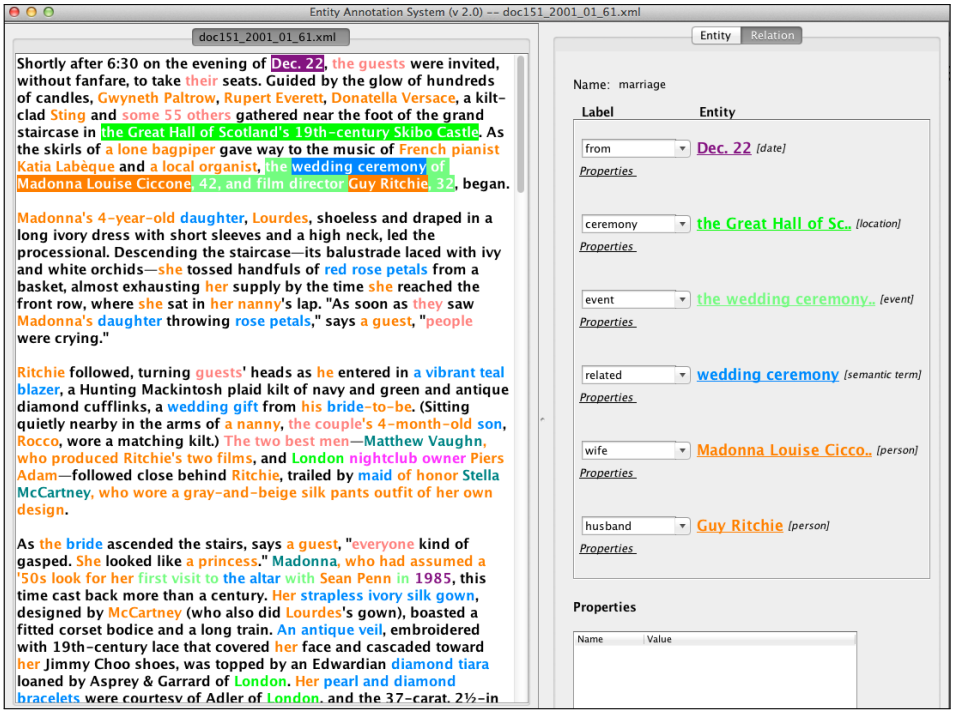
[74] `https://dfki-lt-re-group.bitbucket.io/downloads/`, last access: 2017-05-15.

Corpus Setup The CELEBRITY dataset from Chapter 4 consists of tabloid-press documents taken from a collection of several thousand PEOPLE-magazine articles from the years 2001–2008. For the new corpus, we selected a different subset from the same base collection of articles. To facilitate the manual annotation process, we preprocessed and also preannotated the corpus on several layers: We used the NER component of the WebDARE pipeline (Chapter 3) to highlight mentions of certain types (e.g., person, location). More specifically, we employed the entity recognizer from Stanford CoreNLP (Finkel et al., 2005) as well as a dictionary-based NER module working with Freebase topics. Additionally, a regular-expression based date recognizer was applied. The Stanford CoreNLP package was also used for the segmentation of sentences. We automatically marked potential mentions of the three target relations using a well-performing subset of the extraction patterns from Chapters 3 and 4. Finally, we utilized the relation-specific sub-graphs from Chapter 4 to automatically mark potential trigger terms for mentions of the three target relations. The result of this preprocessing was the automatic annotation of approximately 16,000 sentences, 436 relation mentions, as well as 4,800 mentions of 525 semantic key terms.

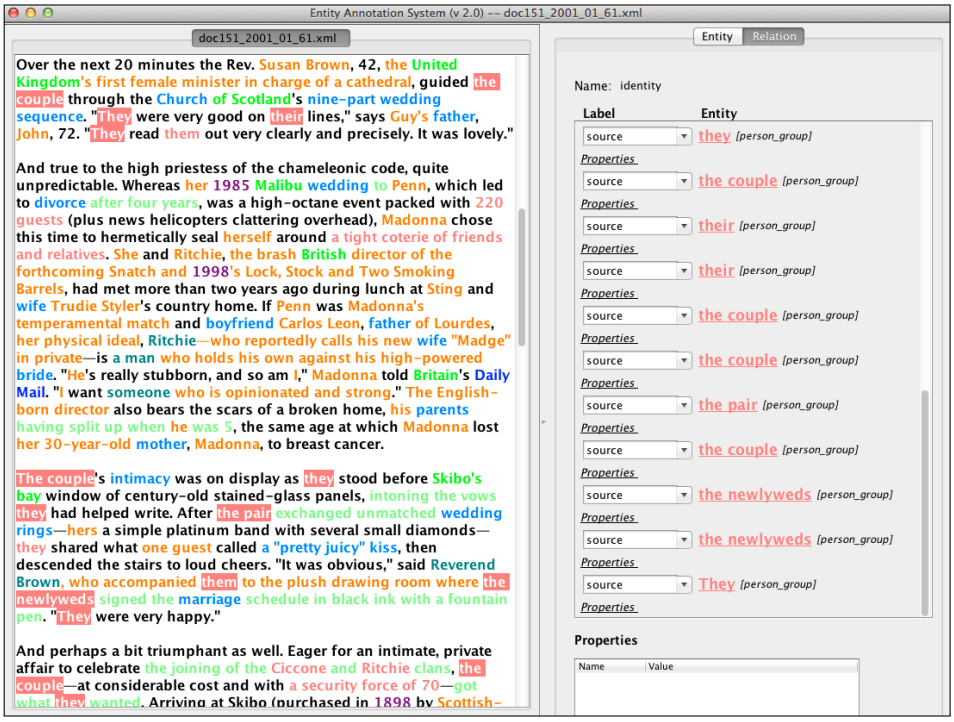
7.2.2 Annotation Process

The corpus was annotated by two human experts, who followed the guidelines outlined above (Text Analytics Group, 2013) and who used the tool Recon (H. Li et al., 2012) for all aspects of the annotation process.[75] Figure 7.1 (p. 159) presents screenshots of an example document from COCKRACE, taken after one of the experts made an initial pass over the document. The first screenshot shows a relation mention which crosses sentence boundaries. More specifically, it contains a paragraph with three sentences, all of which mention one or more arguments of the same *marriage* relation instance. The second figure depicts an excerpt from the same source document. Here, elements of a specific co-reference chain are highlighted (pink-colored boxes), all of which refer to the newlywed couple. This example illustrates why the resolution of co-referential expressions is an important aspect of CS-RE: As the expressions in the chain are distributed over a long span of text, they allow to piece together information bits of the *marriage* instance from all over the document. We make use of co-reference chains in the analysis of Section 7.3.

[75] Recon is a Java-based tool for the annotation of $n \geq 2$-ary relations among arbitrary text spans.



a) Screenshot showing the annotation of a relation mention that extends over three sentences. The arguments are listed on the right-hand side of the picture and highlighted with colored boxes in the text body on the left.



b) Another screenshot, which illustrates the entity-level annotation in Cockrace, namely, a single chain of co-referring noun phrases.

Figure 7.1 – Screenshots of the Cockrace documents in Recon.

Statistics The annotation was carried out over one year. A considerable amount of this time can be accounted to discussions between the annotators on different aspects of the guidelines. The final dataset is comprised of 140 documents with approximately 8,500 paragraphs, for which the experts marked 45,000 concept mentions, more than 1,800 kinship relation mentions, about 4,000 sets of co-referring expressions, and nearly 1,300 bridges between singular/plural entity references.

7.3 Analysis of Challenges in CS-RE

In this section, we report the results of two studies which approached the problem of CS-RE from different angles. In the first analysis (Subsection 7.3.1), we assumed we were given an IE system with oracle components for NER, CR, and intra-sentential RE, i.e., the individual parts are feature-wise on par with state-of-the-art methods and at the same time always return correct results. This pipeline was extended with a straightforward linking strategy for relation mentions, and we examined coverage and limitations of this approach by simulating the extraction process on COCKRACE. Then (Subsection 7.3.2), we focused on the individual components of the IE pipeline and their interplay with RE. We removed the oracle assumption and tried to ascertain how much (CS-)RE performance potentially suffers from limitations and deficiencies in NER, CR, and the linking of extracted relation mentions. In this second analysis, we worked with two standard IE datasets and a web crawl.

7.3.1 Analysis Part 1: COCKRACE

Figure 7.2 (p. 161) illustrates the simple linking strategy for relation mentions that we used on COCKRACE. The text in the example contains two sentences, each with one binary relation mention. One of the arguments of the second sentence is co-referential with an argument from the first sentence. We used this information by merging the intra-sentential relation mentions along the co-reference links, thereby creating cross-sentential relation mentions. As we were interested in the fundamental expressive power and coverage of the baseline linking strategy, we used the provided named-entity annotation along with resolved coreferences, as well as the provided intra-sentential relation annotation. We did not employ the information about singular/plural bridges of entity references, as such information is usually not provided by CR tools in literature.

Results COCKRACE contains 49 relation mentions that connect entities from different sentences. The baseline approach with the naive linking strategy of Figure 7.2 covered

1) NER and 2) CR for entities:

person person
Jane Miller is married to Texan physician Jack.

date
I attended her wedding event in 2011.

3) Intra-sentential RE:

⟨SPOUSE1: Jane Miller, SPOUSE2: Jack, CEREMONY: —, FROM: —, TO: — ⟩

⟨SPOUSE1: her, SPOUSE2: —, CEREMONY: —, FROM: 2011, TO: — ⟩

4) Merging of relation mentions:

⟨SPOUSE1: Jane Miller, SPOUSE2: Jack, CEREMONY: —, FROM: 2011, TO: — ⟩

Figure 7.2 – Steps (1)–(4) of the testbed approach to discourse-level RE. Co-referring entity mentions from step (2) are depicted via underlining. Step (4) merges all mentions from step (3) based on co-reference relations between arguments (her ↪ Jane Miller).

Observation	Affected cross-sentence relation mentions
Disagreement of anaphora	~ 10%
Cross-relation links	~ 20%
Implicit arguments	~ 50%

Table 7.1 – Phenomena not covered by the linking strategy on COCKRACE.

only three of these. This means that out of the annotated cross-sentential mentions, only three can be assembled from their individual binary relation-mention fragments by joining the fragments along (identity) co-reference chains. The remaining mentions bear properties which require more sophisticated processing of the text. Table 7.1 lists the three most important causes of uncovered inter-sentential relation mentions.

Number Disagreement of Anaphora Approximately one tenth of the inter-sentential relation mentions in the corpus contain both plural entity references (e.g., they) and their singular antecedents. Without their successful resolution, the simple CS-RE strategy fails. Consider the three sentences in Example 7.1 (p. 162), taken from COCKRACE, where underlining signifies entities participating in the same co-reference relation:

Example 7.1 But as Lemmon's career flourished, his marriage to Cynthia Stone [...] wilted. "We really had more of a brother-and-sister relationship than a good, solid marriage," he told PEOPLE in 1998. [...] They divorced in 1956, two years after the birth of son Chris.

As in Figure 7.2 (p. 161), there are two binary mentions which could be merged:

- *marriage*: ⟨ SPOUSE1: Lemmon, SPOUSE2: Cynthia Stone ⟩
- *marriage*: ⟨ SPOUSE1+2: They, TO: 1956 ⟩

This time, however, one of the arguments (They) needs to be simultaneously mapped to two arguments (Lemmon, Cynthia Stone), which requires a sophisticated processing of entity references.

Cross-relation Links An observation that affects 20% of the cross-sentence mentions is orthogonal yet still closely related to the use of mutually-exclusive relation definitions for pattern filtering (FO filter, in particular Equation 3.3, p. 66). This filter relied on the insight that particular linguistic constructions can express at most one out of several semantically non-overlapping relations. Cross-relation links involve a different kind of reasoning. Here, annotators decided that several intra-sentential mentions of semantically close relations were arranged and contextualized in a way that allowed to merge them into a cross-sentential mention of one of the considered relation types. As an example, see the text in Example 7.2:

Example 7.2 [76] A year later 10-year-old Jamie endured her own domestic drama in her hometown of Red Deer, Alta., when her father, Gene Sale, an electrician, and her mother, Patti, split. [...] Jamie's older brother Jason, now 30, moved in with Gene, while Jamie moved in with her mom.

The first sentence in this text features two intra-sentential mentions of type *person parent*; the second one mentions a *sibling relationship* (highlighted in the example by underlining):

- *person parent*: ⟨ PARENT: Gene Sale, CHILD: her ⟩
- *person parent*: ⟨ PARENT: Patti, CHILD: her ⟩
- *sibling relationship*: ⟨ SIBLING1: Jamie, SIBLING2: Jason ⟩

[76] Incidentally, this example demonstrates a special case, in which two separate relation mentions occur in the same sentence. We include this case in CS-RE, even though, technically, no inter-sentential reasoning is involved.

The wider context and the type of article (biographic information) in this example made it apparent to the raters that Jason was also a child of Patti and Gene Sales, i.e., that joining these three mentions would result in a relation mention with four arguments for the *parent-child* relation: ⟨PARENT1: Gene Sale, PARENT2: Patti, CHILD1: Jamie, CHILD2: Jason⟩. The baseline strategy followed in this analysis only allowed merging of mentions of the same relation type.

It should be noted that this class of cross-relation links consists of borderline RE cases because a lot of contextual information needs to be considered to make an extraction decision. It is not entirely clear that this kind of inference should actually happen in the RE step, i.e., whether this sort of implicature holds for any text with mentions of these specific kinship relations, or, more generally, how hard it would be to identify which relation types should be handled jointly.

Implicit Arguments The final observation is that approximately fifty percent of the discourse-level relation mentions have arguments which are only referred to implicitly. This is realized via lexical relation triggers, which normally require a particular number of arguments, but which in this case leave one such argument slot unfilled for stylistic reasons. Consider Example 7.3:

Example 7.3 As a teenager she had been arrested for shoplifting; there were busts for DUI and, in 2003, an arrest for battery on her stepfather (the charge was dropped). [...] Daughter Paige had been born 11 weeks premature.

Here, the trigger word Daughter is a relational noun which opens an argument slot for a PARENT of Paige. Note that no entity mention is syntactically linked to the trigger word which could fill the PARENT role. However, the first sentence (and the previous sentences, which are not shown in the example) frequently refer to a specific person (see doubly-underlined words in Example 7.3). This person is the main topic of the text, therefore the apparent interpretation of Daughter assigns this topic to the syntactically unoccupied PARENT role. Another example is depicted in Example 7.4 (p. 164), for the relation *marriage*. Here, the SPOUSE of Julie Wiatt is only implicitly included in the relation mention.

Example 7.4 By year's end he was back at work as an editor for the Bureau of National Affairs-and tossing around a Wiffle ball with his son Zak, 16. [...] "It's like seeing someone reborn," says wife Julie Wiatt, 57, a newspaper editor.

Conclusion from Analysis Part 1 The three observations listed above guided the design of the new CR-RE approach we discuss in Section 7.4.

- The observation on the number disagreement of anaphora emphasizes the importance of CR for classifying the argument roles of entity mentions. A method for CS-RE needs to factor in information about the identity of entity mentions, if available in the respective context. At the same time, a CS-RE method should not assume that all co-reference links between entities were correctly identified, as some links might go beyond the capabilities of state-of-the-art CR tools.
- The occurrence of cross-relation links in the rater annotation of Cockrace indicates that the handling of relation mentions should also account for connections between mentions of different relation types.
- The final observation on the large fraction of implicit arguments in Cockrace suggests that a CS-RE approach should not rely entirely on the presence of argument overlaps to link relation mentions. While such an overlap is indicative of two mentions referring to the same relation instance, co-reference can also occur without it.

7.3.2 Analysis Part 2: Further RE Datasets

In this part of the analysis, we focused on the identification of potential problems for extraction performance that are specific to the individual components of an IE system. For this, we analyzed a corpus of web documents and two datasets from shared tasks. Gold annotations of entities and relations were available for all documents.

Web Documents For nine facts of different semantic relations, we retrieved ten web documents each by querying a search engine with the fact itself. All documents were subsequently annotated by a domain expert for mentions of the respective fact. In total, this dataset comprises 400 relation mentions; we refer to this dataset by Web. The seed facts are listed in Example 7.5 (p. 165):

Example 7.5

- *acquisition*: ⟨ Ebay, Paypal, 2002 ⟩
- *acquisition*: ⟨ Walt Disney Company, Pixar Animation Studios, 2006 ⟩
- *company-end*: ⟨ Pan Am, bankruptcy, 1991 ⟩
- *marriage*: ⟨ Amber Heard, Johnny Depp, 2015 ⟩
- *marriage*: ⟨ Brad Pitt, Angelina Jolie, 2014, Château Miraval ⟩
- *marriage*: ⟨ Ted Hughes, Sylvia Plath, 1956, London ⟩
- *parent-child*: ⟨ Goldie Hawn, Bill Hudson, Kate Hudson, Oliver Hudson ⟩
- *person-death*: ⟨ Paul Walker, 2013-11-30, Santa Clarita, car accident ⟩
- *person-death*: ⟨ Robin Williams, 2014-08-11, Paradise Cay, suicide ⟩

Datasets from Shared Tasks The ACE program provided evaluation data for many facets of IE (Section 2.7); for our analysis, we used the ACE 2005 dataset (Walker et al., 2006) and refer to it as ACE. The TAC KBP 2014 Event track[77] followed ACE and required participants to recognize mentions of n-ary relations in text. We analyzed a version of the input documents annotated with respect to the Rich-ERE guidelines (Song et al., 2015), to which we later refer to as TAC. Both corpora contain two distinct annotation layers for relational information. We used the one with $n \geq 2$-ary relations ("events" in ACE terminology).

Results We started by investigating what fraction of relation information in the datasets is distributed over several sentences, i.e., which amount of relation mentions does not include all arguments of a particular relation instance within one sentence only (e.g., only one SPOUSE out of two in a *marriage* fact). This comprises only those relation mentions where the missing argument is indeed referred to within the respective document, i.e., there is another mention of the same relation instance in the same document which includes an additional argument. The second column of Table 7.2 (p. 166) lists the fraction of relation mentions in this category: 20%–70% of mentions in the three datasets have missing arguments and would, therefore, benefit from an inter-sentential linking. Note that these numbers are not directly comparable to the cross-sentence relation mentions from the COCKRACE corpus. While the latter are inherently inter-sentential, the mentions in WEB, ACE, and TAC are limited to single sentences and in each document there are many mentions of the same relation instance.

The third column of Table 7.2 states the amount of relation mentions with arguments that require co-reference resolution: On every dataset, around 60% or more entity mentions

[77] https://tac.nist.gov/2014/KBP/Event/; last access: 2017-05-17.

Dataset	Requires CS-RE	Requires CR	Requires neither
WEB	68.73%	58.31%	12.16%
ACE	22.99%	≥ 77.20%	≤ 16.26%
TAC	38.20%	≥ 74.86%	≤ 15.44%

Table 7.2 – Distribution of relation mentions across complexity classes. "Requires CS-RE" refers to relation mentions which are incomplete on the argument level. "Requires CR" corresponds to relation mentions with non-proper-name entity mentions as arguments. ≥/≤ in the second/third column are due to certain non-entity argument fillers with unclear realization status.

require CR. For WEB, 50% of entities with argument roles are (proper) name references, 25% are realized as pronouns, and 20% are nominal phrases. This distribution among entity-realization classes is similar in the shared-task datasets, where only approximately 30% to 45% of the argument mentions are realized via proper names. This means that a system which exclusively worked on the intra-sentential level for the handling of entity mentions would be limited in coverage to only half of the argument mentions, since resolving pronouns and nominals to their respective reference entity (most of the time) requires looking at the whole discourse.

The fourth column lists the amount of relation mentions which do not appear in any of the two previous categories. Only this percentage of relation mentions (≈ 15%) would also be extracted by an IE system which processed sentences individually and would neither resolve co-references between entity mentions nor link relation mentions corresponding to the same relation instance. In the following paragraphs, we state three particular observations we made during this analysis, which further illustrate the complexity of CR. Afterwards, we discuss our conclusions for this analysis part.

Number Disagreement of Anaphora Similar to the findings in Subsection 7.3.1, 13% of the argument mentions in WEB were affected by a disagreement of the number of anaphora. We did not carry out this aspect of the analysis for the datasets TAC and ACE, mostly because it is not possible to infer from an argument mention's corresponding part-of-speech whether it is used as a plural reference to several arguments or not.[78]

[78] "NN or NNS: Whether a noun is tagged singular or plural depends not on its semantic properties, but on whether it triggers singular or plural agreement on a verb" (Santorini, 1990, p. 17).

Trigger-Word Overlap The next observation is concerned with cases where an argument mention is realized through the same word that triggers the corresponding relation mention. Here, we follow the ACE notion of triggers (Linguistic Data Consortium, 2005b), which are typically single words of any word class or short phrases. Examples include the following, with triggers being underlined: `the late actor`, `Paul Walker's death` (for the relation *person death*); `the wedding` (for the relation *marriage*). Example 7.6, taken from WEB, illustrates the case of an overlap of trigger and argument:

Example 7.6 `For others, it added substance to the belief that Hughes had information relating to the relationship between him and his estranged wife in the years before her death, that he wished to conceal.`

Here, the trigger word `wife` is congruent with a reference to an argument of the relation, while another argument (`his`) is mentioned in a possessive relation to the trigger word. Capturing such examples requires a joint handling of relation-trigger detection and entity resolution, in contrast to a more standard pipeline approach. Nine percent of argument mentions in WEB were affected, but only less than two percent of argument mentions in ACE and TAC. Note that this phenomenon is distinct from the case of implicit arguments (Subsection 7.3.1), where for relational nouns the argument in the possessive position is omitted. Here, in contrast, no argument is omitted.

Metonymic References An interesting albeit rare finding in WEB is the occurrence of metonymic references, mostly affecting mentions of organizations or geo-political entities, like in the following example for the *employment tenure* relation:

Example 7.7 `... the White House's spokesman ...`

Here, the name of the official residence of the U.S. president is used as a simple way of referring to the US government. Examples like these are common among metonymic references, particularly for relations from the *business* domain. Another instance is shown in the following direct-speech example:

Example 7.8 `We're buying Pixar, we're not buying you.`

This is a statement made by Bob Iger to Steve Jobs; here, `We` refers to the Walt Disney Company, i.e., in the context of an *acquisition* mention, a person is used in place of an organization. This is another kind of resolution that is hard to do automatically.

Conclusion from Analysis Part 2 We drew two conclusions from the observations in this analysis, which confirm the conclusions from Subsection 7.3.1 with respect to the design of the new CS-RE approach. First, a large fraction of single-sentence relation mentions lack arguments that are mentioned somewhere else in the same document. This clearly shows the relevance and necessity of CS-RE. Second, a high percentage of the entity mentions that are arguments of relations require CR. This is noteworthy from a CS-RE point-of-view since CR techniques in literature are far from perfect, in particular in the presence of intricate phenomena like number disagreement, trigger-word overlap, and metonymy. Consequently, a linking strategy for relation mentions that assumes a perfect or near-perfect resolution of entities (like the strategy in Subsection 7.3.1) is doomed to fail. Instead, a strategy should incorporate information about the context and arguments of relation mentions flexibly, such that information from CR is used if available but its absence does not hinder the linking on the level of relations. In the next section, we design a method that meets these requirements.

7.4 Event Linking Approach

In this section, we describe a method for CS-RE which is more advanced than the simple strategy used in Subsection 7.3.1. This method approaches the problem of inter-sentential extraction via the task of event linking, which is concerned with resolving coreferences between recognized event mentions[79] in a document. The candidates for linking are represented via latent-feature vectors, which are produced by a method that learns from data to encode various important aspects of mentions and the surrounding context. This generic method also works in the presence of the phenomena observed in the earlier analysis. Furthermore, while it factors in CR results, it is not strictly dependent on them.

7.4.1 Problem Definition

We follow the notion of events from the ACE 2005 dataset (Linguistic Data Consortium, 2005b; Walker et al., 2006). Consider Example 7.9 (p. 169).

[79] Note again that in this thesis, we do not make a principle distinction between relations and events. Nevertheless, we use the term "event" in this subsection to be in line with related literature and competing systems.

Example 7.9 [80] British bank Barclays had agreed to **buy** Spanish rival Banco Zaragozano for 1.14 billion euros. The **combination** of the banking operations of Barclays Spain and Zaragozano will bring together two complementary businesses and will happen this year, in contrast to Barclays' postponed **merger** with Lloyds.

Processing these sentences in a prototypical IE-pipeline (Section 2.2) involves the following steps:

(a) At first, entity mentions are recognized.

(b) Next, words in the text are processed as to whether they elicit an event reference, i.e., event triggers are identified, and their semantic type is classified.

(c) The task of event extraction further requires that participants of recognized events are determined among the entity mentions in the same sentence, i.e., an event's arguments are identified and their semantic role with respect to the event is classified.

(d) Often, an IE system involves a disambiguation step of the entity mentions against one another in the same document (i.e., CR).

(e) The analogous task on the level of event mentions is called event linking (or: event co-reference resolution) and is the focus of this section.

In Example 7.9, entity mentions (step a) are underlined. Three words in the example sentences trigger event mentions (step b). All mentions are of type *Business.Merge-Org*; the triggers are shown in boldface. Step (c) produces the event mentions in Example 7.10:

Example 7.10

E1: – ⟨ British bank Barclays, Spanish rival Banco Zaragozano, 1.14 billion euros ⟩
  – trigger: **buy**

E2: – ⟨ Barclays Spain, Zaragozano, this year ⟩
  – trigger: **combination**

E3: – ⟨ Barclays, Lloyds ⟩
  – trigger: **merger**

[80] Based on an example by Araki and Mitamura (2015).

CR of entities (step d) allows identifying the three mentions of "*Barclays*" in the text as referring to the same real-world entity. Then, the task of event linking (step e) is to determine that E3 is a singleton reference, while E1 and E2 are co-referential, with the consequence that a document-level event instance can be produced from E1 and E2, listing four arguments (two companies, buying price, and acquisition date).

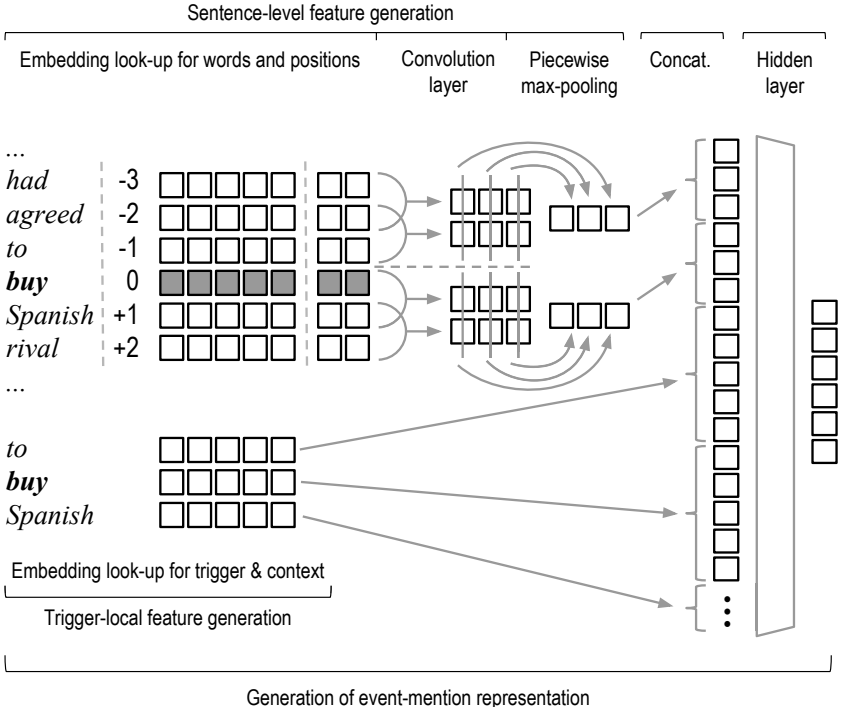
7.4.2 Model Design

This subsection first introduces common features for events, before describing the two-step architecture of our model.

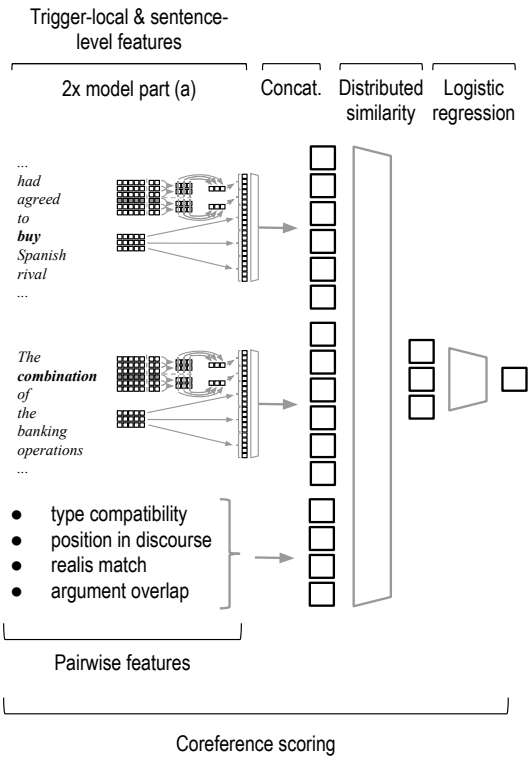
Event Features from Literature So far, a wide range of features was used for the representation of events and relations for extraction purposes (G. Zhou et al., 2005; Mintz et al., 2009; A. Sun et al., 2011) and co-reference resolution (Bejan and Harabagiu, 2010; H. Lee et al., 2012; Z. Liu et al., 2014; Araki and Mitamura, 2015; Cybulska and Vossen, 2015) purposes. The following list is an attempt to classify some of the most common features and examples thereof:

- **lexical:** surface string, lemma, word embeddings, context around trigger
- **syntactic:** depth of trigger in parse tree, dependency arcs from/to trigger
- **discourse:** distance between co-reference candidates, absolute position in document
- **semantic (intrinsic):** comparison of event arguments (entity fillers, present roles), event type of co-reference candidates
- **semantic (external):** similarity of co-reference candidates in lexical-semantic resources (WordNet, FrameNet) and other datasets (VerbOcean corpus), enrichment of arguments with alternative names from external sources (DBpedia, Geonames)

Lexical, discourse, and intrinsic-semantic features are available in virtually all application scenarios of event extraction/linking. Even syntactic parsing is no longer considered an expensive feature source. However, semantic features from external knowledge sources pose a significant burden on the application of event processing systems, as these sources are created at high cost and come with limited domain coverage (Section 2.5). Fortunately, recent work explored the use of a new feature class for tackling relation-/event-extraction related tasks with neural networks: sentential features (Zeng et al., 2014; T. H. Nguyen and Grishman, 2015a; Y. Chen et al., 2015; dos Santos et al., 2015; Zeng et al., 2015). These approaches showed that processing sentences with neural models yields representations suitable for IE, which motivates their use in our approach.



Model part (a).



Model part (b).

Figure 7.3 – The two parts of the event-linking model. The first part computes a representation for a single event mention. The second part is fed with two such event-mention representations plus a number of pairwise features for the input event-mention pair, and calculates a co-reference score.

Learning Event Representations The architecture of the model (Figure 7.3, p. 171) is split into two parts. The first one aims at adequately representing individual event mentions. As is common in literature, words of the whole sentence of an input event mention are represented as real-valued vectors v_w^i of a fixed size d_w, with i being a word's position in the sentence. These *word embeddings* are updated during model training and are stored in a matrix $W_w \in \mathbb{R}^{d_w \times |V|}$; $|V|$ being the vocabulary size of the dataset.

Furthermore, we take the relative position of tokens with respect to the mention into account, as suggested by Collobert et al. (2011) and Zeng et al. (2014). The rationale is that while the absolute position of learned features in a sentence might not be relevant for an event-related decision, the position of them with respect to the event mention is indeed crucial. Embeddings v_p^i of size d_p for relative positions of words are generated in a way similar to word embeddings, i.e., by table lookup from a matrix $W_p \in \mathbb{R}^{d_p \times s_{\max} * 2-1}$ of trainable parameters. Again i denotes the location of a word in a sentence; $s_{\max}$ is the maximum sentence length in the dataset. Embeddings for words and positions are concatenated into vectors v_t^i of size $d_t = d_w + d_p$. This means that now every word in the vocabulary has a different representation for each distinct distance to an event trigger with which it occurs.

A sentence with s words is represented by a matrix of dimensions $s \times d_t$. This matrix serves as input to a convolution layer. In order to compress the semantics of s words into a sentence-level feature vector with constant size, the convolution layer applies d_c filters to each window of n consecutive words, thereby calculating d_c features for each n-gram of a sentence. For a single filter $w_c \in \mathbb{R}^{n * d_t}$ and particular window of n words starting at position i, this operation is defined as

Equation 7.1
$$v_c^i = relu(w_c \cdot v_t^{i:i+n-1} + b_c),$$

where $v_t^{i:i+n-1}$ is the flattened concatenation of vectors $v_t^{(\cdot)}$ for the words in the window, b_c is a bias, and *relu* is the activation function of a rectified linear unit. In Figure 7.3, $d_c = 3$ and $n = 2$.

In order to identify the most indicative features in the sentence and to introduce invariance for the absolute position of these, we feed the n-gram representations to a max-pooling layer, which identifies the maximum value for each filter. We treat n-grams on each side of the trigger word separately during pooling, which allows the model to handle multiple event mentions per sentence, similar in spirit to Y. Chen et al. (2015) and Zeng et al. (2015). The pooling step for a particular convolution filter $j \in [1, d_c]$ and

sentence part $k \in \{\mapsto, \mapsto\}$ is defined as

Equation 7.2
$$v_m^{j,k} = max(v_c^i),$$

where i runs through the convolution windows of k. The output of this step are sentential features $v_{\text{sent}} \in \mathbb{R}^{2*d_c}$ of the input event mention:

Equation 7.3
$$v_{\text{sent}} = (v_m^{1,\mapsto}, \ldots, v_m^{d_c,\mapsto}, v_m^{1,\mapsto}, \ldots, v_m^{d_c,\mapsto})$$

Additionally, we provide the network with trigger-local, lexical-level features by concatenating v_{sent} with the word embeddings $v_w^{(\cdot)}$ of the trigger word and its left and right neighbor, resulting in $v_{\text{sent+lex}} \in \mathbb{R}^{2*d_c+3*d_w}$. This encourages the model to take the lexical semantics of the trigger into account, as these can be a strong indicator for co-reference. The result is processed by an additional hidden layer, generating the final event-mention representation v_e with size d_e used for the subsequent event-linking decision:

Equation 7.4
$$v_e = tanh(W_e v_{\text{sent+lex}} + b_e).$$

Learning Co-reference Decisions The second part of the model, as represented in Figure 7.3b, processes the representations for two event mentions v_e^1, v_e^2, and augments these with pairwise comparison features v_{pairw} to determine the compatibility of the event mentions. The following features are used, in parentheses we give the feature value for the pair E1, E2 from Example 7.10 (p. 169):

- Coarse-grained and/or fine-grained event-type agreement (yes, yes)
- Antecedent event is in first sentence (yes)
- Bagged distance between event mentions in terms of number of sentences and number of intermediate event mentions (1, 0)
- Agreement in event modality (yes)
- Overlap in arguments (two shared arguments)

The concatenation of these vectors

Equation 7.5
$$v_{\text{sent+lex+pairw}} = (v_e^1, v_e^2, v_{\text{pairw}})$$

is processed by a single-layer neural network which calculates a distributed similarity of size d_{sim} for the two event mentions:

Equation 7.6
$$v_{\text{sim}} = square(W_{\text{sim}} v_{\text{sent+lex+pairw}} + b_{\text{sim}}).$$

The use of the square function as the network's non-linearity is backed by the intuition that for measuring similarity, an invariance under polarity changes is desirable. Having d_{sim} similarity dimensions allows the model to learn multiple similarity facets in parallel; in our experiments, this setup outperformed model variants with different activation functions as well as a cosine-similarity based comparison.

To calculate the final output of the model, v_{sim} is fed to a logistic-regression classifier, whose output serves as the co-reference score:

Equation 7.7 $$score = \sigma(W_{\text{out}} v_{\text{sim}} + b_{\text{out}})$$

We train the model parameters in Equation 7.8 by minimizing the logistic loss over shuffled mini-batches with gradient descent using Adam (Kingma and Ba, 2014).

Equation 7.8 $$\theta = \{W_w, W_p, \{w_c\}, \{b_c\}, W_e, b_e, W_{\text{sim}}, b_{\text{sim}}, W_{\text{out}}, b_{\text{out}}\}$$

7.4.3 Example Generation and Clustering

We investigated two alternatives for the generation of examples from documents with recognized event mentions. Algorithm 7.1 (p. 175) shows the strategy we found to perform best. It iterates over the event mentions of a document and pairs each mention (the "anaphors") with all preceding ones (the "antecedent" candidates). This strategy applies to both training and inference time. Soon et al. (2001) proposed an alternative strategy, which during training creates positive examples only for the closest actual antecedent of an anaphoric event mention, with intermediate event mentions serving as negative antecedent candidates. In our experiments, this strategy performed worse than the less elaborate algorithm in Algorithm 7.1.

The pairwise co-reference decisions of our model induce a clustering of a document's event mentions. In order to force the model to output a consistent view on a given document, a strategy for resolving conflicting decisions is needed. We followed the strategy as described in Algorithm 7.2 (p. 175), which builds the transitive closure of all positive links. Additionally, we experimented with V. Ng and Cardie (2002)'s "BestLink" strategy, which discards all but the highest-scoring antecedent of an anaphoric event mention. Z. Liu et al. (2014) reported that for event linking, BestLink outperforms naive transitive closure, however, in our experiments (Subsection 7.4.5) we could not confirm their finding.

1: **procedure** GENERATEEXAMPLES($\mathcal{M}_d$)**:**
2: $\mathcal{M}_d = (m_1, \ldots, m_{|\mathcal{M}_d|})$
3: $\mathcal{P}_d \leftarrow \emptyset$
4: **for** $i = 2, \ldots, |\mathcal{M}_d|$ **do**
5: **for** $j = 1, \ldots, i-1$ **do**
6: $\mathcal{P}_d \leftarrow \mathcal{P}_d \cup \{(m_i, m_j)\}$
7: **return** $\mathcal{P}_d$

Algorithm 7.1 – Generation of examples $\mathcal{P}_d$ for a document d with a sequence of event mentions $\mathcal{M}_d$.

1: **procedure** GENERATECLUSTERS($\mathcal{P}_d$, *score*)**:**
2: $\mathcal{P}_d = \{(m_i, m_j)\}_{i,j}$
3: $score : \mathcal{P}_d \mapsto [0, 1]$
4: $C_d \leftarrow \{(m_i, m_j) \in \mathcal{P}_d : score(m_i, m_j) > 0.5\}$
5: **while** $\exists (m_i, m_k), (m_k, m_j) \in C_d : (m_i, m_j) \notin C_d$ **do**
6: $C_d \leftarrow C_d \cup \{(m_i, m_j)\}$
7: **return** C_d

Algorithm 7.2 – Generation of event clusters C_d for a document d based on the co-reference scores from the model. $\mathcal{P}_d$ is the set of all event-mention pairs from a document, as implemented in Algorithm 7.1.

7.4.4 Experimental Setting and Model Training

We implemented our model using the TensorFlow framework (Abadi et al., 2016a, version 0.6) and chose the ACE 2005 dataset (Walker et al., 2006, later: ACE) as our main testbed. The annotation of this corpus focuses on the event types *Conflict.Attack*, *Movement.Transport*, and *Life.Die*, which report about terrorist attacks, movement of goods and people, as well as of deaths of people; but it also contains many more related event types as well as mentions of business-relevant and judicial events. The corpus consists of merely 599 documents, which is why we created a second dataset that encompasses these documents and additionally contains 1351 more documents annotated analogously with the same set of event types. The additional documents comprise the corpus referred to as TAC in Subsection 7.3.2, as well as further TAC texts with annotation. We refer to this second dataset as ACE^{++}. Both datasets were split 9:1 into a development (*dev*) and *test* partition; we further split *dev* 9:1 into a training (*train*) and validation (*valid*) partition. Table 7.3 (p. 176) lists statistics for the datasets.

There are a number of architectural alternatives in the model as well as hyperparameters to optimize. Apart from varying the size of intermediate representations in the model (d_w, d_p, d_c, d_e, d_{sim}), we experimented with different convolution window sizes n, activation functions for the similarity-function layer in model part (b), whether to use the

	ACE	ACE^{++}
#documents	599	1950
#event instances	3617	7520
#event mentions	4728	9956

Table 7.3 – Properties of datasets in event-linking experiment.

d_w	300	η	10^{-5}
d_p	8	β_1	0.2
d_c	256	β_2	0.999
d_e	50	ϵ	10^{-2}
d_{sim}	2	batch size	512
n	3	epochs	≤ 2000
Dropout	no	$\ell 2$ reg.	no

Table 7.4 – Hyperparameter settings.

dual pooling and final hidden layer in model part (a), whether to apply regularization with $\ell 2$ penalties or Dropout, and parameters to Adam ($\eta, \beta_1, \beta_2, \epsilon$). We started our exploration of this space of possibilities from previously reported hyperparameter values (Y. Zhang and Wallace, 2015; Y. Chen et al., 2015) and followed a combined strategy of random sampling from the hyperparameter space (180 points) and line search. The optimization process involved training on ACE^{++}_{train} and evaluation on ACE^{++}_{valid}. The final settings we used for all of the following experiments are listed in Table 7.4. W_w is initialized with pre-trained embeddings of Mikolov et al. (2013b)[81], while the embedding matrix for relative positions (W_p) and all other model parameters were initialized randomly. Model training was run for 2000 epochs, after which the best model on the respective *valid* partition was selected.

7.4.5 Evaluation

This subsection elaborates on the conducted experiments. First, we compare our approach to state-of-art systems on dataset ACE, after which we report experiments on ACE^{++}, where we contrast variations of our model to gain insights into the impact of the utilized feature classes. We conclude this section with an error analysis.

[81] `https://code.google.com/archive/p/word2vec/`; last access: 2017-05-18.

	BLANC			B-CUBED			MUC			Positive links		
	4 ∗ (Precision / Recall / F1 score) in %											
Our model	**71.80**	**75.16**	**73.31**	**90.52**	86.12	88.26	**61.54**	45.16	**52.09**	47.89	**56.20**	**51.71**
Z. Liu et al. (2014)	70.88	70.01	70.43	89.90	**88.86**	**89.38**	53.42	**48.75**	50.98	**55.86**	40.52	46.97
Bejan and Harabagiu (2010)	—	—	—	83.4	84.2	83.8	—	—	—	43.3	47.1	45.1
Sangeetha and Arock (2012)	—	—	—	—	—	87.7	—	—	—	—	—	—

Table 7.5 – Event-linking performance of several systems on ACE. Best value per metric in bold.

Comparison to State-of-the-Art on ACE Table 7.5 depicts the performance of our model, trained on ACE_{train}, on ACE_{test}, along with the performance of state-of-the-art systems from literature. From the wide range of proposed metrics for the evaluation of co-reference resolution, we believe BLANC (Recasens and Hovy, 2011) has the highest validity, as it balances the impact of positive and negative event-mention links in a document. Negative links and consequently singleton event mentions are more common in this dataset (more than 90% of links are negative). As Recasens and Hovy point out, the informativeness of metrics like MUC (Vilain et al., 1995), B-CUBED (Bagga and Baldwin, 1998), and the naive positive-link metric suffers from this imbalance. Nevertheless, we add these metrics for completeness, and because BLANC scores are not available for all systems.

Unfortunately, there are two caveats to this comparison. First, while a 9:1 train/test split is the commonly accepted way of using ACE, the exact documents in the partitions vary from system to system. We are not aware of any publicized split from previous work on event linking, which is why we create our own.[82] Second, published methods follow different strategies regarding preprocessing components. While *all* systems in Table 7.5 use gold-annotated event-mention *triggers*, Bejan and Harabagiu (2010) and Z. Liu et al. (2014) use a semantic-role labeler and other tools instead of gold-argument information. We argue that using full gold-annotated event mentions is reasonable to mitigate error propagation along the extraction pipeline and make performance values for the task at hand more informative.[83]

We beat Z. Liu et al. (2014)'s system in terms of F1 score on BLANC, MUC, and positive-links, while their system performs better in terms of B-CUBED. Even when taking into account the caveats mentioned above, it seems justified to assess

[82] We announced the list of documents in ACE_{valid}/ACE_{test} at `https://git.io/vwEEP`.

[83] Note that for this reason, we also decided against using the pattern-based methods from Chapters 3 and 4 for the detection of intra-sentential event mentions. Nevertheless, the presented event-linking approach is a downstream step of pattern-based RE and is fully compatible with it.

	Model	Dataset	BLANC (P/R/F1 in %)
1)	Subsection 7.4.2	ACE	71.80 **75.16** **73.31**
2)	Subsec. 7.4.2 + BestLink	ACE	**75.68** 69.72 72.19
3)	Subsection 7.4.2	ACE^{++}	73.22 **83.21** **76.90**
4)	Subsec. 7.4.2 + BestLink	ACE^{++}	**74.24** 68.86 71.09

Table 7.6 – Impact of data amount and clustering.

that our model performs in general on-par with their state-of-the-art system. Their approach involves random-forest classification with best-link clustering and propagation of attributes between event mentions, and is grounded on a manifold of external feature sources, i.e., it uses a "rich set of 105 semantic features". Thus, their approach is strongly tied to domains where these semantic features are available and is potentially hard to port to other text kinds. In contrast, our approach does not depend on resources with restricted domain availability.

Bejan and Harabagiu (2010) proposed a non-parametric Bayesian model with standard lexical-level features and WordNet-based similarity between event elements. We outperform their system in terms of B-CUBED and positive-links, which indicates that their system tends to over-merge event mentions, i.e., has a bias against singletons. They used a slightly bigger variant of ACE with 46 additional documents in their experiments.

Sangeetha and Arock (2012) hand-crafted a similarity metric for event mentions based on the number of shared entities in the respective sentences, lexical terms, synsets in WordNet, which served as input to a mincut-based cluster identification. Their system performs well in terms of B-CUBED F1. However, their paper provided not enough details about the exact experimental setup, which makes it difficult to interpret this result.

Another approach with results on ACE was presented by Z. Chen et al. (2009), who employed a maximum-entropy classifier with agglomerative clustering and lexical, discourse, and semantic features, e.g., also a WordNet-based similarity measure. However, they reported performance using a threshold optimized on the test set, thus we decided to not include the performance here.

Further Evaluation on ACE and ACE^{++} We now look at several aspects of the model performance to gain further insights about it's behavior. Table 7.6 shows the impact of increasing the amount of training data (ACE $\rightarrow$ ACE^{++}). This increase (rows 1, 3) leads to a boost in recall, from 75.16% to 83.21%, at the cost of a small decrease in precision.

	Pw	Loc	Sen	Dataset	BLANC (P/R/F1 in %)
1)	✓			ACE^{++}	57.45 68.16 56.69
2)	✓	✓		ACE^{++}	62.24 76.23 64.12
3)	✓	✓	✓	ACE^{++}	73.22 **83.21** **76.90**
4)	✓		✓	ACE^{++}	**82.60** 70.71 74.97
5)		✓	✓	ACE^{++}	59.67 66.25 61.28
6)			✓	ACE^{++}	58.38 55.85 56.70

Table 7.7 – Impact of feature classes on event-linking performance. "Pw" is short for pairwise features, "Loc" refers to trigger-local lexical features, "Sen" corresponds to sentential features.

This indicates that the model can generalize much better using the additional training data. Looking into the use of the alternative clustering strategy BestLink recommended by Z. Liu et al. (2014), we can confirm the expected observation of a precision improvement (row 1 vs. 2; row 3 vs. 4). This can be explained by the fact that less positive links are used before the transitive-closure clustering takes place. This is however outweighed by a large decline in recall, resulting in a lower F1 score (73.31 → 72.19; 76.90 → 71.09). The better performance of BestLink in Liu et al.'s model suggests that our model already weeds out many low confidence links in the classification step, which makes a downstream filtering unnecessary in terms of precision, and even counter-productive regarding recall.

Table 7.7 shows our model's performance when particular feature classes are removed from the model,[84] row 3 corresponding to the full model as described in Subsection 7.4.2. It is not a surprise that classifying examples with just pairwise features (row 1) results in the worst performance, while adding first trigger-local lexical features (row 2) and sentential features (row 3) subsequently raises both precision and recall. Just using pairwise features and sentential ones (row 4), boosts precision, which is counter-intuitive at first, but may be explained by a different utilization of the sentential-feature part of the model during training. This part is then adapted to focus more on the trigger-word aspect, meaning the sentential features degrade to trigger-local features. While this allows to reach higher precision (since more than fifty percent of positive examples have trigger-word agreement), it substantially limits the model's ability to learn other co-reference-relevant aspects of event-mention pairs, leading to a low recall. Further considering rows 5 and 6, we can conclude that all feature classes indeed positively contribute to the overall model performance.

[84] Feature classes were removed during both training and evaluation.

Model	Dataset	BLANC (P/R/F1 in %)
Subsection 7.4.2	ACE^{++}	**73.22** 83.21 **76.90**
All singletons	ACE^{++}	45.29 50.00 47.53
One instance	ACE^{++}	4.71 50.00 8.60
Same type	ACE^{++}	62.73 **84.75** 61.35

Table 7.8 – Event-linking performance of our model against naive baselines.

The result of applying three naive baselines to ACE^{++} is shown in Table 7.8. The *all singletons*/*one instance* baselines predict every input link to be negative or positive, respectively. In particular, the all-singletons baseline performed well, due to the large fraction of singleton event mentions in the dataset. The third baseline, *same type*, predicts a positive link whenever there is agreement on the event type. Note that this baseline fails for documents with multiple instances of the same event type. For example, a news article that compared a recent terrorist attack to one from the past could not be handled properly. This baseline also performed quite well, in particular in terms of recall, but it also showed low precision.

Error analysis We manually investigated a sample of 100 false positives and 100 false negatives from ACE^{++} to get an understanding of system errors. It turns out that a significant portion of the false negatives would involve the resolution of a pronoun to a previous event mention, a very hard and yet unsolved problem. Consider the following examples:

Example 7.11 "It's crazy that we're **bombing** Iraq. **It** sickens me."

Example 7.12 "Some of the slogans sought to rebut **war** supporters' arguments that the protests are unpatriotic. [...] Nobody questions whether **this** is right or not."

In both examples, the event mentions (trigger words in bold font) are gold-annotated as co-referential, but our model failed to recognize this.

Another observation is that for 17 false negatives, we found analogous cases among the sampled false positives where annotators made a different annotation decision. Consider Examples 7.13 and 7.14:

Example 7.13 The 1860 Presidential **Election**. [...] Lincoln **won** a plurality with about 40% of the vote.

Example 7.14 She **lost** her seat in the 1997 **election**.

Each example has two event mentions (with triggers in bold font) taken from the same document and referring to the same event type, i.e., *Personnel.Elect*. While in the first example, the annotators identified the mentions as co-referential, the second pair of mentions is not annotated as such. Analogously, 22 out of the 100 analyzed false positives were cases where the misclassification of the system was plausible to a human rater. This exemplifies that this task has many boundary cases were a positive/negative decision is difficult to make even for expert annotators.

7.5 Related Work

Here, we briefly point at further resources that were used in literature for work on event linking. Other relevant approaches for this task and similar ones were already surveyed in Section 2.4. Section 2.6 introduced the wide range of NN-based methods for NLP.

Apart from the ACE 2005 corpus, some other datasets with event-co-reference annotation were presented in the past. Hovy et al. (2013a) reported on the annotation process of two corpora from the domains of "violent events" and biographic texts; to our knowledge neither of them is publicly available. OntoNotes (Weischedel et al., 2013) comprises different annotation layers including co-reference (Pradhan et al., 2012), however it intermingles entity and event co-reference. A series of releases of the EventCorefBank corpus (Bejan and Harabagiu, 2010; H. Lee et al., 2012; Cybulska and Vossen, 2014) combine linking of event mentions within and across documents, for which Z. Liu et al. (2014) reported a lack of completeness on the within-document aspect. The ProcessBank dataset (Berant et al., 2014) provides texts with event links from the difficult biological domain.

7.6 Summary

This chapter presented work on CS-RE in three aspects. First, we described the annotation process of a new dataset that was specifically created for the analysis of challenges in document-level fact extraction. Second, we analyzed this new dataset and three additional corpora and reported our observations on the complexity and challenges of CS-RE. Third, we proposed a model for the task of event linking, which achieved state-of-the-art results without relying on external feature sources. With respect to competing systems from literature, this model in particular showed that low linking performance, coming from a

lack of semantic knowledge about a domain, is evitable. In addition, our experiments give further empirical evidence for the significance of neural models for generating latent-feature representations of sentences.

There are several directions for potential future work. As intermediate next steps, the model could be tested on more datasets and task variations, i.e., in a cross-document setting or for joint trigger identification and co-reference resolution. Furthermore, separating anaphoricity detection from antecedent scoring, as is often done for the task of entity co-reference resolution (e.g., by Wiseman et al., 2015), might result in gains in performance; also the generation of sentential features from recurrent neural networks seems promising. Regarding medium-term research, it could be investigated if the model benefited from more fine-grained information about the underlying discourse structure of a text. This could guide the model when encountering the problematic case of pronoun resolution, as described in the error analysis.

Chapter 8

Conclusions

Contents

This thesis described novel methods for IE and related research areas, all of which either employed linguistic patterns or dealt with their limitations. At the beginning of this thesis, we presented **three research questions**. These questions covered different aspects of the use of linguistic patterns: **(a)** how patterns can be generated, filtered, and disambiguated in traditional and open extraction settings, **(b)** how generated patterns can be transformed into a new kind of knowledge resource, and **(c)** how extraction beyond the sentence level should be approached. The questions were addressed by **four main contributions**. We briefly summarize these contributions in the first part of this chapter. Then, we attempt a brief outlook into future research directions.

8.1 Summary of Main Contributions

This dissertation advanced the state-of-the-art in several aspects of the automatic processing of language.

Innovative Filtering and Generation of Patterns The Web-DARE system provides means for collecting linguistic patterns for selected target relations from web texts. It also incorporates a basic filtering strategy for the produced patterns (**Chapter 3**), which, however, does not fully address the high amount of noise, in particular, frequent off-topic patterns. The proposal of **Chapter 4** was to employ lexical-semantic knowledge about the target relations for the identification of such noisy patterns. Experiments showed that this idea indeed results in a performance leap. We subsequently combined both the base Web-DARE filter and the new lexical-semantics filter, which improved performance even further. Another deficiency of the base pattern-collection system was its shortest-path-based heuristic for the generation of patterns from relation mentions. This heuristic failed to capture the relation-relevant parts of sentences in important cases. We extended the pattern-generation algorithm with access to relation-relevant terms, which allowed it to produce more informative patterns, as indicated by experimental results. A further minor contribution in this direction was the creation of an annotated RE dataset (called CELEBRITY) that features mention-level annotation of text and precisely locates the arguments of relations.

Scalable Disambiguation of Open-Domain Patterns We also handled scenarios where a relational schema is either not present or where this schema is not as detailed as in traditional IE settings. Methods for such settings were previously proposed in literature, but they lacked scalability due to overly complex disambiguation approaches or were of generally low accuracy. **Chapter 5** introduced a neural method for the identification of paraphrases that is based on a recent disambiguation approach for words in vast corpora. We defined a training signal that can be derived from the automatic entity annotation of texts and allows to train a model in a weakly supervised way. In this method, observed linguistic patterns are mapped into a latent-feature space, which represents fine-grained semantic similarities between the patterns via their respective distance in the space. In order to produce groups of synonymous patterns, standard clustering techniques can be applied. We reported direct experimental comparisons against two strong baselines and observed that our approach significantly outperformed these on several datasets.

Resource Building from Collected Patterns Linguistic patterns have more applications than just the detection and extraction of relational knowledge from text. For example, they convey interesting information about everyday language use and its properties. The contribution of **Chapter 6** is to show how the collected patterns for a specific set of relations can be compiled into resources called sar-graphs, which make language expressions available to linguists and language enthusiasts. We elaborated in this chapter that sar-graphs fill a gap in the current resource landscape, more precisely, they are situated between factual KBs and repositories of lexical-semantic information. Furthermore, we presented linking strategies that tightly integrate sar-graphs on several levels with other resources. Finally, we conducted and discussed an analysis of the explicitness with which patterns refer to relations, and we showed that various degrees of implicitness are present in sar-graphs.

Extraction Beyond Individual Sentences While the recognition of entity mentions and the handling of their co-reference relationships is traditionally taking place on the level of documents and is not restricted to individual sentences, the prevalent methods for the extraction of relations and events from text still focus on individual sentences. We started **Chapter 7** by creating a corpus of cross-sentential mentions for analysis purposes (called COCKRACE) and continued by conducting an investigation of this corpus and further datasets from RE literature to quantify the need for cross-sentential processing. The result of this analysis is that such processing would facilitate higher extraction performance. Consequently, we designed an approach for the extraction of document-level factual mentions. This approach implemented a neural co-reference resolver for mentions of event information in texts. In our experiments, the new approach performed on par with state-of-the-art systems from literature. However, in contrast to them, our approach does not rely on large sets of domain-specific features. Thus, our system is less restricted in its applications and easier to port to new domains.

8.2 Future Research Directions and Applications

In the following, we describe potential next steps for the further development of the approaches in this thesis. Also, we briefly elaborate on two application domains that motivate future research directions for NLP techniques.

Dependence on Resources and Languages Human language usage exhibits great variety and also gradually changes over time. To cope with these properties, computational strategies should be able to learn from examples in a bottom-up fashion and furthermore should not depend on resources which are highly specific to some domains and individual natural languages. While the methods presented in Chapters 5 and 7 do not rely on domain-specific knowledge, the approaches in Chapters 4 and 6 do employ resources with a particular domain and restricted language availability.[85] Thus, important next steps are work on reducing the dependence on such resources, and additionally the extension of resources to further domains. Another aspect of language limitations is that the methods in all chapters of this thesis rely on language-specific preprocessing tools, for example, POS tagging and parsing. While such tools are often available for languages other than English, their quality is much worse when processing non-English text. Hence, more effort needs to be invested into the development of high-performing tools for more languages (Rehm and Uszkoreit, 2013). Another research avenue is transfer learning between multiple languages (perhaps from the same family), which could help to compensate gaps in language-specific training datasets.

Combinations of Learning Settings and Tasks Some types of data are abundant, for example, plain texts on the web, while others are not that widely available, like fine-grained expert annotation for these web documents. Furthermore, annotated data for some NLP tasks like NER can be created with much less effort and training cost of expert raters than what is required for tasks like parsing. Future methods in NLP need to exploit more than just one kind of such data kind during training (Kaiser et al., 2017). Ideally, they can be optimized via multiple training paradigms, where each paradigm (like supervised learning or bootstrapping) enables the learning from data of the corresponding type. The similar idea of multi-task learning (Collobert et al., 2011) facilitates the production of more expressive representations of linguistic elements in neural methods. These representations are learned by exploiting multiple training signals coming from data for different tasks. The integration of RE methods into such a multi-task training paradigm is a promising direction for future work. Furthermore, the learning from feedback, either via system-initiated requests to label borderline examples in an active-learning fashion, or via the incorporation of interactive forms of reinforcement learning (e.g., in dialogues, Shah et al., 2016) are directions for future work.

[85] The factual and lexical-semantic knowledge employed in these chapters is dominated by what is covered in Wikipedia and hence is restricted to what is of interest to Wikipedia authors.

Implicitness of Language Practical language use for communication purposes is efficient and heavily relies on context for conveying information. Future research should explore how structured and unstructured knowledge could be employed for the reasoning about language utterances, and perhaps find generic approaches for the modeling of contextual information for NLP. Recent research popularized text-contextual representations of words via word2vec or recurrent networks (Section 2.6), which integrate the meaning of words with the semantics of surrounding text. What is still missing is the consideration of common-sense knowledge when text is processed (a classic problem in AI). Moreover, situational knowledge and embodied intelligence are forms of context that can offer important guidance to interpret a text. Finally, new techniques and neural models can help to produce more expressive representations for language fragments. A recent development in this direction is the application of generative adversarial networks (Goodfellow et al., 2016, pp. 690–693) to improve the representations of text (Bowman et al., 2016). Such approaches may also help for the extraction of facts from texts.

IE for Personal Digital Assistants In the introduction (Chapter 1), we mentioned digital assistants as one important application of NLP technology. Many types of dialogues exist in which humans can engage with digital agents. A particularly important type are systems which let users ask questions to explore factual information and other kinds of knowledge. Traditional-IE and open-IE methods are of great utility in this dialogue scenario, as they populate KBs with information that might be of interest to users. On top of that, IE methods allow determining a user's intent in a dialogue, namely to identify the kind of information that users are looking for. As it is reasonable to assume that digital assistants will gain further popularity in the near future, there will likely be high demand for advanced IE technology in the future, too. Another important aspect of dialogues, apart from interpreting user intent and answering questions, is naturalness and fluency. These aspects require understanding not only the prior dialogue but also more generally what makes dialogues sound coherent. For example, reiterating information can produce dialogues that are perceived as repetitive, but completely avoiding redundancy can cause incoherence (Krause et al., 2017). This is an area where computational, data-driven machine learning methods should be augmented with expert linguists' formalizations of how discourse and dialogues are structured.

Journalism and Business Intelligence Fully automated journalism is a vision that requires future breakthroughs in many facets of NLP, most notably generation of long texts.[86] An example of an existing system in this space is Quakebot (Oremus, 2014), a program for the automatic generation of news reports about earthquake events. Current systems produce texts of a quality that clearly exposes them as being algorithmically generated, however, future research on the fluency and naturalness of generation systems could change that. Further research directions are the development of more involved systems that would also double-check news before processing them, as well as an assessment of the news-worthiness of information (urgency, relevance for audience). A more distant vision is the design of systems that automatically identity aspects of a story which lack information and pro-actively identify information sources that could fill this gap.

[86] In Chapter 6, we briefly mentioned the experimental use of sar-graphs for natural-language generation in a system for computer-assisted language learning. Note, however, that in this experiment, only sentence-long texts were produced.

Bibliography

This bibliography is sorted alphabetically by the last name of the authors. Before each entry, we print the citation label that was used in the main part of the thesis to refer to this entry.

[AAAI, 2007]
AAAI (2007). *Machine Reading, Papers from the AAAI Spring Symposium*. Tech. rep. SS-07-06. AAAI Press.

[Abadi et al., 2016a]
Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Gregory S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian J. Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Józefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Gordon Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul A. Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda B. Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng (2016a). "TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems". In: *CoRR* abs/1603.04467. arXiv: 1603.04467 [cs.DC].

[Abadi et al., 2016b]
Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek Gordon Murray, Benoit Steiner, Paul A. Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng (2016b). "TensorFlow: A System for Large-Scale Machine Learning". In: *12th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2016*. USENIX Association, pp. 265–283.

[Adolphs et al., 2011]
Peter Adolphs, Feiyu Xu, Hans Uszkoreit, and Hong Li (Oct. 2011). "Dependency Graphs as a Generic Interface Between Parsers and Relation Extraction Rule Learning". In: *Proceedings of the 34th Annual German Conference on Artificial Intelligence (KI-2011)*. Vol. 7006. Lecture Notes in Computer Science. Springer, pp. 50–62.

[Agichtein, 2006]
Eugene Agichtein (Apr. 2006). "Confidence Estimation Methods for Partially Supervised Information Extraction". In: *Proceedings of the Sixth SIAM International Conference on Data Mining*. SIAM, pp. 539–543. DOI: 10.1137/1.9781611972764.56.

[Agichtein and Gravano, 2000]
Eugene Agichtein and Luis Gravano (2000). "Snowball: Extracting Relations from Large

Plain-Text Collections". In: *DL '00: Proceedings of the Fifth ACM Conference on Digital Libraries*. ACM, pp. 85–94. DOI: 10.1145/336597.336644.

[Agrawal et al., 2012]
Divyakant Agrawal, Philip Bernstein, Elisa Bertino, Susan Davidson, Umeshwar Dayal, Michael Franklin, Johannes Gehrke, Laura Haas, Alon Halevy, Jiawei Han, H. V. Jagadish, Alexandros Labrinidis, Sam Madden, Yannis Papakonstantinou, Jignesh M. Patel, Raghu Ramakrishnan, Kenneth Ross, Cyrus Shahabi, Dan Suciu, Shiv Vaithyanathan, and Jennifer Widom (2012). *Challenges and Opportunities with Big Data — A Community White Paper Developed by Leading Researchers Across the United States*. Tech. rep. http://cra.org/ccc/docs/init/bigdatawhitepaper.pdf, last access: 2015-11-09. Computing Community Consortium committee of the Computing Research Association.

[Aguilar et al., 2014]
Jacqueline Aguilar, Charley Beller, Paul McNamee, Benjamin Van Durme, Stephanie Strassel, Zhiyi Song, and Joe Ellis (June 2014). "A Comparison of the Events and Relations Across ACE, ERE, TAC-KBP, and FrameNet Annotation Standards". In: *Proceedings of the Second Workshop on EVENTS: Definition, Detection, Coreference, and Representation*. ACL, pp. 45–53. DOI: 10.3115/v1/W14-2907.

[Ahn, 2006]
David Ahn (July 2006). "The Stages of Event Extraction". In: *Proceedings of the Workshop on Annotating and Reasoning About Time and Events*. ACL, pp. 1–8.

[Ai and Krause et al., 2015]
Renlong Ai, Sebastian Krause, Walter Kasper, Feiyu Xu, and Hans Uszkoreit (2015). "Semi-Automatic Generation of Multiple-Choice Tests from Mentions of Semantic Relations". In: *Proceedings of Workshop on Natural Language Processing Techniques for Educational Applications*. ACL, pp. 26–33. DOI: 10.18653/v1/W15-4405.

[Akbik, 2016]
Alan Akbik (2016). "Exploratory Relation Extraction in Large Multilingual Data". PhD thesis. Technische Universität Berlin, Institute of Software Engineering and Theoretical Computer Science. DOI: 10.14279/depositonce-5438.

[Akbik and Michael, 2014]
Alan Akbik and Thilo Michael (May 2014). "The Weltmodell: A Data-Driven Commonsense Knowledge Base". In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. ELRA, pp. 3272–3276.

[Akbik et al., 2014]
Alan Akbik, Thilo Michael, and Christoph Boden (Aug. 2014). "Exploratory Relation Extraction in Large Text Corpora". In: *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. ACL, pp. 2087–2096.

[Akbik et al., 2012]
Alan Akbik, Larysa Visengeriyeva, Priska Herger, Holmer Hemsen, and Alexander Löser (2012). "Unsupervised Discovery of Relations and Discriminative Extraction Patterns". In: *COLING 2012, 24th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers*, pp. 17–32.

[Akbik et al., 2013]
Alan Akbik, Larysa Visengeriyeva, Johannes Kirschnick, and Alexander Löser (Oct. 2013). "Effective Selectional Restrictions for Unsupervised Relation Extraction". In: *Proceedings of the Sixth International Joint Conference on Natural Language Processing*. AFNLP, pp. 1312–1320.

[Al-Rfou et al., 2016]
Rami Al-Rfou et al. (2016). "Theano: A Python Framework for Fast Computation of Mathematical Expressions". In: *CoRR* abs/1605.02688. arXiv: 1605.02688 [cs.SC].

[Alfonseca et al., 2012]
Enrique Alfonseca, Katja Filippova, Jean-Yves Delort, and Guillermo Garrido (2012). "Pattern Learning for Relation Extraction with a Hierarchical Topic Model". In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics - Volume 2: Short Papers*, pp. 54–59.

[Alfonseca et al., 2013]
Enrique Alfonseca, Daniele Pighin, and Guillermo Garrido (Aug. 2013). "HEADY: News Headline Abstraction Through Event Pattern Clustering". In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. ACL, pp. 1243–1253.

[Allen, 1987]
James Allen (1987). *Natural Language Understanding*. Benjamin/Cummings Publishing Company.

[Angeli et al., 2015]
Gabor Angeli, Melvin Jose Johnson Premkumar, and Christopher D. Manning (July 2015). "Leveraging Linguistic Structure for Open Domain Information Extraction". In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. ACL, pp. 344–354. DOI: 10.3115/v1/P15-1034.

[Angeli et al., 2014]
Gabor Angeli, Julie Tibshirani, Jean Wu, and Christopher D. Manning (Oct. 2014). "Combining Distant and Partial Supervision for Relation Extraction". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. ACL, pp. 1556–1567. DOI: 10.3115/v1/D14-1164.

[Appelt, 1999]
Douglas E. Appelt (1999). "Introduction to Information Extraction". In: *AI Communications* 12.3, pp. 161–172.

[Araki et al., 2014]
Jun Araki, Zhengzhong Liu, Eduard H. Hovy, and Teruko Mitamura (2014). "Detecting Subevent Structure for Event Coreference Resolution". In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014*. ELRA, pp. 4553–4558.

[Araki and Mitamura, 2015]
Jun Araki and Teruko Mitamura (Sept. 2015). "Joint Event Trigger Identification and

Event Coreference Resolution with Structured Perceptron". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. ACL, pp. 2074–2080. DOI: 10.18653/v1/D15-1247.

[Augenstein, 2014]
Isabelle Augenstein (Aug. 2014). "Seed Selection for Distantly Supervised Web-Based Relation Extraction". In: *Proceedings of the Third Workshop on Semantic Web and Information Extraction*. ACL, pp. 17–24. DOI: 10.3115/v1/W14-6203.

[Augenstein et al., 2014]
Isabelle Augenstein, Diana Maynard, and Fabio Ciravegna (2014). "Relation Extraction from the Web Using Distant Supervision". In: *Knowledge Engineering and Knowledge Management - 19th International Conference, EKAW 2014*. Vol. 8876. Lecture Notes in Computer Science. Springer, pp. 26–41. DOI: 10.1007/978-3-319-13704-9_3.

[Augenstein et al., 2016]
Isabelle Augenstein, Diana Maynard, and Fabio Ciravegna (2016). "Distantly Supervised Web Relation Extraction for Knowledge Base Population". In: *Semantic Web* 7.4, pp. 335–349. DOI: 10.3233/SW-150180.

[Bach and Badaskar, 2007]
Nguyen Bach and Sameer Badaskar (2007). "A Survey on Relation Extraction". http://www.cs.cmu.edu/~nbach/papers/A-survey-on-Relation-Extraction.pdf, last access: 2017-03-24.

[Baeza-Yates and Ribeiro-Neto, 2011]
Ricardo A. Baeza-Yates and Berthier A. Ribeiro-Neto (2011). *Modern Information Retrieval - The Concepts and Technology Behind Search*. Second edition. Harlow, England: Pearson Education Ltd.

[Bagga and Baldwin, 1998]
Amit Bagga and Breck Baldwin (1998). "Algorithms for Scoring Coreference Chains". In: *Proceedings of the LREC 1998 Workshop on Linguistic Coreference*, pp. 563–566.

[Baker et al., 1998]
Collin F. Baker, Charles J. Fillmore, and John B. Lowe (Aug. 1998). "The Berkeley FrameNet Project". In: *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*. ACL, pp. 86–90. DOI: 10.3115/980845.980860.

[Banko and Brill, 2001a]
Michele Banko and Eric Brill (2001a). "Mitigating the Paucity-Of-Data Problem: Exploring the Effect of Training Corpus Size on Classifier Performance for Natural Language Processing". In: *Proceedings of the First International Conference on Human Language Technology Research*. ACL.

[Banko and Brill, 2001b]
Michele Banko and Eric Brill (July 2001b). "Scaling to Very Very Large Corpora for Natural Language Disambiguation". In: *Proceedings of 39th Annual Meeting of the Association for Computational Linguistics*. ACL, pp. 26–33. DOI: 10.3115/1073012.1073017.

[Banko et al., 2007]
Michele Banko, Michael J. Cafarella, Stephen Soderland, Matthew Broadhead, and Oren Etzioni (2007). "Open Information Extraction from the Web". In: *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pp. 2670–2676.

[Banko and Etzioni, 2008]
Michele Banko and Oren Etzioni (2008). "The Tradeoffs Between Open and Traditional Relation Extraction". In: *ACL 2008, Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*, pp. 28–36.

[Barrena et al., 2014]
Ander Barrena, Eneko Agirre, Bernardo Cabaleiro, Anselmo Peñas, and Aitor Soroa (Aug. 2014). "'One Entity Per Discourse' and 'One Entity Per Collocation' Improve Named-Entity Disambiguation". In: *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. ACL, pp. 2260–2269.

[Beedkar and Gemulla, 2015]
Kaustubh Beedkar and Rainer Gemulla (May 2015). "LASH: Large-Scale Sequence Mining with Hierarchies". In: *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*. ACM, pp. 491–503. DOI: 10.1145/2723372.2723724.

[Bejan and Harabagiu, 2010]
Cosmin Adrian Bejan and Sanda M. Harabagiu (July 2010). "Unsupervised Event Coreference Resolution with Rich Linguistic Features". In: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. ACL, pp. 1412–1422.

[Bejan and Harabagiu, 2014]
Cosmin Adrian Bejan and Sanda M. Harabagiu (2014). "Unsupervised Event Coreference Resolution". In: *Computational Linguistics* 40.2, pp. 311–347. DOI: 10.1162/COLI_a_00174.

[Bengio et al., 2003]
Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin (2003). "A Neural Probabilistic Language Model". In: *Journal of Machine Learning Research* 3, pp. 1137–1155.

[Berant et al., 2014]
Jonathan Berant, Vivek Srikumar, Pei-Chun Chen, Abby Vander Linden, Brittany Harding, Brad Huang, Peter Clark, and Christopher D. Manning (2014). "Modeling Biological Processes for Reading Comprehension". In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP*. ACL, pp. 1499–1510. DOI: 10.3115/v1/D14-1159.

[Berners-Lee, 1999]
Tim Berners-Lee (1999). *Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web by Its Inventor*. Harper Collins.

[Betteridge et al., 2009]
Justin Betteridge, Andrew Carlson, Sue A. Hong, Estevam R. Hruschka Jr., Edith L. M. Law, Tom M. Mitchell, and Sophie H. Wang (2009). "Toward Never Ending Language Learning". In: *Proceedings of the 2009 AAAI Spring Symposium on Learning by Reading and Learning to Read*. AAAI, pp. 1–2.

[Bizer et al., 2009a]
Christian Bizer, Tom Heath, and Tim Berners-Lee (2009a). "Linked Data - The Story So Far". In: *International Journal on Semantic Web and Information Systems (IJSWIS)* 5.3, pp. 1–22. DOI: 10.4018/jswis.2009081901.

[Bizer et al., 2009b]
Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann (2009b). "DBpedia - A Crystallization Point for the Web of Data". In: *Journal of Web Semantics* 7.3, pp. 154–165. DOI: 10.1016/j.websem.2009.07.002.

[Bollacker et al., 2008]
Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor (2008). "Freebase: A Collaboratively Created Graph Database for Structuring Human Knowledge". In: *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*. ACM, pp. 1247–1250. DOI: 10.1145/1376616.1376746.

[Bonial et al., 2013]
Claire Bonial, Kevin Stowe, and Martha Palmer (2013). "Renewing and Revising SemLink". In: *Proceedings of the 2nd Workshop on Linked Data in Linguistics (LDL-2013): Representing and Linking Lexicons, Terminologies and Other Language Data*. ACL, pp. 9–17.

[Bowman et al., 2016]
Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew Dai, Rafal Jozefowicz, and Samy Bengio (Aug. 2016). "Generating Sentences from a Continuous Space". In: *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*. ACL, pp. 10–21. DOI: 10.18653/v1/K16-1002.

[Brants, 2000]
Thorsten Brants (Apr. 2000). "TnT – A Statistical Part-Of-Speech Tagger". In: *Proceedings of the Sixth Conference on Applied Natural Language Processing*. ACL, pp. 224–231. DOI: 10.3115/974147.974178.

[Brants et al., 2007]
Thorsten Brants, Ashok C. Popat, Peng Xu, Franz J. Och, and Jeffrey Dean (June 2007). "Large Language Models in Machine Translation". In: *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*. ACL, pp. 858–867.

[Brennan et al., 1987]
Susan E. Brennan, Marilyn W. Friedman, and Carl J. Pollard (July 1987). "A Centering Approach to Pronouns". In: *Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics*. ACL, pp. 155–162. DOI: 10.3115/981175.981197.

[Brin, 1998]
Sergey Brin (1998). "Extracting Patterns and Relations from the World Wide Web". In: *The World Wide Web and Databases, International Workshop WebDB'98*. Vol. 1590. Lecture Notes in Computer Science. Springer, pp. 172–183. DOI: 10.1007/10704656_11.

[Bronstein et al., 2015]
Ofer Bronstein, Ido Dagan, Qi Li, Heng Ji, and Anette Frank (2015). "Seed-Based Event

Trigger Labeling: How Far Can Event Descriptions Get Us?" In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. ACL, pp. 372–376. DOI: 10.3115/v1/P15-2061.

[Bunescu and Mooney, 2005]
Razvan C. Bunescu and Raymond J. Mooney (Oct. 2005). "A Shortest Path Dependency Kernel for Relation Extraction". In: *Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pp. 724–731.

[Bunescu and Mooney, 2007]
Razvan C. Bunescu and Raymond J. Mooney (2007). "Learning to Extract Relations from the Web Using Minimal Supervision". In: *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*. ACL, pp. 576–583.

[Callan et al., 2009]
J. Callan, M. Hoy, C. Yoo, and L. Zhao (2009). *The ClueWeb09 Dataset*. http://lemurproject.org/clueweb09, last access: 2017-06-20.

[Callison-Burch and Dredze, 2010]
Chris Callison-Burch and Mark Dredze (June 2010). "Creating Speech and Language Data with Amazon's Mechanical Turk". In: *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*. ACL, pp. 1–12.

[Cardie, 1997]
Claire Cardie (1997). "Empirical Methods in Information Extraction". In: *AI Magazine* 18.4, pp. 65–80.

[Carlson et al., 2009]
Andrew Carlson, Justin Betteridge, Estevam R. Hruschka Jr., and Tom M. Mitchell (2009). "Coupling Semi-Supervised Learning of Categories and Relations". In: *Proceedings of the NAACL HLT 2009 Workskop on Semi-Supervised Learning for Natural Language Processing*. ACL, pp. 1–9.

[Carlson et al., 2010a]
Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R. Hruschka Jr., and Tom M. Mitchell (2010a). "Toward an Architecture for Never-Ending Language Learning". In: *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2010*. AAAI Press, pp. 1306–1313.

[Carlson et al., 2010b]
Andrew Carlson, Justin Betteridge, Richard C. Wang, Estevam R. Hruschka Jr., and Tom M. Mitchell (2010b). "Coupled Semi-Supervised Learning for Information Extraction". In: *Proceedings of the Third International Conference on Web Search and Web Data Mining, WSDM 2010*. ACM, pp. 101–110. DOI: 10.1145/1718487.1718501.

[Carpuat, 2009]
Marine Carpuat (June 2009). "One Translation Per Discourse". In: *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions (SEW-2009)*. ACL, pp. 19–27.

[Chang et al., 2006]
C.-H. Chang, M. Kayed, R. Girgis, and K.F. Shaalan (2006). "A Survey of Web Information Extraction Systems". In: *IEEE Transactions on Knowledge and Data Engineering* 18.10, pp. 1411–1428.

[Y. Chen et al., 2015]
Yubo Chen, Liheng Xu, Kang Liu, Daojian Zeng, and Jun Zhao (July 2015). "Event Extraction via Dynamic Multi-Pooling Convolutional Neural Networks". In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. ACL, pp. 167–176. DOI: 10.3115/v1/P15-1017.

[Z. Chen and Ji, 2009]
Zheng Chen and Heng Ji (Aug. 2009). "Graph-Based Event Coreference Resolution". In: *Proceedings of the 2009 Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-4)*. ACL, pp. 54–57.

[Z. Chen et al., 2009]
Zheng Chen, Heng Ji, and Robert Haralick (Sept. 2009). "A Pairwise Event Coreference Model, Feature Impact and Evaluation for Event Coreference Resolution". In: *Proceedings of the Workshop on Events in Emerging Text Types*. ACL, pp. 17–22.

[Chiarcos et al., 2013]
Christian Chiarcos, John McCrae, Philipp Cimiano, and Christiane Fellbaum (2013). "Towards Open Data for Linguistics: Linguistic Linked Data". In: *New Trends of Research in Ontologies and Lexical Resources*. Theory and Applications of Natural Language Processing. Berlin, Heidelberg: Springer, pp. 7–25. DOI: 10.1007/978-3-642-31782-8_2.

[Chieu and H. T. Ng, 2002]
Hai Leong Chieu and Hwee Tou Ng (2002). "A Maximum Entropy Approach to Information Extraction from Semi-Structured and Free Text". In: *Proceedings of the Eighteenth National Conference on Artificial Intelligence and Fourteenth Conference on Innovative Applications of Artificial Intelligence*. AAAI Press / The MIT Press, pp. 786–791.

[Chinchor, 1998a]
Nancy A. Chinchor (1998a). "MUC-7 Test Score Reports for All Participants". In: *Proceedings of the 7th Conference on Message Understanding, MUC 1998*. ACL.

[Chinchor, 1998b]
Nancy A. Chinchor (1998b). "MUC-7 Test Scores Introduction". In: *Proceedings of the 7th Conference on Message Understanding, MUC 1998*. ACL.

[Chinchor, 1998c]
Nancy A. Chinchor (1998c). "Overview of MUC-7". In: *Proceedings of the 7th Conference on Message Understanding, MUC 1998*. ACL.

[Chinchor, 2001]
Nancy A. Chinchor (2001). *Message Understanding Conference (MUC) 7 LDC2001T02*. Web Download, https://catalog.ldc.upenn.edu/LDC2001T02, last access: 2017-06-20. Philadelphia: Linguistic Data Consortium.

[Chinchor and Dungca, 1995]
Nancy A. Chinchor and Gary Dungca (1995). "Four Scorers and Seven Years Ago: The Scoring Method for MUC-6". In: *Proceedings of the 6th Conference on Message Understanding, MUC 1995*. Morgan Kaufmann Publishers / ACL, pp. 33–38.

[Chinchor and Sundheim, 1996]
Nancy A. Chinchor and Beth M. Sundheim (1996). *Message Understanding Conference (MUC) 6 Additional News Text LDC96T10*. Web Download, `https://catalog.ldc.upenn.edu/LDC96T10`, last access: 2017-06-20. Philadelphia: Linguistic Data Consortium.

[Chinchor and Sundheim, 2003]
Nancy A. Chinchor and Beth M. Sundheim (2003). *Message Understanding Conference (MUC) 6 LDC2003T13*. Web Download, `https://catalog.ldc.upenn.edu/LDC2003T13`, last access: 2017-06-20. Philadelphia: Linguistic Data Consortium.

[Chiticariu et al., 2013]
Laura Chiticariu, Yunyao Li, and Frederick R. Reiss (Oct. 2013). "Rule-Based Information Extraction Is Dead! Long Live Rule-Based Information Extraction Systems!" In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. ACL, pp. 827–832.

[Chowdhury and Lavelli, 2012]
Md. Faisal Mahbub Chowdhury and Alberto Lavelli (2012). "Combining Tree Structures, Flat Features and Patterns for Biomedical Relation Extraction". In: *EACL 2012, 13th Conference of the European Chapter of the Association for Computational Linguistics*. ACL, pp. 420–429.

[Chrupała, 2006]
Grzegorz Chrupała (2006). "Simple Data-Driven Context-Sensitive Lemmatization". In: *Proceedings of the Sociedad Española Para El Procesamiento Del Lenguaje Natural (SEPLN)*, pp. 121–127.

[Cicchetti, 1994]
Domenic V. Cicchetti (Dec. 1994). "Guidelines, Criteria, and Rules of Thumb for Evaluating Normed and Standardized Assessment Instruments in Psychology". In: *Psychological Assessment* 6.4, pp. 284–290.

[Cimiano et al., 2013]
Philipp Cimiano, Janna Lüker, David Nagel, and Christina Unger (Aug. 2013). "Exploiting Ontology Lexica for Generating Natural Language Texts from RDF Data". In: *Proceedings of the 14th European Workshop on Natural Language Generation*. ACL, pp. 10–19.

[Cimiano et al., 2014]
Philipp Cimiano, Christina Unger, and John McCrae (2014). *Ontology-Based Interpretation of Natural Language*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers. DOI: `10.2200/S00561ED1V01Y201401HLT024`.

[Collobert et al., 2011]
Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel P. Kuksa (2011). "Natural Language Processing (Almost) from Scratch". In: *Journal of Machine Learning Research* 12, pp. 2493–2537.

[Culotta and Sorensen, 2004]
Aron Culotta and Jeffrey S. Sorensen (2004). "Dependency Tree Kernels for Relation Extraction". In: *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*. ACL, pp. 423–429.

[Cunningham, 2006]
H. Cunningham (2006). "Information Extraction, Automatic". In: *Encyclopedia of Language & Linguistics (Second Edition)*. Ed. by Keith Brown. Second Edition. Elsevier, pp. 665–677. DOI: 10.1016/B0-08-044854-2/00960-3.

[Curran and Osborne, 2002]
James R. Curran and Miles Osborne (2002). "A Very Very Large Corpus Doesn't Always Yield Reliable Estimates". In: *Proceedings of the 6th Conference on Natural Language Learning, CoNLL 2002*. ACL.

[Cybulska and Vossen, 2014]
Agata Cybulska and Piek Vossen (2014). "Using a Sledgehammer to Crack a Nut? Lexical Diversity and Event Coreference Resolution". In: *Proceedings of the International Conference on Language Resources and Evaluation, LREC*. ELRA, pp. 4545–4552.

[Cybulska and Vossen, 2015]
Agata Cybulska and Piek Vossen (June 2015). "Translating Granularity of Event Slots Into Features for Event Coreference Resolution." In: *Proceedings of the the 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation*. ACL, pp. 1–10. DOI: 10.3115/v1/W15-0801.

[de Marneffe et al., 2006]
Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning (2006). "Generating Typed Dependency Parses from Phrase Structure Parses". In: *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*. ACL, pp. 449–454.

[de Marneffe and Manning, 2008]
Marie-Catherine de Marneffe and Christopher D. Manning (Sept. 2008). *Stanford Typed Dependencies Manual*. https://nlp.stanford.edu/software/dependencies_manual.pdf, last access: 2017-06-20.

[Dean et al., 2012]
Jeffrey Dean, Greg Corrado, Rajat Monga, Kai Chen, Matthieu Devin, Quoc V. Le, Mark Z. Mao, Marc'Aurelio Ranzato, Andrew W. Senior, Paul A. Tucker, Ke Yang, and Andrew Y. Ng (2012). "Large Scale Distributed Deep Networks". In: *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012*, pp. 1232–1240.

[Deaton et al., 2005]
Chris Deaton, Blake Shepard, Charles Klein, Corrinne Mayans, Brett Summers, Antoine Brusseau, Michael Witbrock, and Doug Lenat (2005). "The Comprehensive Terrorism Knowledge Base in Cyc". In: *Proceedings of the 2005 International Conference on Intelligence Analysis*.

[Delli Bovi et al., 2015]
Claudio Delli Bovi, Luca Telesca, and Roberto Navigli (2015). "Large-Scale Information Extraction from Textual Definitions Through Deep Syntactic and Semantic Analysis". In: *Transactions of the Association for Computational Linguistics* 3, pp. 529–543.

[Doddington et al., 2004]
George R. Doddington, Alexis Mitchell, Mark A. Przybocki, Lance A. Ramshaw, Stephanie Strassel, and Ralph M. Weischedel (2004). "The Automatic Content Extraction (ACE) Program - Tasks, Data, and Evaluation". In: *Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC 2004*. ELRA, pp. 837–840.

[Dojchinovski et al., 2016]
Milan Dojchinovski, Dimitris Kontokostas, Robert Rößling, Magnus Knuth, and Sebastian Hellmann (Sept. 2016). "DBpedia Links: The Crossroad of Links for the Web of Data". In: *Joint Proceedings of the Posters and Demos Track of the 12th International Conference on Semantic Systems - SEMANTiCS2016 and the 1st International Workshop on Semantic Change & Evolving Semantics (SuCCESS'16)*. Vol. 1695. CEUR Workshop Proceedings. CEUR-WS.org.

[Dong et al., 2014]
Xin Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, and Wei Zhang (2014). "Knowledge Vault: A Web-Scale Approach to Probabilistic Knowledge Fusion". In: *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14*. ACM, pp. 601–610. DOI: 10.1145/2623330.2623623.

[dos Santos et al., 2015]
Cicero dos Santos, Bing Xiang, and Bowen Zhou (July 2015). "Classifying Relations by Ranking with Convolutional Neural Networks". In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. ACL, pp. 626–634. DOI: 10.3115/v1/P15-1061.

[Downey et al., 2005]
Doug Downey, Oren Etzioni, and Stephen Soderland (2005). "A Probabilistic Model of Redundancy in Information Extraction". In: *IJCAI-05, Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence*, pp. 1034–1041.

[Downey et al., 2010]
Doug Downey, Oren Etzioni, and Stephen Soderland (2010). "Analysis of a Probabilistic Model of Redundancy in Unsupervised Information Extraction". In: *Artificial Intelligence* 174.11, pp. 726–748.

[Downey et al., 2007]
Doug Downey, Stefan Schoenmackers, and Oren Etzioni (2007). "Sparse Information Extraction: Unsupervised Language Models to the Rescue". In: *ACL 2007, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pp. 696–703.

[Dridan and Oepen, 2012]
Rebecca Dridan and Stephan Oepen (July 2012). "Tokenization: Returning to a Long

Solved Problem — A Survey, Contrastive Experiment, Recommendations, and Toolkit". In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. ACL, pp. 378–382.

[Drozdzynski et al., 2004]
Witold Drozdzynski, Hans-Ulrich Krieger, Jakub Piskorski, Ulrich Schäfer, and Feiyu Xu (2004). "Shallow Processing with Unification and Typed Feature Structures — Foundations and Applications". In: *Künstliche Intelligenz* 18.1, pp. 17–23.

[Ebrahimi and Dou, 2015]
Javid Ebrahimi and Dejing Dou (May 2015). "Chain Based RNN for Relation Classification". In: *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. ACL, pp. 1244–1249. DOI: 10.3115/v1/N15-1133.

[Eckle-Kohler et al., 2015]
Judith Eckle-Kohler, John Philip McCrae, and Christian Chiarcos (2015). "LemonUby - A Large, Interlinked, Syntactically-Rich Lexical Resource for Ontologies". In: *Semantic Web* 6.4, pp. 371–378. DOI: 10.3233/SW-140159.

[Ehrmann et al., 2014]
Maud Ehrmann, Francesco Cecconi, Daniele Vannella, John Philip McCrae, Philipp Cimiano, and Roberto Navigli (May 2014). "Representing Multilingual Data as Linked Data: The Case of BabelNet 2.0". In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. ELRA, pp. 401–408.

[Ehrmann et al., 2016]
Maud Ehrmann, Damien Nouvel, and Sophie Rosset (2016). "Named Entity Resources - Overview and Outlook". In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016*. ELRA, pp. 23–28.

[Ellis et al., 2014]
Joe Ellis, Jeremy Getman, and Stephanie Strassel (Nov. 2014). "Overview of Linguistic Resources for the TAC KBP 2014 Evaluations: Planning, Execution, and Results". In: *Proceedings of the Text Analysis Conference, TAC*. National Institute of Standards and Technology.

[Erxleben et al., 2014]
Fredo Erxleben, Michael Günther, Markus Krötzsch, Julian Mendez, and Denny Vrandecic (2014). "Introducing Wikidata to the Linked Data Web". In: *The Semantic Web - ISWC 2014 - 13th International Semantic Web Conference*. Vol. 8796. Lecture Notes in Computer Science. Springer, pp. 50–65. DOI: 10.1007/978-3-319-11964-9_4.

[Etzioni et al., 2007]
Oren Etzioni, Michele Banko, and Michael J. Cafarella (2007). "Machine Reading". In: *Proceedings of the 2007 AAAI Spring Symposium on Machine Reading*. AAAI, pp. 1–5.

[Etzioni et al., 2004a]
Oren Etzioni, Michael J. Cafarella, Doug Downey, Stanley Kok, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S. Weld, and Alexander Yates (2004a). "Web-Scale Information Extraction in KnowItAll (Preliminary Results)". In: *Proceedings of the 13th*

International Conference on World Wide Web. ACM, pp. 100–110. DOI: 10.1145/988672.988687.

[Etzioni et al., 2004b]
Oren Etzioni, Michael Cafarella, Doug Downey, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S. Weld, and Alexander Yates (2004b). "Methods for Domain-Independent Information Extraction from the Web: An Experimental Comparison". In: *Proceedings of the 19th National Conference on Artificial Intelligence (AAAI 2004)*. AAAI Press / The MIT Press, pp. 391–398.

[Etzioni et al., 2005]
Oren Etzioni, Michael Cafarella, Doug Downey, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S. Weld, and Alexander Yates (2005). "Unsupervised Named-Entity Extraction from the Web: An Experimental Study". In: *Artificial Intelligence* 165.1, pp. 91–134. DOI: 10.1016/j.artint.2005.03.001.

[Etzioni et al., 2011]
Oren Etzioni, Anthony Fader, Janara Christensen, Stephen Soderland, and Mausam (2011). "Open Information Extraction: The Second Generation". In: *Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI 2011)*. IJCAI/AAAI, pp. 3–10. DOI: 10.5591/978-1-57735-516-8/IJCAI11-012.

[Evang et al., 2013]
Kilian Evang, Valerio Basile, Grzegorz Chrupała, and Johan Bos (Oct. 2013). "Elephant: Sequence Labeling for Word and Sentence Segmentation". In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. ACL, pp. 1422–1426.

[Fader et al., 2011]
Anthony Fader, Stephen Soderland, and Oren Etzioni (2011). "Identifying Relations for Open Information Extraction". In: *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP 2011)*. ACL, pp. 1535–1545.

[Fader et al., 2013]
Anthony Fader, Luke S. Zettlemoyer, and Oren Etzioni (2013). "Paraphrase-Driven Learning for Open Question Answering". In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013*. ACL, pp. 1608–1618.

[Fellbaum, 1998]
Christiane Fellbaum, ed. (May 1998). *WordNet: An Electronic Lexical Database*. Language, Speech, and Communication. Cambrige, MA: MIT Press.

[Ferrucci et al., 2010]
David Ferrucci, Eric Brown, Jennifer Chu-Carroll, James Fan, David Gondek, Aditya A. Kalyanpur, Adam Lally, J. William Murdock, Eric Nyberg, John Prager, Nico Schlaefer, and Chris Welty (2010). "Building Watson: An Overview of the DeepQA Project". In: *AI Magazine* 31.3, pp. 59–79.

[Filippova and Altun, 2013]
Katja Filippova and Yasemin Altun (2013). "Overcoming the Lack of Parallel Data in Sentence Compression". In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013*. ACL, pp. 1481–1491.

[Finkel et al., 2005]
Jenny Rose Finkel, Trond Grenager, and Christopher Manning (2005). "Incorporating Non-Local Information into Information Extraction Systems by Gibbs Sampling". In: *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*. ACL, pp. 363–370. DOI: 10.3115/1219840.1219885.

[Fujiwara and Sekine, 2011]
Shoji Fujiwara and Satoshi Sekine (2011). "Self-Adjusting Bootstrapping". In: *12th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2011)*. Vol. 6609. Lecture Notes in Computer Science. Springer, pp. 188–201. DOI: 10.1007/978-3-642-19437-5_15.

[Gabrilovich et al., 2013]
Evgeniy Gabrilovich, Michael Ringgaard, and Amarnag Subramanya (June 2013). *FACC1: Freebase Annotation of ClueWeb Corpora, Version 1 (Release Date 2013-06-26, Format Version 1, Correction Level 0)*. http://lemurproject.org/clueweb09/FACC1/, last access: 2017-06-20.

[Gabryszak and Krause et al., 2016]
Aleksandra Gabryszak, Sebastian Krause, Leonhard Hennig, Feiyu Xu, and Hans Uszkoreit (2016). "Relation- And Phrase-Level Linking of FrameNet with Sar-Graphs". In: *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC'16)*. ELRA, pp. 2419–2424.

[Gale et al., 1992]
William A. Gale, Kenneth W. Church, and David Yarowsky (Feb. 1992). "One Sense Per Discourse". In: *Proceedings of the Speech and Natural Language Workshop*. HLT. ACL.

[Gandomi and Haider, 2015]
Amir Gandomi and Murtaza Haider (2015). "Beyond the Hype: Big Data Concepts, Methods, and Analytics". In: *International Journal of Information Management* 35.2, pp. 137–144. DOI: 10.1016/j.ijinfomgt.2014.10.007.

[Gardner et al., 2013]
Matt Gardner, Partha Pratim Talukdar, Bryan Kisiel, and Tom M. Mitchell (2013). "Improving Learning and Inference in a Large Knowledge-Base Using Latent Syntactic Cues". In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013*. ACL, pp. 833–838.

[Gardner et al., 2014]
Matt Gardner, Partha Pratim Talukdar, Jayant Krishnamurthy, and Tom M. Mitchell (2014). "Incorporating Vector Space Similarity in Random Walk Inference Over Knowledge Bases". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014*. ACL, pp. 397–406. DOI: 10.3115/v1/D14-1044.

[Gerber and Ngomo, 2011]
Daniel Gerber and Axel-Cyrille Ngonga Ngomo (2011). "Bootstrapping the Linked Data Web". In: *Workshop on Web Scale Knowledge Extraction (WeKEx), International Semantic Web Conference*.

[Gesmundo and Samardzic, 2012]
Andrea Gesmundo and Tanja Samardzic (2012). "Lemmatisation as a Tagging Task". In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. ACL, pp. 368–372.

[Girju et al., 2007]
Roxana Girju, Preslav Nakov, Vivi Nastase, Stan Szpakowicz, Peter Turney, and Deniz Yuret (June 2007). "SemEval-2007 Task 04: Classification of Semantic Relations Between Nominals". In: *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*. ACL, pp. 13–18.

[Goldberg, 2016]
Yoav Goldberg (2016). "A Primer on Neural Network Models for Natural Language Processing". In: *J. Artif. Intell. Res. (JAIR)* 57, pp. 345–420. DOI: 10.1613/jair.4992.

[Goldberg and Orwant, 2013]
Yoav Goldberg and Jon Orwant (2013). "A Dataset of Syntactic-Ngrams Over Time from a Very Large Corpus of English Books". In: *Proceedings of the Second Joint Conference on Lexical and Computational Semantics, *SEM 2013*. ACL, pp. 241–247.

[Goodfellow et al., 2016]
Ian Goodfellow, Yoshua Bengio, and Aaron Courville (2016). *Deep Learning*. Cambridge, MA: MIT Press.

[Gormley et al., 2015]
Matthew R. Gormley, Mo Yu, and Mark Dredze (Sept. 2015). "Improved Relation Extraction with Feature-Rich Compositional Embedding Models". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. ACL, pp. 1774–1784.

[Greenwood and Stevenson, 2006]
Mark A. Greenwood and Mark Stevenson (July 2006). "Improving Semi-Supervised Acquisition of Relation Extraction Patterns". In: *Proceedings of the Workshop on Information Extraction Beyond the Document*. ACL, pp. 29–35.

[Grishman, 1997]
Ralph Grishman (1997). "Information Extraction: Techniques and Challenges". In: *Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology, International Summer School (SCIE-97)*. Vol. 1299. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, pp. 10–27. DOI: 10.1007/3-540-63438-X_2.

[Grishman, 2012]
Ralph Grishman (Jan. 2012). "Information Extraction: Capabilities and Challenges". http://cs.nyu.edu/cs/faculty/grishman/tarragona.pdf, last access: 2017-06-20.

[Grishman and Sundheim, 1995]
Ralph Grishman and Beth M. Sundheim (1995). "Design of the MUC-6 Evaluation". In: *Proceedings of the 6th Conference on Message Understanding, MUC 1995*, pp. 1–11.

[Grishman and Sundheim, 1996]
Ralph Grishman and Beth M. Sundheim (1996). "Message Understanding Conference - 6:

A Brief History". In: *Proceedings of the 16th International Conference on Computational Linguistics (COLING 1996)*. ACL, pp. 466–471.

[Grishman et al., 2005]
Ralph Grishman, David Westbrook, and Adam Meyers (2005). *NYU's English ACE 2005 System Description*. Tech. rep. 05-019. Proteus Project, Department of Computer Science, New York University.

[Grouin et al., 2011]
Cyril Grouin, Sophie Rosset, Pierre Zweigenbaum, Karën Fort, Olivier Galibert, and Ludovic Quintard (June 2011). "Proposal for an Extension of Traditional Named Entities: from Guidelines to Evaluation, an Overview". In: *Proceedings of the 5th Linguistic Annotation Workshop*. ACL, pp. 92–100.

[Grycner and Weikum, 2014]
Adam Grycner and Gerhard Weikum (2014). "HARPY: Hypernyms and Alignment of Relational Paraphrases". In: *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. ACL, pp. 2195–2204.

[Grycner and Weikum, 2016]
Adam Grycner and Gerhard Weikum (Nov. 2016). "POLY: Mining Relational Paraphrases from Multilingual Sentences". In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. ACL, pp. 2183–2192.

[Grycner et al., 2015]
Adam Grycner, Gerhard Weikum, Jay Pujara, James Foulds, and Lise Getoor (Sept. 2015). "RELLY: Inferring Hypernym Relationships Between Relational Phrases". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. ACL, pp. 971–981.

[Gupta et al., 2014]
Rahul Gupta, Alon Y. Halevy, Xuezhi Wang, Steven Euijong Whang, and Fei Wu (2014). "Biperpedia: An Ontology for Search Applications". In: *Proceedings of the VLDB Endowment* 7.7, pp. 505–516.

[Gurevych et al., 2012]
Iryna Gurevych, Judith Eckle-Kohler, Silvana Hartmann, Michael Matuschek, Christian M. Meyer, and Christian Wirth (Apr. 2012). "Uby: A Large-Scale Unified Lexical-Semantic Resource Based on LMF". In: *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. ACL, pp. 580–590.

[Hallgren, 2012]
K. A. Hallgren (2012). "Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial". In: *Tutorials in Quantitative Methods for Psychology* 8.1, pp. 23–34.

[Hashem et al., 2015]
Ibrahim Abaker Targio Hashem, Ibrar Yaqoob, Nor Badrul Anuar, Salimah Mokhtar, Abdullah Gani, and Samee Ullah Khan (2015). "The Rise of 'Big Data' on Cloud Computing: Review and Open Research Issues". In: *Information Systems* 47, pp. 98–115. DOI: 10.1016/j.is.2014.07.006.

[Hashimoto et al., 2013]
Kazuma Hashimoto, Makoto Miwa, Yoshimasa Tsuruoka, and Takashi Chikayama (Oct. 2013). "Simple Customization of Recursive Neural Networks for Semantic Relation Classification". In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. ACL, pp. 1372–1376.

[Hashimoto et al., 2015]
Kazuma Hashimoto, Pontus Stenetorp, Makoto Miwa, and Yoshimasa Tsuruoka (July 2015). "Task-Oriented Learning of Word Embeddings for Semantic Relation Classification". In: *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*. ACL, pp. 268–278.

[Hasler et al., 2006]
Laura Hasler, Constantin Orasan, and Karin Naumann (2006). "NPs for Events: Experiments in Coreference Annotation". In: *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*. ELRA, pp. 1167–1172.

[Hearst, 1992]
Marti A. Hearst (1992). "Automatic Acquisition of Hyponyms from Large Text Corpora". In: *COLING 1992 Volume 2: The 15th International Conference on Computational Linguistics*. ACL, pp. 539–545.

[Hendrickx et al., 2010]
Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz (July 2010). "SemEval-2010 Task 8: Multi-Way Classification of Semantic Relations Between Pairs of Nominals". In: *Proceedings of the 5th International Workshop on Semantic Evaluation*. ACL, pp. 33–38.

[Hennig, H. Li, and Krause, et al., 2015]
Leonhard Hennig, Hong Li, Sebastian Krause, Feiyu Xu, and Hans Uszkoreit (2015). "A Web-Based Collaborative Evaluation Tool for Automatically Learned Relation Extraction Patterns". In: *Proceedings of ACL-IJCNLP 2015 System Demonstrations*. ACL/AFNLP, pp. 43–48.

[Hirschberg and Manning, 2015]
Julia Hirschberg and Christopher D. Manning (2015). "Advances in Natural Language Processing". In: *Science* 349.6245, pp. 261–266. DOI: 10.1126/science.aaa8685.

[Hobbs, 1978]
Jerry R. Hobbs (1978). "Resolving Pronoun References". In: *Lingua* 44.4, pp. 311–338. DOI: 10.1016/0024-3841(78)90006-2.

[Hoffart et al., 2011]
Johannes Hoffart, Fabian M. Suchanek, Klaus Berberich, Edwin Lewis Kelham, Gerard de Melo, and Gerhard Weikum (2011). "YAGO2: Exploring and Querying World Knowledge in Time, Space, Context, and Many Languages". In: *Proceedings of the 20th International Conference on World Wide Web, WWW 2011, Companion Volume*. ACM, pp. 229–232. DOI: 10.1145/1963192.1963296.

[Hoffart et al., 2010]
Johannes Hoffart, Fabian M. Suchanek, Klaus Berberich, and Gerhard Weikum (2010). *YAGO2: A Spatially and Temporally Enhanced Knowledge Base from Wikipedia*. Tech. rep. MPI-I-2010-5-007. Saarbrücken: Max-Planck-Institut für Informatik.

[Hoffart et al., 2013]
Johannes Hoffart, Fabian M. Suchanek, Klaus Berberich, and Gerhard Weikum (2013). "YAGO2: A Spatially and Temporally Enhanced Knowledge Base from Wikipedia". In: *Artificial Intelligence* 194, pp. 28–61. DOI: 10.1016/j.artint.2012.06.001.

[Hoffmann et al., 2011]
Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke Zettlemoyer, and Daniel S. Weld (2011). "Knowledge-Based Weak Supervision for Information Extraction of Overlapping Relations". In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. ACL, pp. 541–550.

[Hoffmann et al., 2010]
Raphael Hoffmann, Congle Zhang, and Daniel S. Weld (July 2010). "Learning 5000 Relational Extractors". In: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. ACL, pp. 286–295.

[Hovy et al., 2009]
Eduard H. Hovy, Zornitsa Kozareva, and Ellen Riloff (2009). "Toward Completeness in Concept Extraction and Classification". In: *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, EMNLP 2009*. ACL, pp. 948–957.

[Hovy et al., 2013a]
Eduard H. Hovy, Teruko Mitamura, Felisa Verdejo, Jun Araki, and Andrew Philpot (June 2013a). "Events Are Not Simple: Identity, Non-Identity, and Quasi-Identity". In: *Workshop on Events: Definition, Detection, Coreference, and Representation*. ACL, pp. 21–28.

[Hovy et al., 2013b]
Eduard H. Hovy, Roberto Navigli, and Simone Paolo Ponzetto (2013b). "Collaboratively Built Semi-Structured Content and Artificial Intelligence: The Story So Far". In: *Artificial Intelligence* 194, pp. 2–27. DOI: 10.1016/j.artint.2012.10.002.

[Hu et al., 2014]
Han Hu, Yonggang Wen, Tat-Seng Chua, and Xuelong Li (2014). "Toward Scalable Systems for Big Data Analytics: A Technology Tutorial". In: *IEEE Access* 2, pp. 652–687. DOI: 10.1109/ACCESS.2014.2332453.

[Huang and Riloff, 2012]
Ruihong Huang and Ellen Riloff (2012). "Modeling Textual Cohesion for Event Extraction". In: *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*. AAAI Press, pp. 1664–1670.

[Illig et al., 2014]
Jens Illig, Benjamin Roth, and Dietrich Klakow (Apr. 2014). "Unsupervised Parsing for Generating Surface-Based Relation Extraction Patterns". In: *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, Volume 2: Short Papers*. ACL, pp. 100–105.

[Jain and Pantel, 2010]
Alpa Jain and Patrick Pantel (Aug. 2010). "FactRank: Random Walks on a Web of Facts". In: *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*. Coling 2010 Organizing Committee, pp. 501–509.

[Jean-Louis et al., 2013]
Ludovic Jean-Louis, Romaric Besançon, Olivier Ferret, and Adrien Durand (2013). "Using Distant Supervision for Extracting Relations on a Large Scale". In: *Knowledge Discovery, Knowledge Engineering and Knowledge Management: Third International Joint Conference, IC3K 2011. Revised Selected Papers*. Communications in Computer and Information Science. Berlin, Heidelberg: Springer, pp. 141–155. DOI: 10.1007/978-3-642-37186-8_9.

[Ji, 2010]
Heng Ji (Aug. 2010). "Challenges from Information Extraction to Information Fusion". In: *Coling 2010: Posters*. Coling 2010 Organizing Committee, pp. 507–515.

[Ji et al., 2010]
Heng Ji, Zheng Chen, Jonathan Feldman, Antonio Gonzalez, Ralph Grishman, and Vivek Upadhyay (June 2010). "Utility Evaluation of Cross-Document Information Extraction". In: *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. ACL, pp. 285–288.

[Ji and Grishman, 2008]
Heng Ji and Ralph Grishman (June 2008). "Refining Event Extraction Through Cross-Document Inference". In: *ACL 2008, Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*. ACL, pp. 254–262.

[Ji and Grishman, 2011]
Heng Ji and Ralph Grishman (June 2011). "Knowledge Base Population: Successful Approaches and Challenges". In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. ACL, pp. 1148–1158.

[Ji et al., 2009]
Heng Ji, Ralph Grishman, Zheng Chen, and Prashant Gupta (Sept. 2009). "Cross-Document Event Extraction and Tracking: Task, Evaluation, Techniques and Challenges". In: *Proceedings of the International Conference RANLP-2009*. ACL, pp. 166–172.

[Ji et al., 2014]
Heng Ji, Joel Nothman, and Ben Hachey (2014). "Overview of TAC-KBP2014 Entity Discovery and Linking Tasks". In: *Proceedings of the Text Analysis Conference, TAC*.

[Ji et al., 2015]
Heng Ji, Joel Nothman, Ben Hachey, and Radu Florian (2015). "Overview of TAC-KBP2015 Tri-Lingual Entity Discovery and Linking". In: *Proceedings of the Text Analysis Conference, TAC*.

[Jia et al., 2014]
Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross B. Girshick, Sergio Guadarrama, and Trevor Darrell (2014). "Caffe: Convolutional Architecture for Fast Feature Embedding". In: *CoRR* abs/1408.5093. arXiv: 1408.5093 [cs.CV].

[Jiang and Zhai, 2007]
Jing Jiang and ChengXiang Zhai (2007). "A Systematic Exploration of the Feature Space for Relation Extraction". In: *Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*. ACL, pp. 113–120.

[Jurafsky and Martin, 2009]
Daniel Jurafsky and James H. Martin (2009). *Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*. Second Edition. Pearson International Edition. Upper Saddle River, New Jersey: Pearson Education.

[Kaiser et al., 2017]
Lukasz Kaiser, Aidan N. Gomez, Noam Shazeer, Ashish Vaswani, Niki Parmar, Llion Jones, and Jakob Uszkoreit (June 2017). "One Model to Learn Them All". In: *CoRR* abs/1706.05137. arXiv: 1706.05137 [cs.LG].

[Kalchbrenner et al., 2014]
Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom (2014). "A Convolutional Neural Network for Modelling Sentences". In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014*. ACL, pp. 655–665.

[Kambhatla, 2004]
Nanda Kambhatla (2004). "Combining Lexical, Syntactic, and Semantic Features with Maximum Entropy Models for Information Extraction". In: *The Companion Volume to the Proceedings of 42st Annual Meeting of the Association for Computational Linguistics*. ACL, pp. 178–181.

[Keller et al., 2002]
Frank Keller, Maria Lapata, and Olga Ourioupina (July 2002). "Using the Web to Overcome Data Sparseness". In: *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*. ACL, pp. 230–237. DOI: 10.3115/1118693.1118723.

[Kim et al., 2011]
Seokhwan Kim, Minwoo Jeong, and Gary Geunbae Lee (July 2011). "A Local Tree Alignment Approach to Relation Extraction of Multiple Arguments". In: *Information Processing & Management* 47.4, pp. 593–605. DOI: 10.1016/j.ipm.2010.12.002.

[Kingma and Ba, 2014]
Diederik P. Kingma and Jimmy Ba (2014). "Adam: A Method for Stochastic Optimization". In: *CoRR* abs/1412.6980. arXiv: 1412.6980 [cs.LG].

[Kiros et al., 2015]
Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler (2015). "Skip-Thought Vectors". In: *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015*, pp. 3294–3302.

[Kirschnick et al., 2014]
Johannes Kirschnick, Alan Akbik, and Holmer Hemsen (May 2014). "Freepal: A Large Collection of Deep Lexico-Syntactic Patterns for Relation Extraction". In: *Proceedings of the*

Ninth International Conference on Language Resources and Evaluation (LREC'14). ELRA, pp. 2071–2075.

[Komen, 2009]
Erwin R. Komen (June 2009). *Coreference Annotation Guidelines*. Tech. rep. available at `http://erwinkomen.ruhosting.nl/doc/2009_CorefCodingManual_V2-0.pdf`, last access: 2017-05-11. Nijmegen, Netherlands: Radboud University.

[Kozareva and Hovy, 2010a]
Zornitsa Kozareva and Eduard H. Hovy (2010a). "A Semi-Supervised Method to Learn and Construct Taxonomies Using the Web". In: *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, EMNLP 2010*. ACL, pp. 1110–1118.

[Kozareva and Hovy, 2010b]
Zornitsa Kozareva and Eduard H. Hovy (2010b). "Learning Arguments and Supertypes of Semantic Relations Using Recursive Patterns". In: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. ACL, pp. 1482–1491.

[Kozareva and Hovy, 2011]
Zornitsa Kozareva and Eduard H. Hovy (Sept. 2011). "Learning Temporal Information for States and Events". In: *Proceedings of the 5th IEEE International Conference on Semantic Computing (ICSC 2011)*. IEEE Computer Society, pp. 424–429. DOI: `10.1109/ICSC.2011.94`.

[Kozareva et al., 2008]
Zornitsa Kozareva, Ellen Riloff, and Eduard H. Hovy (2008). "Semantic Class Learning from the Web with Hyponym Pattern Linkage Graphs". In: *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*. ACL, pp. 1048–1056.

[Krause, 2012]
Sebastian Krause (2012). "Relation Extraction with Massive Seed and Large Corpora". Diplom thesis, Department of Computer Science, Humboldt-Universität zu Berlin.

[Krause et al., 2015a]
Sebastian Krause, Enrique Alfonseca, Katja Filippova, and Daniele Pighin (2015a). "Idest: Learning a Distributed Representation for Event Patterns". In: *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. ACL, pp. 1140–1149.

[Krause et al., 2015b]
Sebastian Krause, Leonhard Hennig, Aleksandra Gabryszak, Feiyu Xu, and Hans Uszkoreit (2015b). "Sar-Graphs: A Linked Linguistic Knowledge Resource Connecting Facts with Language". In: *Proceedings of the 4th Workshop on Linked Data in Linguistics: Resources and Applications*. ACL, pp. 30–38.

[Krause et al., 2016a]
Sebastian Krause, Leonhard Hennig, Andrea Moro, Dirk Weissenborn, Feiyu Xu, Hans Uszkoreit, and Roberto Navigli (2016a). "Sar-Graphs: A Language Resource Connecting Linguistic Knowledge with Semantic Relations from Knowledge Graphs". In: *Web Semantics: Science, Services and Agents on the World Wide Web* 37-38, pp. 112–131. DOI: `10.1016/j.websem.2016.03.004`.

[Krause et al., 2017]
Sebastian Krause, Mikhail Kozhevnikov, Daniele Pighin, and Eric Malmi (2017). "Redundancy Localization for the Conversationalization of Unstructured Responses". In: *Proceedings of the SIGDIAL 2017 Conference, The 18th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. To appear. ACL.

[Krause et al., 2012]
Sebastian Krause, Hong Li, Hans Uszkoreit, and Feiyu Xu (Nov. 2012). "Large-Scale Learning of Relation-Extraction Rules with Distant Supervision from the Web". In: *The Semantic Web – ISWC 2012 – 11th International Semantic Web Conference*. Vol. 7649. Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 263–278. DOI: 10.1007/978-3-642-35176-1_17.

[Krause et al., 2014]
Sebastian Krause, Hong Li, Feiyu Xu, Hans Uszkoreit, Robert Hummel, and Luise Spielhagen (2014). "Language Resources and Annotation Tools for Cross-Sentence Relation Extraction". In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. ELRA, pp. 4320–4325.

[Krause et al., 2016b]
Sebastian Krause, Feiyu Xu, Hans Uszkoreit, and Dirk Weissenborn (Aug. 2016b). "Event Linking with Sentential Features from Convolutional Neural Networks". In: *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning (CoNLL '16)*. ACL, pp. 239–249.

[Krizhevsky et al., 2012]
Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton (2012). "ImageNet Classification with Deep Convolutional Neural Networks". In: *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012*, pp. 1106–1114.

[Kübler et al., 2009]
Sandra Kübler, Ryan McDonald, and Joakim Nivre (2009). *Dependency Parsing*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool.

[Labrinidis and Jagadish, 2012]
Alexandros Labrinidis and H. V. Jagadish (Aug. 2012). "Challenges and Opportunities with Big Data". In: *Proceedings of the VLDB Endowment* 5.12, pp. 2032–2033. DOI: 10.14778/2367502.2367572.

[Lao et al., 2011]
Ni Lao, Tom Mitchell, and William W. Cohen (July 2011). "Random Walk Inference and Learning in a Large Scale Knowledge Base". In: *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. ACL, pp. 529–539.

[Lao et al., 2012]
Ni Lao, Amarnag Subramanya, Fernando Pereira, and William W. Cohen (July 2012). "Reading the Web with Learned Syntactic-Semantic Inference Rules". In: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. ACL, pp. 1017–1026.

[Lapata and Keller, 2004]
Mirella Lapata and Frank Keller (May 2004). "The Web as a Baseline: Evaluating the Performance of Unsupervised Web-Based Models for a Range of NLP Tasks". In: *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, HLT-NAACL*. ACL, pp. 121–128.

[Le and Mikolov, 2014]
Quoc V. Le and Tomas Mikolov (2014). "Distributed Representations of Sentences and Documents". In: *Proceedings of the 31th International Conference on Machine Learning, ICML 2014*. Vol. 32. JMLR.org, pp. 1188–1196.

[Le et al., 2012]
Quoc V. Le, Marc'Aurelio Ranzato, Rajat Monga, Matthieu Devin, Greg Corrado, Kai Chen, Jeffrey Dean, and Andrew Y. Ng (2012). "Building High-Level Features Using Large Scale Unsupervised Learning". In: *Proceedings of the 29th International Conference on Machine Learning, ICML 2012*. icml.cc / Omnipress.

[Lebret and Collobert, 2014]
Rémi Lebret and Ronan Collobert (2014). "Word Embeddings Through Hellinger PCA". In: *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2014*. ACL, pp. 482–490.

[Lebret et al., 2013]
Rémi Lebret, Joël Legrand, and Ronan Collobert (Dec. 2013). *Is Deep Learning Really Necessary for Word Embeddings?* Tech. rep. Idiap-RR-44-2013. Idiap.

[Lebrun and Team Wit, 2016]
Alex Lebrun and Team Wit (Apr. 2016). *Bot Engine*. wit.ai blog, `https://wit.ai/blog/2016/04/12/bot-engine`, last access: 2016-05-02.

[LeCun et al., 2015]
Yann LeCun, Yoshua Bengio, and Geoffrey Hinton (May 2015). "Deep Learning". In: *Nature* 521.7553, pp. 436–444. DOI: `10.1038/nature14539`.

[LeCun et al., 1998]
Yann LeCun, Léon Bottou, Yoshua Bengio, and Pattrick Haffner (Nov. 1998). "Gradient-Based Learning Applied to Document Recognition". In: *Proceedings of the IEEE* 86.11, pp. 2278–2324. DOI: `10.1109/5.726791`.

[A. Lee et al., 2010]
Adam Lee, Marissa Passantino, Heng Ji, Guojun Qi, and Thomas Huang (Aug. 2010). "Enhancing Multi-Lingual Information Extraction via Cross-Media Inference and Fusion". In: *Coling 2010: Posters*. Coling 2010 Organizing Committee, pp. 630–638.

[H. Lee et al., 2013]
Heeyoung Lee, Angel Chang, Yves Peirsman, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky (2013). "Deterministic Coreference Resolution Based on Entity-Centric, Precision-Ranked Rules". In: *Computational Linguistics* 39.4, pp. 885–916. DOI: `10.1162/COLI_a_00152`.

[H. Lee et al., 2011]
Heeyoung Lee, Yves Peirsman, Angel Chang, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky (June 2011). "Stanford's Multi-Pass Sieve Coreference Resolution System at the CoNLL-2011 Shared Task". In: *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*. ACL, pp. 28–34.

[H. Lee et al., 2012]
Heeyoung Lee, Marta Recasens, Angel Chang, Mihai Surdeanu, and Dan Jurafsky (July 2012). "Joint Entity and Event Coreference Resolution across Documents". In: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. ACL, pp. 489–500.

[Lehmann et al., 2015]
Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Christian Bizer (2015). "DBpedia - A Large-Scale, Multilingual Knowledge Base Extracted from Wikipedia". In: *Semantic Web Journal* 6.2, pp. 167–195. DOI: 10.3233/SW-140134.

[Lenat, 1995]
Douglas B. Lenat (1995). "CYC: A Large-Scale Investment in Knowledge Infrastructure". In: *Communications of the ACM* 38.11, pp. 33–38. DOI: 10.1145/219717.219745.

[Lenat and Durlach, 2014]
Douglas B. Lenat and Paula J. Durlach (2014). "Reinforcing Math Knowledge by Immersing Students in a Simulated Learning-By-Teaching Experience". In: *International Journal of Artificial Intelligence in Education* 24.3, pp. 216–250. DOI: 10.1007/s40593-014-0016-x.

[Levy and Goldberg, 2014a]
Omer Levy and Yoav Goldberg (June 2014a). "Dependency-Based Word Embeddings". In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. ACL, pp. 302–308.

[Levy and Goldberg, 2014b]
Omer Levy and Yoav Goldberg (2014b). "Linguistic Regularities in Sparse and Explicit Word Representations". In: *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*. ACL, pp. 171–180.

[Levy and Goldberg, 2014c]
Omer Levy and Yoav Goldberg (2014c). "Neural Word Embedding as Implicit Matrix Factorization". In: *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014*, pp. 2177–2185.

[Levy et al., 2015]
Omer Levy, Steffen Remus, Chris Biemann, and Ido Dagan (June 2015). "Do Supervised Distributional Methods Really Learn Lexical Inference Relations?" In: *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. ACL, pp. 970–976.

[C. Li et al., 2013]
Chen Li, Maria Liakata, and Dietrich Rebholz-Schuhmann (2013). "Biological Network

Extraction from Scientific Literature: State of the Art and Challenges". In: *Briefings in Bioinformatics* 15.5, pp. 856–877. DOI: 10.1093/bib/bbt006.

[H. Li et al., 2012]
Hong Li, Xiwen Cheng, Kristina Adson, Tal Kirshboim, and Feiyu Xu (May 2012). "Annotating Opinions in German Political News". In: *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*. ELRA, pp. 1183–1188.

[H. Li and Krause et al., 2015]
Hong Li, Sebastian Krause, Feiyu Xu, Andrea Moro, Hans Uszkoreit, and Roberto Navigli (2015). "Improvement of n-ary Relation Extraction by Adding Lexical Semantics to Distant-Supervision Rule Learning". In: *Proceedings of the 7th International Conference on Agents and Artificial Intelligence (ICAART-15)*. SciTePress, pp. 317–324.

[H. Li and Krause et al., 2014]
Hong Li, Sebastian Krause, Feiyu Xu, Hans Uszkoreit, Robert Hummel, and Veselina Mironova (2014). "Annotating Relation Mentions in Tabloid Press". In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. ELRA, pp. 3253–3257.

[J. Li et al., 2015]
Jiwei Li, Thang Luong, Dan Jurafsky, and Eduard Hovy (Sept. 2015). "When Are Tree Structures Necessary for Deep Learning of Representations?" In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. ACL, pp. 2304–2314.

[Q. Li et al., 2011]
Qi Li, Sam Anzaroot, Wen-Pin Lin, Xiang Li, and Heng Ji (2011). "Joint Inference for Cross-Document Information Extraction". In: *Proceedings of the 20th ACM Conference on Information and Knowledge Management (CIKM 2011)*, pp. 2225–2228. DOI: 10.1145/2063576.2063932.

[X. Li et al., 2015]
Xiang Li, Thien Huu Nguyen, Kai Cao, and Ralph Grishman (July 2015). "Improving Event Detection with Abstract Meaning Representation". In: *Proceedings of the First Workshop on Computing News Storylines*. ACL, pp. 11–15.

[Liao and Grishman, 2010a]
Shasha Liao and Ralph Grishman (Aug. 2010a). "Filtered Ranking for Bootstrapping in Event Extraction". In: *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*. Coling 2010 Organizing Committee, pp. 680–688.

[Liao and Grishman, 2010b]
Shasha Liao and Ralph Grishman (July 2010b). "Using Document Level Cross-Event Inference to Improve Event Extraction". In: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. ACL, pp. 789–797.

[Liao and Grishman, 2011]
Shasha Liao and Ralph Grishman (June 2011). "Can Document Selection Help Semi-Supervised Learning? A Case Study on Event Extraction". In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. ACL, pp. 260–265.

[Lin and Pantel, 2001]
Dekang Lin and Patrick Pantel (2001). "Discovery of Inference Rules for Question-Answering". In: *Natural Language Engineering* 7.4, pp. 343–360.

[Linguistic Data Consortium, 2004a]
Linguistic Data Consortium (2004a). *ACE (Automatic Content Extraction) Annotation Guidelines for Entity Detection and Tracking (EDT)*. Version 4.2.6, 200400401. https://catalog.ldc.upenn.edu/docs/LDC2005T09/guidelines/EnglishEDTV4-2-6.PDF, last access: 2017-05-15.

[Linguistic Data Consortium, 2004b]
Linguistic Data Consortium (2004b). *ACE (Automatic Content Extraction) Annotation Guidelines for Relation Detection and Characterization (RDC)*. Version 4.3.2, 20040401. https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-rdc-v4.3.2.PDF, last access: 2017-05-15.

[Linguistic Data Consortium, 2005a]
Linguistic Data Consortium (2005a). *ACE (Automatic Content Extraction) English Annotation Guidelines for Entities*. Version 5.6.6, 2005.08.01. https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-entities-guidelines-v5.6.6.pdf, last access: 2017-02-16.

[Linguistic Data Consortium, 2005b]
Linguistic Data Consortium (2005b). *ACE (Automatic Content Extraction) English Annotation Guidelines for Events*. Version 5.4.3, 2005.07.01. https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-events-guidelines-v5.4.3.pdf, last access: 2017-03-08.

[Linguistic Data Consortium, 2005c]
Linguistic Data Consortium (2005c). *ACE (Automatic Content Extraction) English Annotation Guidelines for Relations*. Version 5.8.3, 2005.07.01. https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-relations-guidelines-v5.8.3.pdf, last access: 2017-02-16.

[Y. Liu et al., 2015]
Yang Liu, Furu Wei, Sujian Li, Heng Ji, Ming Zhou, and Houfeng WANG (July 2015). "A Dependency-Based Neural Network for Relation Classification". In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. ACL, pp. 285–290.

[Z. Liu et al., 2014]
Zhengzhong Liu, Jun Araki, Eduard H. Hovy, and Teruko Mitamura (May 2014). "Supervised Within-Document Event Coreference Using Information Propagation". In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. ELRA, pp. 4539–4544.

[J. Lu and V. Ng, 2015]
Jing Lu and Vincent Ng (Nov. 2015). "UTD's Event Nugget Detection and Coreference

System at KBP 2015". In: *Proceedings of the Eighth Text Analysis Conference (TAC 2015)*. National Institute of Standards and Technology.

[J. Lu et al., 2016]
Jing Lu, Deepak Venugopal, Vibhav Gogate, and Vincent Ng (Dec. 2016). "Joint Inference for Event Coreference Resolution". In: *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. The COLING 2016 Organizing Committee, pp. 3264–3275.

[W. Lu and Roth, 2012]
Wei Lu and Dan Roth (July 2012). "Automatic Event Extraction with Structured Preference Modeling". In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. ACL, pp. 835–844.

[Mahdisoltani et al., 2015]
Farzaneh Mahdisoltani, Joanna Biega, and Fabian M. Suchanek (2015). "YAGO3: A Knowledge Base from Multilingual Wikipedias". In: *CIDR 2015, Seventh Biennial Conference on Innovative Data Systems Research*. www.cidrdb.org.

[Manning, 2015]
Christopher D. Manning (2015). "Computational Linguistics and Deep Learning". In: *Computational Linguistics* 41.4, pp. 701–707. DOI: 10.1162/COLI_a_00239.

[Manning and Schütze, 1999]
Christopher D. Manning and Hinrich Schütze (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, MA: The MIT Press.

[Manning et al., 2014]
Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky (June 2014). "The Stanford CoreNLP Natural Language Processing Toolkit". In: *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. ACL, pp. 55–60.

[D. Marcus, 2015]
David Marcus (Aug. 2015). Blog entry on Facebook, http://newsroom.fb.com/news/2016/04/messenger-platform-at-f8/, last access: 2016-05-02.

[D. Marcus, 2016]
David Marcus (Apr. 2016). *Messenger Platform at F8*. Blog entry on "Facebook Newsroom", http://newsroom.fb.com/news/2016/04/messenger-platform-at-f8/, last access: 2016-05-02.

[M. P. Marcus et al., 1993]
Mitchell P Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini (1993). "Building a Large Annotated Corpus of English: The Penn Treebank". In: *Computational linguistics* 19.2, pp. 313–330.

[Matuszek et al., 2006]
Cynthia Matuszek, John Cabral, Michael J. Witbrock, and John DeOliveira (2006). "An Introduction to the Syntax and Content of Cyc". In: *Formalizing and Compiling Background*

Knowledge and Its Applications to Knowledge Representation and Question Answering, Papers from the 2006 AAAI Spring Symposium, Technical Report SS-06-05. AAAI, pp. 44–49.

[Mausam et al., 2012]
Mausam, Michael Schmitz, Stephen Soderland, Robert Bart, and Oren Etzioni (2012). "Open Language Learning for Information Extraction". In: *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, EMNLP-CoNLL*. ACL, pp. 523–534.

[McCrae et al., 2011]
John P. McCrae, Dennis Spohr, and Philipp Cimiano (2011). "Linking Lexical Resources and Ontologies on the Semantic Web with Lemon". In: *The Semantic Web: Research and Applications - 8th Extended Semantic Web Conference, ESWC 2011*. Vol. 6643. Lecture Notes in Computer Science. Springer, pp. 245–259. DOI: 10.1007/978-3-642-21034-1_17.

[Melli et al., 2007]
Gabor Melli, Martin Ester, and Anoop Sarkar (2007). "Recognition of Multi-Sentence N-Ary Subcellular Localization Mentions in Biomedical Abstracts". In: *Short Paper Proceedings of the 2nd International Symposium on Languages in Biology and Medicine (LBM 2007)*. Vol. 319. CEUR Workshop Proceedings. CEUR-WS.org.

[Melo and Weikum, 2009]
Gerard de Melo and Gerhard Weikum (2009). "Towards a Universal Wordnet by Learning from Combined Evidence". In: *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM)*. ACM, pp. 513–522. DOI: 10.1145/1645953.1646020.

[Mikolov et al., 2013a]
Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean (2013a). "Efficient Estimation of Word Representations in Vector Space". In: *CoRR* abs/1301.3781. arXiv: 1301.3781 [cs.CL].

[Mikolov et al., 2015]
Tomas Mikolov, Armand Joulin, and Marco Baroni (2015). "A Roadmap towards Machine Intelligence". In: *CoRR* abs/1511.08130. arXiv: 1511.08130 [cs.AI].

[Mikolov et al., 2013b]
Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean (2013b). "Distributed Representations of Words and Phrases and Their Compositionality". In: *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013*, pp. 3111–3119.

[Mikolov et al., 2013c]
Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig (June 2013c). "Linguistic Regularities in Continuous Space Word Representations". In: *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. ACL, pp. 746–751.

[G. A. Miller, 1995]
George A. Miller (Nov. 1995). "WordNet: A Lexical Database for English". In: *Communications of the ACM* 38 (11), pp. 39–41. DOI: 10.1145/219717.219748.

[S. Miller et al., 2000]
Scott Miller, Heidi Fox, Lance A. Ramshaw, and Ralph M. Weischedel (2000). "A Novel Use of Statistical Parsing to Extract Information from Text". In: *Sixth Applied Natural Language Processing Conference, ANLP*, pp. 226–233.

[Min and Grishman, 2012]
Bonan Min and Ralph Grishman (May 2012). "Challenges in the Knowledge Base Population Slot Filling Task". In: *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. ELRA, pp. 1137–1142.

[Min et al., 2013]
Bonan Min, Ralph Grishman, Li Wan, Chang Wang, and David Gondek (2013). "Distant Supervision for Relation Extraction with an Incomplete Knowledge Base". In: *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics*. ACL, pp. 777–782.

[Mintz et al., 2009]
Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky (2009). "Distant Supervision for Relation Extraction without Labeled Data". In: *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*. Vol. 2. ACL, pp. 1003–1011.

[A. Mitchell et al., 2005]
Alexis Mitchell, Stephanie Strassel, Shudong Huang, and Ramez Zakhary (Mar. 2005). *ACE 2004 Multilingual Training Corpus LDC2005T09*. Web Download, `https://catalog.ldc.upenn.edu/LDC2005T09`, last access: 2017-06-20. Philadelphia: Linguistic Data Consortium.

[A. Mitchell et al., 2004]
Alexis Mitchell, Stephanie Strassel, Mark Przybocki, JK Davis, George Doddington, Ralph Grishman, Adam Meyers, Ada Brunstein, Lisa Ferro, and Beth Sundheim (Feb. 2004). *TIDES Extraction (ACE) 2003 Multilingual Training Data LDC2004T09*. Web Download, `https://catalog.ldc.upenn.edu/LDC2004T09`, last access: 2017-06-20. Philadelphia: Linguistic Data Consortium.

[T. M. Mitchell et al., 2015]
Tom M. Mitchell, William W. Cohen, Estevam R. Hruschka Jr., Partha Pratim Talukdar, Justin Betteridge, Andrew Carlson, Bhavana Dalvi Mishra, Matthew Gardner, Bryan Kisiel, Jayant Krishnamurthy, Ni Lao, Kathryn Mazaitis, Thahir Mohamed, Ndapandula Nakashole, Emmanouil Antonios Platanios, Alan Ritter, Mehdi Samadi, Burr Settles, Richard C. Wang, Derry Tanti Wijaya, Abhinav Gupta, Xinlei Chen, Abulhair Saparov, Malcolm Greaves, and Joel Welling (2015). "Never-Ending Learning". In: *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*. AAAI Press, pp. 2302–2310.

[Mitkov et al., 2000]
Ruslan Mitkov, Richard Evans, Constantin Orasan, Catalina Barbu, Lisa Jones, and Violeta Sotirova (2000). "Coreference and Anaphora: Developing Annotating Tools, Annotated Resources and Annotation Strategies". In: *Proceedings of the Discourse, Anaphora and Reference Resolution Conference (DAARC2000)*, pp. 49–58.

[Miwa and Bansal, 2016]
Makoto Miwa and Mohit Bansal (Aug. 2016). "End-To-End Relation Extraction Using LSTMs on Sequences and Tree Structures". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. ACL, pp. 1105–1116.

[Miyao et al., 2009]
Yusuke Miyao, Kenji Sagae, Rune Sætre, Takuya Matsuzaki, and Jun'ichi Tsujii (2009). "Evaluating Contributions of Natural Language Parsers to Protein–Protein Interaction Extraction". In: *Bioinformatics* 25.3, pp. 394–400. DOI: 10.1093/bioinformatics/btn631.

[Mnih et al., 2015]
Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis (Feb. 2015). "Human-Level Control Through Deep Reinforcement Learning". In: *Nature* 518.7540, pp. 529–533. DOI: 10.1038/nature14236.

[Moens, 2006]
Marie-Francine Moens (2006). *Information Extraction: Algorithms and Prospects in a Retrieval Context*. Vol. 21. The Information Retrieval Series. Springer Netherlands. DOI: 10.1007/978-1-4020-4993-4.

[Mohamed et al., 2011]
Thahir Mohamed, Estevam Hruschka, and Tom Mitchell (July 2011). "Discovering Relations Between Noun Categories". In: *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. ACL, pp. 1447–1455.

[Mohri et al., 2012]
Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar (2012). *Foundations of Machine Learning*. Adaptive Computation and Machine Learning. Cambridge, MA: MIT Press.

[Moro, H. Li, and Krause, et al., 2013]
Andrea Moro, Hong Li, Sebastian Krause, Feiyu Xu, Roberto Navigli, and Hans Uszkoreit (2013). "Semantic Rule Filtering for Web-Scale Relation Extraction". In: *The Semantic Web – ISWC 2013 – Proceedings of 12th International Semantic Web Conference*. Vol. 8218. Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 347–362. DOI: 10.1007/978-3-642-41335-3_22.

[Moro and Navigli, 2012]
Andrea Moro and Roberto Navigli (2012). "WiSeNet: Building a Wikipedia-Based Semantic Network with Ontologized Relations". In: *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*. ACM, pp. 1672–1676. DOI: 10.1145/2396761.2398495.

[Moro and Navigli, 2013]
Andrea Moro and Roberto Navigli (2013). "Integrating Syntactic and Semantic Analysis Into the Open Information Extraction Paradigm". In: *IJCAI 2013, Proceedings of the 23rd International Joint Conference on Artificial Intelligence*. IJCAI/AAAI, pp. 2148–2154.

[Moro et al., 2014]
Andrea Moro, Alessandro Raganato, and Roberto Navigli (2014). "Entity Linking Meets Word Sense Disambiguation: A Unified Approach". In: *Transactions of the Association for Computational Linguistics* 2, pp. 231–244.

[Movshovitz-Attias and Cohen, 2015]
Dana Movshovitz-Attias and William W. Cohen (July 2015). "KB-LDA: Jointly Learning a Knowledge Base of Hierarchy, Relations, and Facts". In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. ACL, pp. 1449–1459.

[MSR, 2014]
MSR (Apr. 2014). *Anticipating More from Cortana*. Microsoft Research Blog, `http://research.microsoft.com/en-us/news/features/cortana-041614.aspx`, last access: 2016-05-02.

[MUC, 1991]
MUC (1991). *Proceedings of the 3rd Conference on Message Understanding, MUC 1991, San Diego, California, USA, May 21-23, 1991*. San Mateo, CA: Morgan Kaufmann Publishers / ACL.

[MUC, 1992]
MUC (1992). *Proceedings of the 4th Conference on Message Understanding, MUC 1992, McLean, Virginia, USA, June 16-18, 1992*. San Mateo, CA: Morgan Kaufmann Publishers / ACL.

[MUC, 1993]
MUC (1993). *Proceedings of the 5th Conference on Message Understanding, MUC 1993, Baltimore, Maryland, USA, August 25-27, 1993*. San Francisco, CA: Morgan Kaufmann Publishers / ACL.

[MUC, 1995]
MUC (1995). *Proceedings of the 6th Conference on Message Understanding, MUC 1995, Columbia, Maryland, USA, November 6-8, 1995*. San Francisco, CA: Morgan Kaufmann Publishers / ACL.

[MUC, 1998]
MUC (1998). *Proceedings of the 7th Conference on Message Understanding, MUC 1998, Fairfax, Virginia, April 29 - May 1, 1998*. ACL.

[Muslea, 1999]
Ion Muslea (1999). "Extraction Patterns for Information Extraction Tasks: A Survey". In: *Papers from the AAAI Workshop on Machine Learning for Information Extraction*. The AAAI Press.

[Nadeau and Sekine, 2007]
David Nadeau and Satoshi Sekine (Aug. 2007). "A Survey of Named Entity Recognition and Classification". In: *Lingvisticae Investigationes* 30.1, pp. 3–26.

[Nakashole et al., 2010]
Ndapandula Nakashole, Martin Theobald, and Gerhard Weikum (2010). "Find Your Advisor:

Robust Knowledge Gathering from the Web". In: *Proceedings of the 13th International Workshop on the Web and Databases 2010, WebDB 2010*. DOI: 10.1145/1859127.1859136.

[Nakashole et al., 2011]
Ndapandula Nakashole, Martin Theobald, and Gerhard Weikum (2011). "Scalable Knowledge Harvesting with High Precision and High Recall". In: *Proceedings of the Forth International Conference on Web Search and Web Data Mining, WSDM 2011*, pp. 227–236.

[Nakashole et al., 2012a]
Ndapandula Nakashole, Gerhard Weikum, and Fabian M. Suchanek (2012a). "Discovering and Exploring Relations on the Web". In: *Proceedings of the VLDB Endowment* 5.12, pp. 1982–1985.

[Nakashole et al., 2012b]
Ndapandula Nakashole, Gerhard Weikum, and Fabian M. Suchanek (July 2012b). "PATTY: A Taxonomy of Relational Patterns with Semantic Types". In: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. ACL, pp. 1135–1145.

[Nastase et al., 2013]
Vivi Nastase, Preslav Nakov, Diarmuid Ó Séaghdha, and Stan Szpakowicz (2013). *Semantic Relations Between Nominals*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers. DOI: 10.2200/S00489ED1V01Y201303HLT019.

[Nastase et al., 2010]
Vivi Nastase, Michael Strube, Benjamin Boerschinger, Caecilia Zirn, and Anas Elghafari (May 2010). "WikiNet: A Very Large Scale Multi-Lingual Concept Network". In: *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10)*. ELRA.

[Navigli, 2009]
Roberto Navigli (2009). "Word Sense Disambiguation: A Survey". In: *ACM Computing Surveys* 41.2, 10:1–10:69. DOI: 10.1145/1459352.1459355.

[Navigli and Ponzetto, 2012]
Roberto Navigli and Simone Paolo Ponzetto (2012). "BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network". In: *Artificial Intelligence* 193, pp. 217–250.

[V. Ng, 2010]
Vincent Ng (July 2010). "Supervised Noun Phrase Coreference Research: The First Fifteen Years". In: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. ACL, pp. 1396–1411.

[V. Ng and Cardie, 2002]
Vincent Ng and Claire Cardie (July 2002). "Improving Machine Learning Approaches to Coreference Resolution". In: *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*. ACL, pp. 104–111. DOI: 10.3115/1073083.1073102.

[Q. Nguyen et al., 2010]
Quang Nguyen, Domonkos Tikk, and Ulf Leser (2010). "Simple Tricks for Improving Pattern-

Based Information Extraction from the Biomedical Literature". In: *Journal of Biomedical Semantics* 1.9. DOI: 10.1186/2041-1480-1-9.

[T. H. Nguyen and Grishman, 2014]
Thien Huu Nguyen and Ralph Grishman (June 2014). "Employing Word Representations and Regularization for Domain Adaptation of Relation Extraction". In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. ACL, pp. 68–74.

[T. H. Nguyen and Grishman, 2015a]
Thien Huu Nguyen and Ralph Grishman (July 2015a). "Event Detection and Domain Adaptation with Convolutional Neural Networks". In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. ACL, pp. 365–371.

[T. H. Nguyen and Grishman, 2015b]
Thien Huu Nguyen and Ralph Grishman (June 2015b). "Relation Extraction: Perspective from Convolutional Neural Networks". In: *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*. ACL, pp. 39–48.

[T. H. Nguyen et al., 2015]
Thien Huu Nguyen, Barbara Plank, and Ralph Grishman (July 2015). "Semantic Representations for Domain Adaptation: A Case Study on the Tree Kernel-Based Method for Relation Extraction". In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. ACL, pp. 635–644.

[T. V. T. Nguyen and Moschitti, 2011a]
Truc Vien T. Nguyen and Alessandro Moschitti (June 2011a). "End-To-End Relation Extraction Using Distant Supervision from External Semantic Repositories". In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. ACL, pp. 277–282.

[T. V. T. Nguyen and Moschitti, 2011b]
Truc Vien T. Nguyen and Alessandro Moschitti (Nov. 2011b). "Joint Distant and Direct Supervision for Relation Extraction". In: *Proceedings of 5th International Joint Conference on Natural Language Processing*. AFNLP, pp. 732–740.

[Nickel et al., 2016]
Maximilian Nickel, Kevin Murphy, Volker Tresp, and Evgeniy Gabrilovich (Jan. 2016). "A Review of Relational Machine Learning for Knowledge Graphs". In: *Proceedings of the IEEE* 104.1, pp. 11–33. DOI: 10.1109/JPROC.2015.2483592.

[Niklaus et al., 2016]
Christina Niklaus, Bernhard Bermeitinger, Siegfried Handschuh, and André Freitas (Dec. 2016). "A Sentence Simplification System for Improving Relation Extraction". In: *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*. The COLING 2016 Organizing Committee, pp. 170–174.

[NIST, 2003]
NIST (Aug. 2003). *The ACE 2003 Evaluation Plan (DRAFT, but Almost Official) - Entity*

Detection and Tracking (EDT) - Relation Detection and Characterization (RDC). Version 1. National Institute of Standards and Technology.

[NIST, 2004]
NIST (July 2004). *The ACE 2004 Evaluation Plan - Evaluation of the Recognition of ACE Entities, ACE Relations and ACE Events*. Version 7. National Institute of Standards and Technology.

[NIST, 2005]
NIST (Oct. 2005). *The ACE 2005 (ACE05) Evaluation Plan - Evaluation of the Detection and Recognition of ACE Entities, Values, Temporal Expressions, Relations, and Events*. Version 3. National Institute of Standards and Technology.

[Nivre et al., 2016]
Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman (May 2016). "Universal Dependencies V1: A Multilingual Treebank Collection". In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. ELRA, pp. 1659–1666.

[Nivre et al., 2007]
Joakim Nivre, Johan Hall, Jens Nilsson, Atanas Chanev, Gülsen Eryigit, Sandra Kübler, Svetoslav Marinov, and Erwin Marsi (2007). "MaltParser: A Language-Independent System for Data-Driven Dependency Parsing". In: *Natural Language Engineering* 13.2, pp. 95–135. DOI: 10.1017/S1351324906004505.

[Oord et al., 2016]
Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W. Senior, and Koray Kavukcuoglu (2016). "WaveNet: A Generative Model for Raw Audio". In: *CoRR* abs/1609.03499. arXiv: 1609.03499 [cs.SD].

[Oremus, 2014]
Will Oremus (Mar. 2014). *The First News Report on the L.A. Earthquake Was Written by a Robot*. Published online in Slate magazine, http://www.slate.com/blogs/future_tense/2014/03/17/quakebot_los_angeles_times_robot_journalist_writes_article_on_la_earthquake.html, last access: 2017-05-30.

[Pantel and Pennacchiotti, 2006]
Patrick Pantel and Marco Pennacchiotti (July 2006). "Espresso: Leveraging Generic Patterns for Automatically Harvesting Semantic Relations". In: *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*. ACL, pp. 113–120.

[Pantel et al., 2004]
Patrick Pantel, Deepak Ravichandran, and Eduard H. Hovy (Aug. 2004). "Towards Terascale Semantic Acquisition". In: *Proceedings of Coling 2004*. COLING, pp. 771–777.

[Parker et al., 2011]
Robert Parker, David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda (2011). *English Gigaword Fifth Edition LDC2011T07*. Web Download, https://catalog.ldc.upenn.edu/LDC2011T07, last access: 2017-06-20. Philadelphia: Linguistic Data Consortium.

[Pasca et al., 2006a]
Marius Pasca, Dekang Lin, Jeffrey Bigham, Andrei Lifchits, and Alpa Jain (July 2006a). "Names and Similarities on the Web: Fact Extraction in the Fast Lane". In: *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*. ACL, pp. 809–816.

[Pasca et al., 2006b]
Marius Pasca, Dekang Lin, Jeffrey Bigham, Andrei Lifchits, and Alpa Jain (2006b). "Organizing and Searching the World Wide Web of Facts - Step One: The One-Million Fact Extraction Challenge". In: *Proceedings of the 21st National Conference on Artificial Intelligence and the 18th Innovative Applications of Artificial Intelligence Conference (AAAI 2006)*. AAAI Press, pp. 1400–1405.

[Peng et al., 2016]
Haoruo Peng, Yangqiu Song, and Dan Roth (Nov. 2016). "Event Detection and Co-Reference with Minimal Supervision". In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. ACL, pp. 392–402.

[Pennington et al., 2014]
Jeffrey Pennington, Richard Socher, and Christopher Manning (Oct. 2014). "Glove: Global Vectors for Word Representation". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. ACL, pp. 1532–1543.

[Petrov et al., 2012]
Slav Petrov, Dipanjan Das, and Ryan McDonald (May 2012). "A Universal Part-Of-Speech Tagset". In: *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*. ELRA, pp. 2089–2096.

[Petrov and McDonald, 2012]
Slav Petrov and Ryan McDonald (2012). "Overview of the 2012 Shared Task on Parsing the Web". In: *First Workshop on Syntactic Analysis of Non-Canonical Language (SANCL)*.

[Pighin et al., 2014]
Daniele Pighin, Marco Cornolti, Enrique Alfonseca, and Katja Filippova (June 2014). "Modelling Events Through Memory-Based, Open-IE Patterns for Abstractive Summarization". In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014), Volume 1: Long Papers*. ACL, pp. 892–901.

[Piskorski and Yangarber, 2013]
Jakub Piskorski and Roman Yangarber (2013). "Information Extraction: Past, Present and Future". In: *Multi-Source, Multilingual Information Extraction and Summarization*. Ed. by Thierry Poibeau, Horacio Saggion, Jakub Piskorski, and Roman Yangarber. Theory and Applications of Natural Language Processing. Berlin, Heidelberg: Springer, pp. 23–49. DOI: 10.1007/978-3-642-28569-1_2.

[Plank and Moschitti, 2013]
Barbara Plank and Alessandro Moschitti (Aug. 2013). "Embedding Semantic Similarity in Tree Kernels for Domain Adaptation of Relation Extraction". In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. ACL, pp. 1498–1507.

[Poon et al., 2010]
Hoifung Poon, Janara Christensen, Pedro Domingos, Oren Etzioni, Raphael Hoffmann, Chloe Kiddon, Thomas Lin, Xiao Ling, Mausam, Alan Ritter, Stefan Schoenmackers, Stephen Soderland, Dan Weld, Fei Wu, and Congle Zhang (2010). "Machine Reading at the University of Washington". In: *Proceedings of the Workshop on Formalisms and Methodology for Learning by Reading (FAM-LbR) at NAACL 2010*. ACL, pp. 87–95.

[Poon and Domingos, 2009]
Hoifung Poon and Pedro M. Domingos (2009). "Unsupervised Semantic Parsing". In: *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, EMNLP 2009*. ACL, pp. 1–10.

[Porter, 1980]
M.F. Porter (1980). "An Algorithm for Suffix Stripping". In: *Program* 14.3, pp. 130–137. DOI: 10.1108/eb046814.

[Pradhan et al., 2012]
Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang (July 2012). "CoNLL-2012 Shared Task: Modeling Multilingual Unrestricted Coreference in OntoNotes". In: *Joint Conference on EMNLP and CoNLL - Shared Task*. ACL, pp. 1–40.

[Pradhan et al., 2011]
Sameer Pradhan, Lance Ramshaw, Mitchell Marcus, Martha Palmer, Ralph Weischedel, and Nianwen Xue (June 2011). "CoNLL-2011 Shared Task: Modeling Unrestricted Coreference in OntoNotes". In: *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*. ACL, pp. 1–27.

[Qian, 2013]
Richard Qian (Mar. 2013). *Understand Your World with Bing*. Bing blogs, http://blogs.bing.com/search/2013/03/21/understand-your-world-with-bing/, last access: 2017-03-17.

[Quirk and Poon, 2016]
Chris Quirk and Hoifung Poon (2016). "Distant Supervision for Relation Extraction Beyond the Sentence Boundary". In: *CoRR* abs/1609.04873. arXiv: 1609.04873 [cs.CL].

[Raghunathan et al., 2010]
Karthik Raghunathan, Heeyoung Lee, Sudarshan Rangarajan, Nathanael Chambers, Mihai Surdeanu, Dan Jurafsky, and Christopher Manning (2010). "A Multi-Pass Sieve for Coreference Resolution". In: *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. ACL, pp. 492–501.

[Ravichandran and Hovy, 2002]
Deepak Ravichandran and Eduard H. Hovy (2002). "Learning Surface Text Patterns for a Question Answering System". In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 41–47.

[Ray and Craven, 2001]
Soumya Ray and Mark Craven (2001). "Representing Sentence Structure in Hidden Markov Models for Information Extraction". In: *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence, IJCAI 2001*. Morgan Kaufmann, pp. 1273–1279.

[Read et al., 2012]
Jonathon Read, Rebecca Dridan, Stephan Oepen, and Lars Jørgen Solberg (Dec. 2012). "Sentence Boundary Detection: A Long Solved Problem?" In: *Proceedings of COLING 2012: Posters*. The COLING 2012 Organizing Committee, pp. 985–994.

[Recasens and Hovy, 2011]
Marta Recasens and Eduard H. Hovy (2011). "BLANC: Implementing the Rand Index for Coreference Evaluation". In: *Natural Language Engineering* 17.4, pp. 485–510.

[Recasens et al., 2010]
Marta Recasens, Lluís Màrquez, Emili Sapena, M. Antònia Martí, Mariona Taulé, Véronique Hoste, Massimo Poesio, and Yannick Versley (July 2010). "SemEval-2010 Task 1: Coreference Resolution in Multiple Languages". In: *Proceedings of the 5th International Workshop on Semantic Evaluation*. ACL, pp. 1–8.

[Rehm and Uszkoreit, 2013]
Georg Rehm and Hans Uszkoreit, eds. (2013). *META-NET Strategic Research Agenda for Multilingual Europe 2020*. White Paper Series. Berlin Heidelberg: Springer. DOI: 10.1007/978-3-642-36349-8.

[Riedel et al., 2010]
Sebastian Riedel, Limin Yao, and Andrew McCallum (2010). "Modeling Relations and Their Mentions without Labeled Text". In: *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD '10)*. Vol. 6323. Lecture Notes in Computer Science. Springer, pp. 148–163. DOI: 10.1007/978-3-642-15939-8_10.

[Riedel et al., 2013]
Sebastian Riedel, Limin Yao, Andrew McCallum, and Benjamin M. Marlin (June 2013). "Relation Extraction with Matrix Factorization and Universal Schemas". In: *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. ACL, pp. 74–84.

[Riloff, 1996]
Ellen Riloff (1996). "Automatically Generating Extraction Patterns from Untagged Text". In: *Proceedings of Thirteenth National Conference on Artificial Intelligence (AAAI-96)*. The AAAI Press/MIT Press, pp. 1044–1049.

[Rocktäschel et al., 2015]
Tim Rocktäschel, Sameer Singh, and Sebastian Riedel (May 2015). "Injecting Logical Background Knowledge Into Embeddings for Relation Extraction". In: *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. ACL, pp. 1119–1129.

[Roller et al., 2015]
Roland Roller, Eneko Agirre, Aitor Soroa, and Mark Stevenson (July 2015). "Improving Distant Supervision Using Inference Learning". In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. ACL, pp. 273–278.

[Rosario and Hearst, 2004]
Barbara Rosario and Marti A. Hearst (July 2004). "Classifying Semantic Relations in Bioscience Texts". In: *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04), Main Volume*. ACL, pp. 430–437. DOI: 10.3115/1218955.1219010.

[Rosenberg, 2016]
Seth Rosenberg (Apr. 2016). *How to Build Bots for Messenger*. Blog entry on 'facebook for developers', https://developers.facebook.com/blog/post/2016/04/12/bots-for-messenger/, last access: 2016-05-02.

[Ruppenhofer et al., 2010]
Josef Ruppenhofer, Michael Ellsworth, Miriam R. L. Petruck, Christopher R. Johnson, and Jan Scheffczyk (2010). *FrameNet II: Extended Theory and Practice*. September 14, 2010. https://framenet2.icsi.berkeley.edu/docs/r1.5/book.pdf, last access: 2017-02-25. International Computer Science Institute, Berkeley.

[Sachan et al., 2015]
Mrinmaya Sachan, Eduard H. Hovy, and Eric P. Xing (2015). "An Active Learning Approach to Coreference Resolution". In: *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015*. AAAI Press, pp. 1312–1318.

[Sangeetha and Arock, 2012]
S. Sangeetha and Michael Arock (2012). "Event Coreference Resolution Using Mincut Based Graph Clustering". In: *The Fourth International Workshop on Computer Networks & Communications*, pp. 253–260. DOI: 10.5121/csit.2012.2422.

[Santorini, 1990]
Beatrice Santorini (July 1990). *Part-Of-Speech Tagging Guidelines for the Penn Treebank Project*. Tech. rep. MS-CIS-90-47. http://repository.upenn.edu/cis_reports/570/, last access: 2017-05-11. University of Pennsylvania, Department of Computer and Information Science.

[Sarawagi, 2008]
Sunita Sarawagi (2008). "Information Extraction". In: *Foundations and Trends in Databases* 1.3, pp. 261–377.

[Sarikaya et al., 2016]
Ruhi Sarikaya, Paul A. Crook, Alex Marin, Minwoo Jeong, Jean-Philippe Robichaud, Asli Çelikyilmaz, Young-Bum Kim, Alexandre Rochette, Omar Zia Khan, Xiaohu Liu, Daniel Boies, Tasos Anastasakos, Zhaleh Feizollahi, Nikhil Ramesh, H. Suzuki, Roman Holenstein, Elizabeth Krawczyk, and Vasiliy Radostev (Dec. 2016). "An Overview of End-To-End Language Understanding and Dialog Management for Personal Digital Assistants". In: *2016 IEEE Spoken Language Technology Workshop (SLT 2016)*. IEEE, pp. 391–397. DOI: 10.1109/SLT.2016.7846294.

[Sasano et al., 2009]
Ryohei Sasano, Daisuke Kawahara, and Sadao Kurohashi (June 2009). "The Effect of Corpus Size on Case Frame Acquisition for Discourse Analysis". In: *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. ACL, pp. 521–529.

[Scheffczyk et al., 2006]
Jan Scheffczyk, Adam Pease, and Michael Ellsworth (2006). "Linking FrameNet to the Suggested Upper Merged Ontology". In: *Formal Ontology in Information Systems, Proceedings of the Fourth International Conference, FOIS 2006*. Vol. 150. Frontiers in Artificial Intelligence and Applications. IOS Press, pp. 289–300.

[Schmidek and Barbosa, 2014]
Jordan Schmidek and Denilson Barbosa (May 2014). "Improving Open Relation Extraction via Sentence Re-Structuring". In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. ELRA, pp. 3720–3723.

[Schmidhuber, 2015]
Jürgen Schmidhuber (2015). "Deep Learning in Neural Networks: An Overview". In: *Neural Networks* 61, pp. 85–117. DOI: 10.1016/j.neunet.2014.09.003.

[Schuler, 2005]
Karin Kipper Schuler (2005). "Verbnet: A Broad-Coverage, Comprehensive Verb Lexicon". PhD thesis. Philadelphia, PA, USA: University of Pennsylvania.

[Sekine, 2006]
Satoshi Sekine (July 2006). "On-Demand Information Extraction". In: *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*. ACL, pp. 731–738.

[Sekine and Nobata, 2004]
Satoshi Sekine and Chikashi Nobata (2004). "Definition, Dictionaries and Tagger for Extended Named Entity Hierarchy". In: *Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC 2004*. ELRA, pp. 1977–1980.

[Sekine et al., 2002]
Satoshi Sekine, Kiyoshi Sudo, and Chikashi Nobata (2002). "Extended Named Entity Hierarchy". In: *Proceedings of the Third International Conference on Language Resources and Evaluation, LREC 2002*. ELRA, pp. 1818–1824.

[Sergienya and Schütze, 2015]
Irina Sergienya and Hinrich Schütze (Sept. 2015). "Learning Better Embeddings for Rare Words Using Distributional Representations". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. ACL, pp. 280–285.

[Shah et al., 2016]
Pararth Shah, Dilek Hakkani-Tür, and Larry Heck (2016). "Interactive Reinforcement Learning for Task-Oriented Dialogue Management". In: *Workshop on Deep Learning for Action and Interaction at NIPS 2016*.

[Shinyama and Sekine, 2006]
Yusuke Shinyama and Satoshi Sekine (June 2006). "Preemptive Information Extraction Using Unrestricted Relation Discovery". In: *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL 2006)*. ACL, pp. 304–311.

[Siefkes and Siniakov, 2005]
Christian Siefkes and Peter Siniakov (2005). "An Overview and Classification of Adaptive

Approaches to Information Extraction". In: *Journal on Data Semantics IV*. Lecture Notes in Computer Science book series (LNCS, volume 3730). Springer, pp. 172–212. DOI: 10.1007/11603412_6.

[Silver et al., 2016]
David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis (Jan. 2016). "Mastering the Game of Go with Deep Neural Networks and Tree Search". In: *Nature* 529.7587, pp. 484–489. DOI: 10.1038/nature16961.

[Singh et al., 2013]
Sameer Singh, Sebastian Riedel, Brian Martin, Jiaping Zheng, and Andrew McCallum (2013). "Joint Inference of Entities, Relations, and Coreference". In: *Proceedings of the 2013 Workshop on Automated Knowledge Base Construction, AKBC@CIKM 13*. ACM, pp. 1–6. DOI: 10.1145/2509558.2509559.

[Singhal, 2012]
Amit Singhal (May 2012). *Introducing the Knowledge Graph: Things, Not Strings*. Google Official Blog, https://googleblog.blogspot.de/2012/05/introducing-knowledge-graph-things-not.html, last access: 2017-06-20.

[Skounakis et al., 2003]
Marios Skounakis, Mark Craven, and Soumya Ray (2003). "Hierarchical Hidden Markov Models for Information Extraction". In: *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence (IJCAI-03)*. Morgan Kaufmann, pp. 427–433.

[Snijders et al., 2012]
Chris Snijders, Uwe Matzat, and Ulf-Dietrich Reips (2012). "'Big Data': Big Gaps of Knowledge in the Field of Internet Science". In: *International Journal of Internet Science* 7.1, pp. 1–5.

[Snow et al., 2004]
Rion Snow, Daniel Jurafsky, and Andrew Y. Ng (2004). "Learning Syntactic Patterns for Automatic Hypernym Discovery". In: *Advances in Neural Information Processing Systems 17, NIPS 2004*, pp. 1297–1304.

[Socher et al., 2012]
Richard Socher, Brody Huval, Christopher D. Manning, and Andrew Y. Ng (July 2012). "Semantic Compositionality Through Recursive Matrix-Vector Spaces". In: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. ACL, pp. 1201–1211.

[Socher et al., 2013]
Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts (Oct. 2013). "Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank". In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. ACL, pp. 1631–1642.

[Soderland et al., 2010]
Stephen Soderland, Brendan Roof, Bo Qin, Shi Xu, Mausam, and Oren Etzioni (2010). "Adapting Open Information Extraction to Domain-Specific Relations". In: *AI Magazine* 31.3, pp. 93–102.

[Song et al., 2016]
Zhiyi Song, Ann Bies, Stephanie Strassel, Joe Ellis, Teruko Mitamura, Hoa Trang Dang, Yukari Yamakawa, and Sue Holm (2016). "Event Nugget and Event Coreference Annotation". In: *Proceedings of the Fourth Workshop on Events*. ACL, pp. 37–45.

[Song et al., 2015]
Zhiyi Song, Ann Bies, Stephanie Strassel, Tom Riese, Justin Mott, Joe Ellis, Jonathan Wright, Seth Kulick, Neville Ryant, and Xiaoyi Ma (June 2015). "From Light to Rich ERE: Annotation of Entities, Relations, and Events". In: *Proceedings of the the 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation*. ACL, pp. 89–98.

[Soon et al., 2001]
Wee Meng Soon, Hwee Tou Ng, and Chung Yong Lim (2001). "A Machine Learning Approach to Coreference Resolution of Noun Phrases". In: *Computational Linguistics* 27.4, pp. 521–544.

[Speer et al., 2016]
Robert Speer, Joshua Chin, and Catherine Havasi (2016). "ConceptNet 5.5: An Open Multilingual Graph of General Knowledge". In: *CoRR* abs/1612.03975. arXiv: 1612.03975 [cs.CL].

[Speer and Havasi, 2012]
Robert Speer and Catherine Havasi (May 2012). "Representing General Relational Knowledge in ConceptNet 5". In: *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*. ELRA, pp. 3679–3686.

[Speer and Havasi, 2013]
Robert Speer and Catherine Havasi (2013). "ConceptNet 5: A Large Semantic Network for Relational Knowledge". In: *The People's Web Meets NLP*. Ed. by Iryna Gurevych and Jungi Kim. Theory and Applications of Natural Language Processing. Berlin, Heidelberg: Springer, pp. 161–176. DOI: 10.1007/978-3-642-35085-6_6.

[Stevenson, 2004]
Mark Stevenson (Aug. 2004). "Information Extraction from Single and Multiple Sentences". In: *Proceedings of Coling 2004*. COLING, pp. 875–881.

[Stevenson, 2006]
Mark Stevenson (2006). "Fact Distribution in Information Extraction". In: *Language Resources and Evaluation* 40.2, pp. 183–201.

[Stevenson and Greenwood, 2005]
Mark Stevenson and Mark A. Greenwood (June 2005). "A Semantic Approach to IE Pattern Induction". In: *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*. ACL, pp. 379–386. DOI: 10.3115/1219840.1219887.

[Suchanek et al., 2006a]
Fabian M. Suchanek, Georgiana Ifrim, and Gerhard Weikum (2006a). "Combining Linguistic

and Statistical Analysis to Extract Relations from Web Documents". In: *Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, pp. 712–717.

[Suchanek et al., 2006b]
Fabian M. Suchanek, Georgiana Ifrim, and Gerhard Weikum (2006b). "LEILA: Learning to Extract Information by Linguistic Analysis". In: *Proceedings of the 2nd Workshop on Ontology Learning and Population (OLP2)*. ACL, pp. 18–25.

[Suchanek et al., 2007]
Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum (2007). "YAGO: A Core of Semantic Knowledge". In: *16th International World Wide Web Conference (WWW 2007)*. ACM Press, pp. 697–706.

[Suchanek et al., 2008]
Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum (2008). "YAGO: A Large Ontology from Wikipedia and WordNet". In: *Web Semantics: Science, Services and Agents on the World Wide Web* 6.3, pp. 203–217. DOI: 10.1016/j.websem.2008.06.001.

[Suchanek et al., 2009]
Fabian M. Suchanek, Mauro Sozio, and Gerhard Weikum (2009). "SOFIE: A Self-Organizing Framework for Information Extraction". In: *Proceedings of the 18th International Conference on World Wide Web, WWW 2009*. ACM, pp. 631–640.

[A. Sun and Grishman, 2010]
Ang Sun and Ralph Grishman (Aug. 2010). "Semi-Supervised Semantic Pattern Discovery with Guidance from Unsupervised Pattern Clusters". In: *Coling 2010: Posters*. Coling 2010 Organizing Committee, pp. 1194–1202.

[A. Sun et al., 2011]
Ang Sun, Ralph Grishman, and Satoshi Sekine (June 2011). "Semi-Supervised Relation Extraction with Large-Scale Word Clustering". In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. ACL, pp. 521–529.

[Y. Sun et al., 2015]
Yaming Sun, Lei Lin, Duyu Tang, Nan Yang, Zhenzhou Ji, and Xiaolong Wang (2015). "Modeling Mention, Context and Entity with Neural Networks for Entity Disambiguation". In: *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015*. AAAI Press, pp. 1333–1339.

[Sundheim, 1995]
Beth M. Sundheim (1995). "Overview of the Results of the MUC-6 Evaluation". In: *Proceedings of the 6th Conference on Message Understanding, MUC 1995*, pp. 13–31.

[Surdeanu, 2013]
Mihai Surdeanu (2013). "Overview of the TAC 2013 Knowledge Base Population Evaluation: English Slot Filling and Temporal Slot Filling". In: *Proceedings of the Text Analysis Conference*.

[Surdeanu and Ji, 2014]
Mihai Surdeanu and Heng Ji (2014). "Overview of the English Slot Filling Track at the TAC2014 Knowledge Base Population Evaluation". In: *Proceedings of the TAC-KBP 2014 Workshop*.

[Swampillai and Stevenson, 2010]
Kumutha Swampillai and Mark Stevenson (2010). "Inter-Sentential Relations in Information Extraction Corpora". In: *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010*. ELRA, pp. 2637–2641.

[Swampillai and Stevenson, 2011]
Kumutha Swampillai and Mark Stevenson (2011). "Extracting Relations Within and Across Sentences". In: *Proceedings of Recent Advances in Natural Language Processing, RANLP 2011*. RANLP 2011 Organising Committee, pp. 25–32.

[Swartz, 2013]
Aaron Swartz (2013). *A Programmable Web: An Unfinished Work*. Ed. by James Hendler and Ying Ding. Synthesis Lectures on The Semantic Web: Theory and Technology. Morgan & Claypool Publishers.

[Takamatsu et al., 2012]
Shingo Takamatsu, Issei Sato, and Hiroshi Nakagawa (July 2012). "Reducing Wrong Labels in Distant Supervision for Relation Extraction". In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. ACL, pp. 721–729.

[Taylor et al., 2003]
Ann Taylor, Mitchell Marcus, and Beatrice Santorini (2003). "The Penn Treebank: An Overview". In: *Treebanks: Building and Using Parsed Corpora*. Ed. by Anne Abeillé. Dordrecht: Springer Netherlands, pp. 5–22. DOI: 10.1007/978-94-010-0201-1_1.

[Text Analytics Group, 2013]
Text Analytics Group (2013). *Annotation Guidelines for the CockrACE Corpus*. version 2014-01-16; available at https://dfki-lt-re-group.bitbucket.io/downloads/cockrace/annotation-guidelines-2014-01-16.pdf, last access: 2017-05-15. Berlin, Germany.

[Thomas, 2015]
Philippe Thomas (2015). "Robust Relationship Extraction in the Biomedical Domain". PhD thesis. Humboldt University of Berlin, Faculty of Mathematics and Natural Sciences.

[Tjong Kim Sang and De Meulder, 2003]
Erik F. Tjong Kim Sang and Fien De Meulder (2003). "Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition". In: *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pp. 142–147.

[Toutanova et al., 2015]
Kristina Toutanova, Danqi Chen, Patrick Pantel, Hoifung Poon, Pallavi Choudhury, and Michael Gamon (Sept. 2015). "Representing Text for Joint Embedding of Text and Knowledge Bases". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. ACL, pp. 1499–1509.

[Turmo et al., 2006]
Jordi Turmo, Alicia Ageno, and Neus Català (July 2006). "Adaptive Information Extraction". In: *ACM Computing Surveys* 38.2 (2). DOI: 10.1145/1132956.1132957.

[Unger et al., 2013]
Christina Unger, John P. McCrae, Sebastian Walter, Sara Winter, and Philipp Cimiano (2013). "A Lemon Lexicon for DBpedia". In: *Proceedings of the NLP & DBpedia Workshop Co-Located with the 12th International Semantic Web Conference (ISWC 2013)*. Vol. 1064. CEUR Workshop Proceedings. CEUR-WS.org.

[Uszkoreit, 2011]
Hans Uszkoreit (2011). "Learning Relation Extraction Grammars with Minimal Human Intervention: Strategy, Results, Insights and Plans". In: *Computational Linguistics and Intelligent Text Processing*. Ed. by Alexander Gelbukh. Vol. 6609. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, pp. 106–126.

[Vilain et al., 1995]
Marc B. Vilain, John D. Burger, John S. Aberdeen, Dennis Connolly, and Lynette Hirschman (1995). "A Model-Theoretic Coreference Scoring Scheme". In: *Proceedings of the 6th Conference on Message Understanding, MUC 1995*, pp. 45–52.

[Volokh, 2010]
Alexander Volokh (Feb. 2010). *MDParser - Technical Report for Version 4.0*. Tech. rep. http://mdparser.sb.dfki.de/report.pdf, last access: 2017-06-28. DFKI.

[Volokh and Neumann, 2010]
Alexander Volokh and Günter Neumann (2010). "372:Comparing the Benefit of Different Dependency Parsers for Textual Entailment Using Syntactic Constraints Only". In: *Proceedings of the 5th International Workshop on Semantic Evaluation*. ACL, pp. 308–312.

[Vrandečić and Krötzsch, 2014]
Denny Vrandečić and Markus Krötzsch (2014). "Wikidata: A Free Collaborative Knowledgebase". In: *Communications of the ACM* 57 (10), pp. 78–85.

[Walker et al., 2006]
Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda (Feb. 2006). *ACE 2005 Multilingual Training Corpus LDC2006T06*. Web Download, https://catalog.ldc.upenn.edu/LDC2006T06, last access: 2017-06-20. Philadelphia: Linguistic Data Consortium.

[Walter et al., 2014a]
Sebastian Walter, Christina Unger, and Philipp Cimiano (Nov. 2014a). "ATOLL - A Framework for the Automatic Induction of Ontology Lexica". In: *Data & Knowledge Engineering* 94, Part B. S.I. following NLDB 2013, pp. 148–162. DOI: 10.1016/j.datak.2014.09.003.

[Walter et al., 2014b]
Sebastian Walter, Christina Unger, and Philipp Cimiano (2014b). "M-Atoll: A Framework for the Lexicalization of Ontologies in Multiple Languages". In: *The Semantic Web - ISWC 2014*. Vol. 8796. Lecture Notes in Computer Science. Springer International Publishing, pp. 472–486. DOI: 10.1007/978-3-319-11964-9_30.

[Wang et al., 2010]
Yafang Wang, Mingjie Zhu, Lizhen Qu, Marc Spaniol, and Gerhard Weikum (2010). "Timely YAGO: Harvesting, Querying, and Visualizing Temporal Knowledge from Wikipedia". In: *EDBT 2010, 13th International Conference on Extending Database Technology*. Vol. 426. ACM International Conference Proceeding Series. ACM, pp. 697–700. DOI: 10.1145/1739041.1739130.

[Weikum and Theobald, 2010]
Gerhard Weikum and Martin Theobald (2010). "From Information to Knowledge: Harvesting Entities and Relationships from Web Sources". In: *Proceedings of the Twenty-Ninth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, PODS 2010*. ACM, pp. 65–76.

[Weischedel et al., 2013]
Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, Mohammed El-Bachouti, Robert Belvin, and Ann Houston (2013). *OntoNotes Release 5.0 LDC2013T19*. Web Download, https://catalog.ldc.upenn.edu/LDC2013T19, last access: 2017-06-20. Philadelphia: Linguistic Data Consortium.

[Weissenborn et al., 2015a]
Dirk Weissenborn, Leonhard Hennig, Feiyu Xu, and Hans Uszkoreit (July 2015a). "Multi-Objective Optimization for the Joint Disambiguation of Nouns and Named Entities". In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. ACL, pp. 596–605.

[Weissenborn et al., 2015b]
Dirk Weissenborn, Feiyu Xu, and Hans Uszkoreit (June 2015b). "DFKI: Multi-Objective Optimization for the Joint Disambiguation of Entities and Nouns & Deep Verb Sense Disambiguation". In: *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*. ACL, pp. 335–339.

[Weston et al., 2013]
Jason Weston, Antoine Bordes, Oksana Yakhnenko, and Nicolas Usunier (Oct. 2013). "Connecting Language and Knowledge Bases with Embedding Models for Relation Extraction". In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. ACL, pp. 1366–1371.

[Wiebe et al., 2005]
Janyce Wiebe, Theresa Wilson, and Claire Cardie (2005). "Annotating Expressions of Opinions and Emotions in Language". In: *Language Resources and Evaluation* 39.2-3, pp. 165–210. DOI: 10.1007/s10579-005-7880-9.

[Wiseman et al., 2015]
Sam Wiseman, Alexander M. Rush, Stuart Shieber, and Jason Weston (July 2015). "Learning Anaphoricity and Antecedent Ranking Features for Coreference Resolution". In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. ACL, pp. 1416–1426.

[Witbrock et al., 2015]
Michael J. Witbrock, Karen Pittman, Jessica Moszkowicz, Andrew Beck, Dave Schneider, and Douglas B. Lenat (2015). "Cyc and the Big C: Reading That Produces and Uses Hypotheses About Complex Molecular Biology Mechanisms". In: *Scholarly Big Data: AI Perspectives, Challenges, and Ideas, Papers from the 2015 AAAI Workshop*. Vol. WS-15-13. AAAI Workshops. AAAI Press.

[F. Wu and Weld, 2007]
Fei Wu and Daniel S. Weld (2007). "Autonomously Semantifying Wikipedia". In: *Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management, CIKM 2007*. ACM, pp. 41–50. DOI: 10.1145/1321440.1321449.

[F. Wu and Weld, 2010]
Fei Wu and Daniel S. Weld (July 2010). "Open Information Extraction Using Wikipedia". In: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. ACL, pp. 118–127.

[Y. Wu et al., 2016]
Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean (Sept. 2016). "Google's Neural Machine Translation System: Bridging the Gap Between Human and Machine Translation". In: *CoRR* abs/1609.08144. arXiv: 1609.08144 [cs.CL].

[F. Xu, 2007]
Feiyu Xu (2007). "Bootstrapping Relation Extraction from Semantic Seeds". PhD thesis. Saarland University.

[F. Xu et al., 2002]
Feiyu Xu, Daniela Kurz, Jakub Piskorski, and Sven Schmeier (2002). "A Domain Adaptive Approach to Automatic Acquisition of Domain Relevant Terms and Their Relations with Bootstrapping". In: *Proceedings of the Third International Conference on Language Resources and Evaluation, LREC 2002*. ELRA, pp. 224–230.

[F. Xu et al., 2011]
Feiyu Xu, Hong Li, Yi Zhang, Hans Uszkoreit, and Sebastian Krause (2011). "Minimally Supervised Domain-Adaptive Parse Reranking for Relation Extraction". In: *Proceedings of International Conference on Parsing Technologies (IWPT 2011)*. ACL, pp. 118–128.

[F. Xu et al., 2014]
Feiyu Xu, Hong Li, Yi Zhang, Hans Uszkoreit, and Sebastian Krause (Apr. 2014). "Parse Reranking for Domain-Adaptative Relation Extraction". In: *Journal of Logic and Computation* 24.2, pp. 413–431. DOI: 10.1093/logcom/exs055.

[F. Xu, Uszkoreit, and Krause, et al., 2010]
Feiyu Xu, Hans Uszkoreit, Sebastian Krause, and Hong Li (Aug. 2010). "Boosting Relation Extraction with Limited Closed-World Knowledge". In: *Coling 2010: Posters*. Coling 2010 Organizing Committee, pp. 1354–1362.

[F. Xu et al., 2007]
Feiyu Xu, Hans Uszkoreit, and Hong Li (June 2007). "A Seed-Driven Bottom-Up Machine Learning Framework for Extracting Relations of Various Complexity". In: *Proceedings of ACL 2007, 45th Annual Meeting of the Association for Computational Linguistics*. ACL, pp. 584–591.

[H. Xu et al., 2009]
Hongzhi Xu, Changjian Hu, and Guoyang Shen (2009). "Discovery of Dependency Tree Patterns for Relation Extraction". In: *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation, PACLIC 23*. City University of Hong Kong Press, pp. 851–858.

[Y. Xu et al., 2015]
Yan Xu, Lili Mou, Ge Li, Yunchuan Chen, Hao Peng, and Zhi Jin (Sept. 2015). "Classifying Relations via Long Short Term Memory Networks Along Shortest Dependency Paths". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. ACL, pp. 1785–1794.

[Yang et al., 2015]
Bishan Yang, Claire Cardie, and Peter I. Frazier (2015). "A Hierarchical Distance-Dependent Bayesian Model for Event Coreference Resolution". In: *Transactions of the Association for Computational Linguistics* 3, pp. 517–528.

[Yangarber, 2001]
Roman Yangarber (2001). "Scenario Customization for Information Extraction". PhD thesis. Department of Computer Science, Graduate School of Arts and Science, New York University.

[Yangarber, 2003]
Roman Yangarber (July 2003). "Counter-Training in Discovery of Semantic Patterns". In: *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*. ACL, pp. 343–350.

[Yangarber et al., 2000]
Roman Yangarber, Ralph Grishman, and Pasi Tapanainen (2000). "Automatic Acquisition of Domain Knowledge for Information Extraction". In: *Proceedings of the 18th International Conference on Computational Linguistics*. ACL, pp. 940–946.

[Yao et al., 2010]
Limin Yao, Sebastian Riedel, and Andrew McCallum (Oct. 2010). "Collective Cross-Document Relation Extraction without Labelled Data". In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '10)*. ACL, pp. 1013–1023.

[Yao et al., 2012]
Limin Yao, Sebastian Riedel, and Andrew McCallum (2012). "Unsupervised Relation Discovery with Sense Disambiguation". In: *The 50th Annual Meeting of the Association for Computational Linguistics - Volume 1: Long Papers*. ACL, pp. 712–720.

[Yarowsky, 1993]
David Yarowsky (Mar. 1993). "One Sense Per Collocation". In: *Proceedings of the Human Language Technology Workshop*. HLT. ACL.

[Yates et al., 2007]
Alexander Yates, Michael Cafarella, Michele Banko, Oren Etzioni, Matthew Broadhead, and Stephen Soderland (2007). "TextRunner: Open Information Extraction on the Web". In: *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*. NAACL-Demonstrations '07. Rochester, New York: ACL, pp. 25–26.

[Yates and Etzioni, 2007]
Alexander Yates and Oren Etzioni (2007). "Unsupervised Resolution of Objects and Relations on the Web". In: *Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*. ACL, pp. 121–130.

[Yates and Etzioni, 2009]
Alexander Yates and Oren Etzioni (Mar. 2009). "Unsupervised Methods for Determining Object and Relation Synonyms on the Web". In: *Journal of Artificial Intelligence Research* 34, pp. 255–296.

[Yu et al., 2015]
Mo Yu, Matthew R. Gormley, and Mark Dredze (June 2015). "Combining Word Embeddings and Feature Embeddings for Fine-Grained Relation Extraction". In: *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. ACL, pp. 1374–1379.

[Zeiler and Fergus, 2014]
Matthew D. Zeiler and Rob Fergus (2014). "Visualizing and Understanding Convolutional Networks". In: *Computer Vision - ECCV 2014 - 13th European Conference*. Vol. 8689. Lecture Notes in Computer Science. Springer, pp. 818–833. DOI: 10.1007/978-3-319-10590-1_53.

[Zelenko et al., 2003]
Dmitry Zelenko, Chinatsu Aone, and Anthony Richardella (2003). "Kernel Methods for Relation Extraction". In: *Journal of Machine Learning Research* 3, pp. 1083–1106.

[Zeng et al., 2015]
Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao (Sept. 2015). "Distant Supervision for Relation Extraction via Piecewise Convolutional Neural Networks". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. ACL, pp. 1753–1762.

[Zeng et al., 2014]
Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao (Aug. 2014). "Relation Classification via Convolutional Deep Neural Network". In: *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. Dublin City University and ACL, pp. 2335–2344.

[Chiyuan Zhang et al., 2016]
Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals (2016). "Understanding Deep Learning Requires Rethinking Generalization". In: *CoRR* abs/1611.03530. arXiv: 1611.03530 [cs.LG].

[Congle Zhang et al., 2015]
Congle Zhang, Stephen Soderland, and Daniel Weld (2015). "Exploiting Parallel News Streams for Unsupervised Event Extraction". In: *Transactions of the Association for Computational Linguistics* 3, pp. 117–129.

[Congle Zhang and Weld, 2013]
Congle Zhang and Daniel S. Weld (Oct. 2013). "Harvesting Parallel News Streams to Generate Paraphrases of Event Relations". In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. ACL, pp. 1776–1786.

[K. Zhang and Shasha, 1989]
Kaizhong Zhang and Dennis Shasha (Dec. 1989). "Simple Fast Algorithms for the Editing Distance Between Trees and Related Problems". In: *SIAM Journal on Computing* 18.6, pp. 1245–1262. DOI: 10.1137/0218082.

[M. Zhang et al., 2006]
Min Zhang, Jie Zhang, Jian Su, and GuoDong Zhou (July 2006). "A Composite Kernel to Extract Relations Between Entities with Both Flat and Structured Features". In: *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*. ACL, pp. 825–832. DOI: 10.3115/1220175.1220279.

[Q. Zhang et al., 2008]
Qi Zhang, Fabian M. Suchanek, Lihua Yue, and Gerhard Weikum (2008). "TOB: Timely Ontologies for Business Relations". In: *11th International Workshop on the Web and Databases, WebDB 2008*.

[T. Zhang et al., 2015]
Tongtao Zhang, Hongzhi Li, Heng Ji, and Shih-Fu Chang (Sept. 2015). "Cross-Document Event Coreference Resolution Based on Cross-Media Features". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. ACL, pp. 201–206.

[Y. Zhang and Wallace, 2015]
Ye Zhang and Byron C. Wallace (2015). "A Sensitivity Analysis of (And Practitioners' Guide To) Convolutional Neural Networks for Sentence Classification". In: *CoRR* abs/1510.03820. arXiv: 1510.03820 [cs.CL].

[Z. Zhang, 2004]
Zhu Zhang (2004). "Weakly-Supervised Relation Classification for Information Extraction". In: *Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management (CIKM '04)*. ACM, pp. 581–588. DOI: 10.1145/1031171.1031279.

[D. Zhou and He, 2008]
Deyu Zhou and Yulan He (Apr. 2008). "Extracting Interactions Between Proteins from the Literature". In: *Journal of Biomedical Informatics* 41 (2), pp. 393–407. DOI: 10.1016/j.jbi.2007.11.008.

[G. Zhou et al., 2010]
Guodong Zhou, Longhua Qian, and Jianxi Fan (2010). "Tree Kernel-Based Semantic Relation Extraction with Rich Syntactic and Semantic Information". In: *Information Sciences* 180.8, pp. 1313–1325. DOI: 10.1016/j.ins.2009.12.006.

[G. Zhou et al., 2005]
Guodong Zhou, Jian Su, Jie Zhang, and Min Zhang (2005). "Exploring Various Knowledge in Relation Extraction". In: *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*. ACL, pp. 427–434.

[G. Zhou and M. Zhang, 2007]
Guodong Zhou and Min Zhang (2007). "Extracting Relation Information from Text Documents by Exploring Various Types of Knowledge". In: *Information Processing & Management* 43.4, pp. 969–982. DOI: 10.1016/j.ipm.2006.09.012.