Linguistics to Structure Unstructured Information

Authors: Günter Neumann (DFKI), Gerhard Paaß (Fraunhofer IAIS), David van den Akker (Attensity Europe GmbH)

Abstract

The extraction of semantics of unstructured documents requires the recognition and classification of textual patterns, their variability and their inter-relationships, i.e. the analysis of the linguistic structure of documents. Being the integral part of a larger real-life application, this linguistic analysis process must be robust, fast and adaptable. This creates a big challenge for the development of the necessary linguistic base components. In this drill-down we present several dimensions of this challenge and show how they have been successfully tackled in ORDO.

Introduction

According to Patrick Thibodeau in an article in *Computerworld*, October 2010, "data is expected to grow by 800% over the next five years, and 80% of it will be unstructured data." Methods for extracting useful information from such unstructured sources are still dominated by simple word-level strategies, either based on simple linguistic operations or on no linguistics at all. They mostly just pick up the isolated words, and consequently apply some word-based statistics to generate useful information. For coarse-grained applications like word-based indexing, word-based spelling correction, and word-cloud generation, such techniques have been proven to be satisfactory.

It has been observed, however, that the higher one's demands are for the extraction of meaningful structure from the original text, the less satisfying any word-level analysis becomes. For example, in the case of the extraction of named entities or relations, it is not just the individual words that are important, but also the syntactic structures in which they are configured, as well as the functional roles which these structures have in their clause, and last but not least the semantic context in which these roles and structures appear. Consider as an exemplification an application that is designed to update a company's database with the latest revenue numbers of the competition as they get reported on the public portals that these companies typically maintain. For example, "According to IDC's Worldwide Quarterly Server Tracker, IBM led in factory revenue gains to achieve 30.5 per cent market share in the second quarter of 2011, compared with HP's 29.8 per cent share."

Actually, this integrated word-level and syntax-level approach of Natural Language Processing (NLP) has been a driving power followed in almost all application scenarios and realised in ORDO, as we will make clear in the following sections.

From Symbol to Sense: Applied Linguistics

Natural language produces a staggering amount of ambiguity, already at the levels of word sense, phrase structure and syntactic roles, but most dramatically in the transformation of structure into the meaning that the sentence conveys. Mastering the high degree of potential ambiguity is the main challenge for achieving efficient, and hence, usable NLP technology.

Ambiguities may often be resolved within the sentence scope. For instance, the choice between "lie" for "tell untrue things" and for "take a horizontal position" might be resolved locally with the help of contextual clues. Sometimes though, such ambiguity types cannot be resolved within the sentence scope. Worse, other ambiguity types can never be resolved at that level, such as most pronoun ("him", "it") and common noun ("the company", "the defendant") references. Even if resolution may be achieved elsewhere in the same text, NLP is still in charge: in recent years, the extraction of cross-sentence syntactic and semantic relations has come within reach. Some encouraging results will be presented later in this chapter.

However, even optimal extraction of cross-sentence semantic relations will leave a residue of unresolved ambiguities. Every text contains references that cannot be understood just by reading that text. One text will assume that you know that "Arab Spring" denotes a political uprising, not a climatological season. Another will trust that you grasp the strictly metaphorical use of the term "Titanic". A third presupposes the knowledge that one can see through glasses, but not through wood. Knowledge of these pragmatic relations is required for understanding the full meaning of a text, yet they are for the large part unattainable by software today. Broad-domain ontologies can help in bridging this gap, but much of our common knowledge happens to be stored only in our minds.

Clearly, the road from symbol to sense is long and windy. NLP is steadily moving along, though. It is able to grasp more sense today than was even considered theoretically possible ten years ago. Three independent factors have enabled this development: (1) a revolutionary increase in affordable computational power, thanks to faster processors and the arrival of cloud computing; (2) an equally revolutionary increase in the availability of multilingual data, thanks notably to the emergence of social media on the web; (3) inspired by the former factors, an unprecedented focus on NLP semantics. In the following sections, we will specify the results of this focus, starting with the underlying classical components, moving further to recent developments in automatically trainable multilingual syntactic analysers and finally to their integration and exploitation for the recognition and extraction of domain-specific entities and relations among them. All these technologies have been exploited and used in the different application scenarios of ORDO, for example in the area of innovation monitoring or semantic analysis of web content etc.

What Linguistics Can Do For You

Automatic morphosyntactic analysis has found its way into a host of text-oriented applications. The one component not generally considered a commodity today is the analysis of phrases and syntactic functions. We will address it in some detail at the end of this section, after a brief description of the classical components.

Widely used classical components are *language identification* (determination of the source language of the input text), *tokenisation* (segmentation of a document input into a stream of linguistic units – tokens, words, numbers, punctuation etc.) and *POS-tagging* (selection of the correct part-of-speech using statistical models and/or heuristic rules). *Stemming* is the process of the identification of the stem of a given word form with a morphological analyser and a lexicon lookup. A stem carries the core syntactic and semantic properties of all its word form variations (e.g. "buys", "bought", "buying" have the same core semantics as the stem "to buy"). *Decompounding* segments a word into a sequence of stems; e.g. the German noun "Hybridelektrokraftfahrzeug" ("hybrid electric vehicle") can be decomposed into the following stems: "Hybrid#elektro#kraft#fahrzeug". Compounding is a very productive process and cannot simply be handled by lexical lookup. Furthermore, compounds are notoriously ambiguous, e.g. "Verdichtung#verhältnis" and "Ver#dichtung#verhältnis" are two possible segmentations of the German noun "Verdichtungsverhältnis" ("compression ratio"), but only the first one has an acceptable meaning. Heuristically specified semantic rules are usually needed for identifying the best decomposition.

Almost all of this standard functionality is widely available either as Open Source packages or as professional software development kits, e.g. Attensity's Text Mining Engine (TME), which provides these classical components for 32 languages. The software has already had a big impact in boosting the on-market development in a wide and diverse range of areas of the semantic analysis of unstructured documents like text analytics, information extraction, semantic search, deep-question answering and ontology learning.

Syntactic analysis

As stated in the introduction, the analysis of phrase structure and syntactic functions is a dynamic field in NLP, which is in contrast to the components listed above. Syntactic analysis has only relatively recently started to find its way into commercial NLP application, for the simple reason that it is a computationally heavy operation, precluding acceptable performance of NLP systems until even a few years ago. However, with the fast-growing demand for semantically driven applications as those exploited in ORDO, the status of syntactic annotation changed from unaffordable to indispensable almost overnight. Practically all intelligent semantic analysis requires some level of syntactic annotation for satisfactory accuracy.

A large and widely varying number of theories and formalisms are available for syntactic parsing and function assignment of natural language text. Most of them, however, aim at full coverage, and are consequently quite elaborate, even by today's standards of processing speed. For TME, a more goal-oriented formalism has been chosen. Its syntactic analysis just provides the coverage and level of detail that subsequent semantic components need and no more than that. Currently, TME offers such annotation classes for five languages: English, German, French, Spanish and Simplified Chinese.

These syntactic annotation classes have been developed in a proprietary formalism, called SALSA (Simple and Adaptive Language for Semantic Annotation). It is originally a regular expression language, allowing reference to any and all linguistic annotations already produced. For the purpose of syntactic function assignment and coreference resolution, SALSA has been extended with a global cache mechanism. Details follow further down.

Trainable fast multilingual dependency parsing

In the previous section, we have learned that goal-oriented formalisms for syntactic analysis, like the one used in TME, are promising approaches for reaching the high demands of semantically driven applications. In this context, rule-based approaches are often used where the rules are defined manually. Although it is known that rule-based approaches can achieve high levels of accuracy, they usually lack a wide coverage, i.e. they might not be able to identify enough syntactic variations. Consequently, weak robustness is a critical issue. As an alternative, in recent years a lot of research energy has been brought into the development of robust trainable syntactic parsers that are able to compute the complete syntactic relationships of arbitrary sentences quickly, robustly and accurately

Such trainable parsers receive as input a (usually very large) set of sentences (also called a "*treebank"*) each of which is already (manually or semi-automatically) annotated with the correct syntactic tree. A parsing engine together with a statistical-based learning algorithm is applied to automatically learn a model of all possible syntactic decisions that can be induced from the treebank. This acquired model is then used to determine the syntactic structure of any new sentence. Since the annotation schema is basically the only language-specific parameter, such a trainable parsing system is inherently multilingual, because it can process treebanks of any language.

The syntactic annotation schema of a treebank usually follows a linguistic theory or formalism. However, dependency theory in particular has shown recently to be very suitable for achieving the necessary degree of robustness and efficiency in such a learning environment [1]. The dependency structure of a sentence is a rooted tree (more precisely, a rooted acyclic graph) where the nodes are labelled with the words of the sentence and the directed edges between the nodes are labelled with the grammatical relations that hold between pairs of words. Figure x is an example of this.

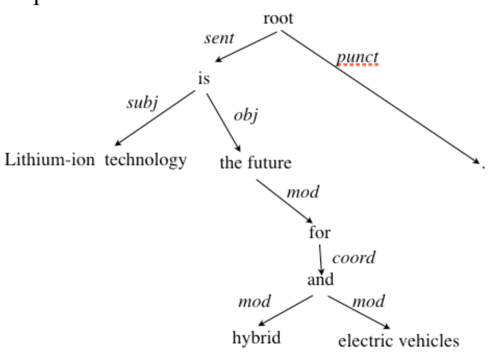


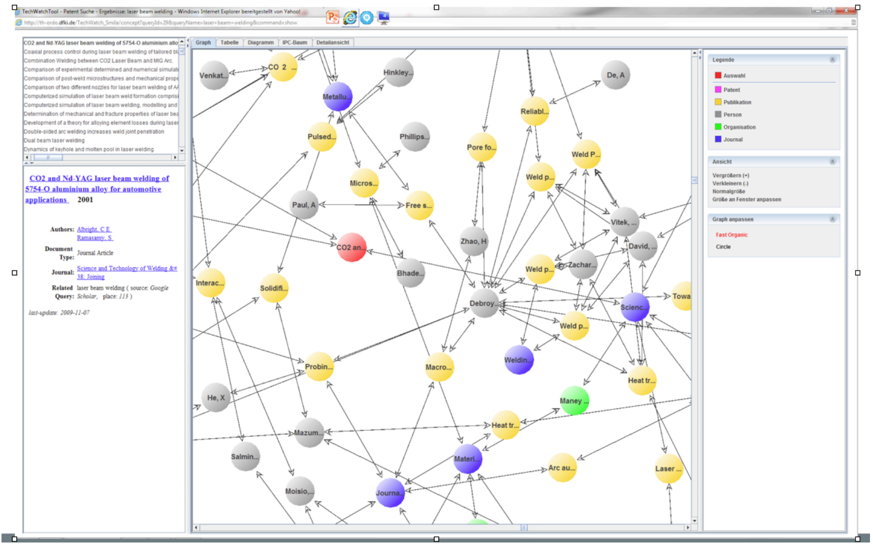
Fig. X: Simplified dependency tree for sentence "Lithium-ion technology is the future for hybrid and electric vehicles." Edges are decorated with labels for grammatical functions.

Dependency structures are appealing because they already represent a "shallow" semantic relationship. They are a suited data structure for many semantic applications, e.g. semantic search and relation extraction.

As part of ORDO, DFKI has developed the MDParser, which is a very fast multilingual trainable dependency parser, which also exists as a SMILA component [2]. MDParser has been adapted to a new highly efficient linear multiclass classifier to obtain high speeds. It is now able to process up to 50,000 tokens (~2,000 sentences) per second. For the traditional English WSJ test data, MDParser scores 89.7 % UAS (Unlabelled Accuracy Score) and 87.7% LAS (Labelled Accuracy Score). This performance is comparable with other state-of-the-art parsers; it is about three to five times faster than the widely used and previously fastest known dependency parser MaltParser [3].

The MDParser has been exploited successfully in a number of different semantic applications, such as recognising textual entailment [4] and trend analysis [5]. As an example for the latter case, we developed a software demonstrator called "TechWatchTool" using SMILA. It aids companies in detecting emergent technologies in a particular field and in identifying associated key players and their cooperative networks. It currently supports three scenarios: 1) retrieval of patents and publications; 2) ontological presentation of a knowledge domain; and 3) identification of new trends in relevant documents.

The system has been developed in collaboration with ThyssenKrupp Steel AG and combines various methods used in bibliometrics, information extraction and knowledge technologies. It integrates advanced NLP for detecting and extracting relevant facts from unstructured documents. Especially for relation extraction, the MDParser is used as a basis for the extraction of relational features. TechWatchTool provides personal and group-level access via a browser application and an interactive graphic interface, as shown in Figure x.



TechWatchTool - Patent Suche - Ergebnisse: laser beam welding - Windows Internet Explorer bereitgestellt von Yahoo!

http://th-ordo.dfki.de/TechWatch_Smila/concept?queryId=29&queryName=laser+beam+welding&command=show

CO2 and Nd-YAG laser beam welding of 5754-O aluminium allo
Coaxial process control during laser beam welding of tailored bl
Combination Welding between CO2 Laser Beam and MIG Arc.
Comparison of experimental determined and numerical simulat
Comparison of post-weld microstructures and mechanical prope
Comparison of two different nozzles for laser beam welding of A
Computerized simulation of laser beam weld formation compris
Computerized simulation of laser beam welding, modelling and
Determination of mechanical and fracture properties of laser bea
Development of a theory for alloying element losses during laser
Double-sided arc welding increases weld joint penetration
Dual beam laser welding
Dynamics of keyhole and molten pool in laser welding

CO2 and Nd-YAG laser beam welding of 5754-O aluminium alloy for automotive applications 2001

Authors: Albright, C.E. Ramasamy, S.
Document Type: Journal Article
Journal: Science and Technology of Welding &#38; Joining
Related Query: laser beam welding (source: Google Scholar, place: 113)

last-update: 2009-11-07

Fig. X: Screenshot of TechWatchTool – retrieval and navigation in patents and publications.

The extraction of named entities and semantic relations

A central annotation task is the extraction of meaningful semantic information from text. Important subtasks are the extraction of named entities (names of persons, locations and organisations) as well as the extraction of semantic relations between these entities. For example, the sentence "Metro-Chef Eckhard Cordes hat von der Konzerntochter Real im März 2008 gefordert, eine Umsatzrendite von drei Prozent zu erwirtschaften." We have the person "Eckhard Cordes", the companies "Metro" and "Real", the date "März 2008" and the percentage "drei Prozent". As nouns in German are in general capitalised the detection of names is much more difficult than in English. In addition, names often contain words which also have other meanings. Therefore, we require a context-sensitive detection algorithm. We used conditional random fields [6] to detect the name phrases of a sentence by taking into account the roles and attributes of consecutive words in a sentence. A major challenge for their application is the extraction of relevant input features for the German language.

Often extracted entities have several possible meanings. In Wikipedia the term "Metro", for instance, can be a rapid transport system, an airport, a city, an administration, a newspaper, a retailer, a theatre, a film, etc. To be able to interpret an entity we have to disambiguate a mention of an entity using its context. We have developed an advanced disambiguation system [7] based on latent representations, which is able to find the correct meaning of an entity with high reliability and assign it to its corresponding article in Wikipedia. The system also detects if an entity is not covered in Wikipedia. The approach is language independent, as it works for the English, German and French Wikipedia.

Relation detection requires that two named entities are selected as arguments. In the above example, for instance, we have the relations "Eckhard Cordes" is-leader-of "Metro" and "Real" is-subsidiary-of "Metro". To detect such relations it is no longer sufficient to exploit the sequence of words but we need the structure of a sentence as described by a parser. In THESEUS we evaluated dependency parsers and phrase grammar parsers and found that a combination of both structures gives the best results. Using annotated training examples, we trained a classifier using these structures, which predicts whether the target relation holds between these arguments [8]. For the member-of relation, for instance, we achieved an F-value of more than 80 %.

As an application within ORDO, Fraunhofer IAIS has developed a tool for Commerzbank, the second largest bank in Germany, which has its headquarters in Frankfurt. Similar to many other large banks, Commerzbank has only a few administrative centres that handle the credit business with medium and small enterprises. Although this structure is efficient, the centres often lack sufficient knowledge about the local markets and enterprises within small regions. The tool extracts information about the economic situation of markets and enterprises from local sources.

An important indicator of the economic situation of a company is the economic success of its business partners or its competitors. This posed the task of extracting and monitoring the relationships between companies from local news, especially online pages of local newspapers. To allow screening of large text repositories and the regular monitoring of relations, machine learning approaches from text mining were used. The figure below shows a network of relations between businesses extracted for the region of Bonn from the *General-Anzeiger* (the regional newspaper).

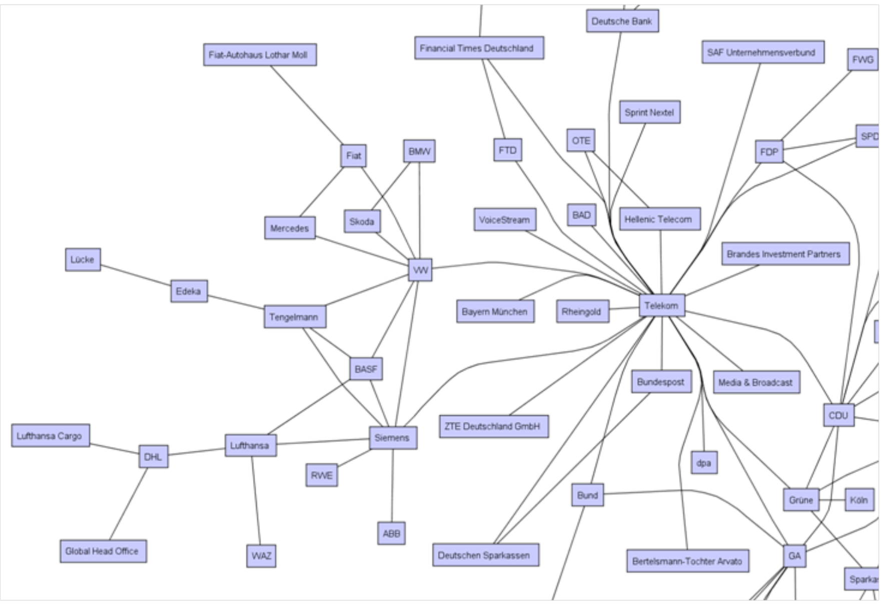


Fig. X: Company relations extracted from the *General-Anzeiger* in Bonn.

After downloading the web pages and cleaning the html pages, two text mining tools were employed. By using named entity recognition models, the names of companies were extracted from the text (2,209 companies for Bonn). Subsequently a relation extractor was trained and applied to extract relations between companies, e.g. company A does business with company B, A is a competitor of B, or A acquires B. As an input for the relation extractor, the text of sentences is enhanced by many features that are indicators of a relation. As an example Figure X shows a parse tree of a sentence stating the acquisition of a firm.

Different types of information relevant for the bankers could be extracted by using the detected relations:

- Slump in sales or net loss for the year for an enterprise. This also indicates possible problems for business partners and competitors.
- Cooperation between companies, acquisition of companies.
- Increase in turnover, growth of staff.
- Good order situation in the building sector.

If critical developments are detected this information is given to the representatives of the bank who take appropriate measures. Currently the extension of the system to include more relationships as well as its application to a larger number of news sources is being prepared.

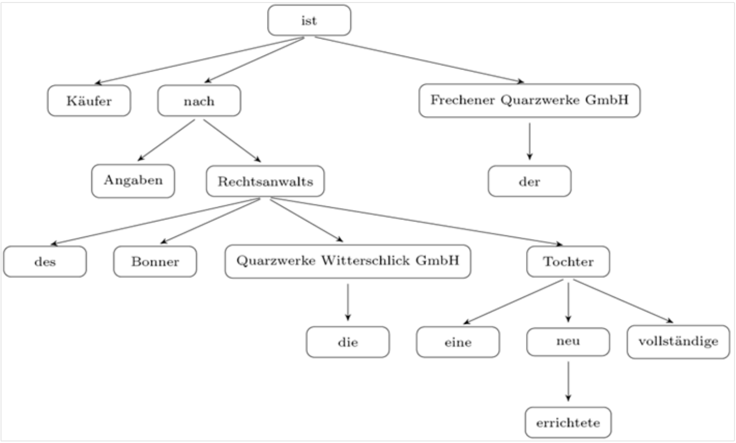


Fig. X: Parse tree of a sentence concerning a company acquisition.

Cross-sentence analysis

As mentioned in the introduction, one of the more difficult tasks in text mining is the resolution of ambiguity outside the sentence scope. A typical example is pronominal reference, which usually crosses sentence boundaries, as in: "George1 sent Martha2 flowers. He1 wanted to make up with her2." It is obvious that the level of understanding of a text is enhanced by the proper resolution of pronouns like "He" and "her" in the example above.

First and foremost, this applies to the resolution of person names. Standard named entity extraction operating at the sentence level is able to identify only proper names as person referents, not pronouns. It is not able either to link a variant of a person's name (called alias) to its full (or canonical) referent. Consequently, standard methods fail to generalise mentions such as "George Burton", "Mr Burton", "George", "he" and "him" by linking them to the same canonical referent, "George Burton".

Issues like these can be solved by the introduction of a cache mechanism that stores likely candidates with their attribute sets for any type of disambiguation resolution. Such a cache has been developed for TME. Feature matching and stem similarity are the predominant heuristics used by the cache.

In addition to named entity extraction, TME deploys the SALSA cache for certain syntactic function annotations, and for semantic applications such as sentiment analysis and resume annotation.

A functionality called "Dossier Creation" is currently under development, which heavily relies on the cache mechanism. Its general idea is to create automatically qualified structured information about a certain coherent class of objects from unstructured or semi-structured data sources. Dossier Creation promises a large variety of applications, from targeted data mining for research (intelligence, finance, patents) to normalisation of large chaotic – i.e. unstructured – data repositories (HR, medical, technical, legal).

Summary

This chapter has outlined the needs and challenges for extracting semantics from unstructured documents from a linguistic perspective. All of the described language technology has been exploited and realised in ORDO, where some of the most important results and their embedding into innovative applications have been described in more detail. These results allow us to draw valuable conclusions from otherwise unstructured texts. Linguistics provides the basis for dealing with the Big Data challenge on the content level.

From an application perspective, future work will further develop the creation of dossiers. Here, it might also be interesting to take into account cross-lingual phenomena since several of the above-mentioned linguistic components can already process several natural languages. Another promising line of R&D is the new field of "textual inference", a sort of applied textual semantics. A sub-area of this new field, "recognising textual entailment", has already been explored successfully in ORDO with promising results.

Beyond the mere research, ORDO has demonstrated that text analysis with linguistics is a robust technology. Its integration in day-to-day business processes creates added value and facilitates the automation of tasks, which often still require human intervention.

References

1. Kübler, S., McDonald, R., Nivre, J.: Dependency Parsing. Morgan & Claypool.
2. Volokh, A., Neumann, G.: Automatic Detection and Correction of Errors in Dependency Treebanks. 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL) (2011).
3. Hall, J., Nilsson, J., Nivre, J.: MaltParser - a data-driven dependency parser, http://maltparser.org/index.html.
4. Volokh, A., Neumann, G.: Using MT-Based Metrics for RTE. Fourth Text Analysis Conference , Gaithersburg, MD, USA. (2011).

5. Hong, L., Xu, F., Uszkoreit, H.: TechWatchTool: Innovation and Trend Monitoring. International Conference on Recent Advances in Natural Language Processing, Hissar, Bulgaria (2011).
6. Sutton, C., McCallum, A.: An introduction to conditional random fields for relational learning. Introduction to statistical relational learning. MIT Press (2006).
7. Pilz, A., Paaß, G.: From names to entities using thematic context distance. Proceedings of the 20th ACM international conference on Information and knowledge management. S. 857–866 (2011).
8. Reichartz, F., Korte, H., Paaß, G.: Semantic relation extraction with kernels over typed dependency trees. Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. S. 773–782 (2010).