Incorporating ontological background knowledge into Information Extraction

Benjamin Adrian

Knowledge Management Department, DFKI, Kaiserslautern, Germany
benjamin.adrian@dfki.de

Abstract. In my PhD work I apply formal domain knowledge to support domain specific Information Extraction tasks. My main research goal is revealing strategies incorporating domain ontologies for: (i) Interchanging domain ontologies by letting systems adapt on new domains without any additional engineering effort. (ii) Allowing extraction templates to be specified in the ontology's vocabulary with the canonical query language SPARQL. (iii) Improving Ontology-based Information Extraction approaches by making extraction pipelines access existing knowledge in the earliest possible stages. (iv) Returning potential query results in RDF (graphs of facts and instances) as scenarios weighted with textual and ontological evidences. (v) Improving methods using instance-knowledge to train statistical models for extraction. In summary, my PhD thesis' contribution is a system letting users load up ontologies about their domains of interest, query domain relevant text with SPARQL and get results as weighted RDF graphs.

1 Introduction

Whereas, the vast majority of information in WWW is written in unstructured text, the Semantic Web extends this and makes WWW data machine understandable by using formal knowledge in form of semantic annotations about WWW data. Existing ontologies formalize these annotations allowing reasoning and deduction about known knowledge. In order to use Semantic Web applications, these annotations have to be created. A promising approach for this is applying Information Extraction (IE) techniques to extracting semantic annotations about unstructured text. Intended use cases are e.g., instance recognition, fact extraction, or scenario extraction combining both. In general, two major IE disciplines exist:

- Knowledge-engineering approaches focus on hand crafted extraction rules (e.g., grammars, regular expressions) about language patterns. Thus, they are unusable in evolving domains, as maintenance is too expensive.
- Machine-learning based approaches depend on training extraction models with example data. But, these training examples are expensive to build.

As result, traditional IE systems either tend to provide rich functionality for single domains (e.g., gene extraction) or they provide a small set of functionalities

for domain independent information (e.g., Named Entity Recognition). When applying IE systems to creating semantic annotations in the Semantic Web use case, these IE systems have to adapt to the content of domain ontologies. The types of annotations generated should also be configurable by using template mechanisms defined similar way to querying Semantic Web ontologies.

The claim of my PhD thesis is letting users customize their information extraction system by loading up RDFS ontologies about domains of interest, process text with SPARQL templates and gain results as weighted RDF graphs.

This work has been done in project Nepomuk[1] where I generated tag and fact recommendations for text and will proceed in Perspecting[2] where I will explore generating extraction scenarios and supporting IE templates in search processes. This thesis supports the topic of Ontology-based Information Extraction (OBIE) as base for Semantic Web applications grounding on semantic annotations.

The structure of my PhD proposal is as follows: At first, related work about IE is given by mentioning state-of-the-art taken and research activities that compete with mine in at least single topics. Next, main research goals are summarized. These goals are organized in work packages listed in the following section, giving details for each with its progress state and existing publications. Finally, I summarize this PhD proposal and comment on its role for Grishman's inspiring long term IE goal.

2 Related Work

This thesis applies well established IE methodologies i.e., IE templates and a pipeline system architecture consisting of finite state transducers. Three publications form the foundation of my research i.e., Embley's first application of ontologies to IE [1], the template design principles [2] by Hobbs and Israel describing template design as ontology engineering, and finally the use case by Sintek et. al. [3] for populating domain ontologies with IE results. An IE system taken as reference is ANNIE based on the GATE framework [4]. Prototypes developed in my thesis comply with ANNIE according to techniques such as finite state transducer, gazetteers, or writing extraction rules with regular expressions. I also build upon the work done by Dayne Freitag who applied machine learning to IE [5], training models for POS tagging, noun phrase chunking, or instance disambiguation. McCallums MALLET[3] framework is used as base for statistical models such as Naive Bayes or Conditional Random Fields. Performing comparative evaluations between OBIE systems is still a problem. The community has not yet agreed on a standard corpus or at least a common evaluation methodology. Fortunately, Maynard analyzed metrics, accounting hierarchies in ontologies [6] that will be used in this thesis' evaluation. She also proposed benchmarking hints for annotation tools [7] that are being considered. My thesis focus on

[1] Project NEPOMUK, grant FP6-027705, `http://nepomuk.semanticdesktop.org`

[2] Project Perspecting, grant 01IW08002, `http://www.dfki.uni-kl.de/perspecting`

[3] McCallum, Andrew Kachites. "MALLET: A Machine Learning for Language Toolkit." `http://mallet.cs.umass.edu`. 2002.

Ontology-based Information Extraction first mentioned by Maedche, Neumann and Staab [8] and later ellaborated on during the SEKT project[4].

Comparable OBIE systems are S-Cream [9], SOBA [10], or an early bootstrapping approach [8]. These approaches extend standard IE systems and populate domain ontologies or generate annotations with extraction results. In consequence they deal with the problem of aligning plain extraction results such as Named Entities to ontological instances. Solutions cover mapping strategies requiring costly post-processing efforts in disambiguation or discourse analysis. To avoid this overhead, my approach integrates ontologies into extraction pipelines as early as possible. As result, ontologies can be preprocessed statically once before starting further extraction steps. This relates the OntoRootGazetteer, a GATE plugin, analyzing existing concept labels with tokenizers, POS taggers and stemmers. GATE has also been extended with so called OntoGazetteers for mapping gazetteer lists to ontology classes [11], manually. I will extend this, by classifying gazetteers as potential datatype properties automatically.

Labsky et al. use specialized ontologies for extraction purpose [12]. These ontologies specify what and how to extract from text. In contrast, my approach focus reuse of existing ontologies, e.g., from Open Linked Data, without any changes.

Regarding machine learning and ontologies, Li and Bontcheva already showed the use of ontologies' instance knowledge for training machine learning models [13], effectively. As this strategy is promising, I will apply it on several problems (e.g., noun phrase chunking, instance recognition and resolution, instance disambiguation, fact extraction, etc.).

In addition to these related works, my thesis will cover using SPARQL as ontology based extraction templates.

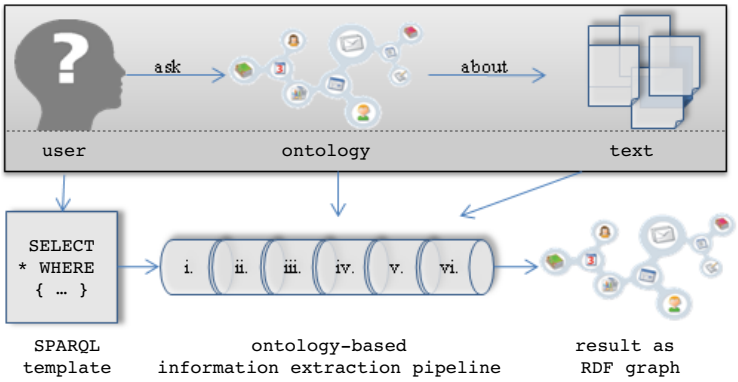

Fig. 1. Scenario performing Ontology-based Information Extraction

[4] please refer to: `http://www.sekt-project.com/rd/deliverables/wp02`

3 The Approach

In general, my PhD work reveals strategies on how and when to incorporate ontological knowledge into a holistic OBIE scenario. Figure 1 covers this overall OBIE scenario of my approach.

3.1 General Scenario

By using an existing ontology and documents about a certain domain (e.g., music, US politics, etc.), users specify their information demand in a SPARQL query by using the ontology's vocabulary. For example, demanding information from news stories about occurring politicians, their names, age, and electoral district, and related parties with names might be specified as template in SPARQL: *SELECT * WHERE {?politician a foaf:Person; foaf:name ?name; foaf:age ?age; pol:district ?district; pol:related ?party. ?party skos:prefLabel ?partyName. ?district skos:prefLabel ?districtLabel.}* The OBIE pipeline uses the ontology, existing instance knowledge, and the template for extracting this information from text. Finally, it generates potential results as weighted scenarios in RDF graphs. Please consider that instance knowledge may even entail a generated scenario with facts that extend the amount of information of the underlying text.

By changing the domain ontology, the user is allowed to "ask" different questions covering other domains without any implementation or rule engineering efforts. It is just expected that users pass parameters about which classes, object and datatype properties to be considered by the OBIE system.

3.2 Preprocessing the ontology

In a preprocessing step the OBIE system analyses the ontology, existing instances, their datatype properties, and object properties to other instances. Values of datatype properties are then converted to efficient data structures (e.g., B*-Trees, Suffix Arrays). Instances' properties are represented in adjacency lists stored as bitmaps. Additional statistics are computed about property value cardinalities and instances' inter-relations. It is also possible to pass additional documents, gazetteers or regular expressions about a domain. Gazetteers values and regular expressions are analyzed, whether they match with certain datatype property values. Documents are used for learning phrase patterns of matching datatype property values.

3.3 OBIE Pipeline

After finishing preprocessing once, the OBIE pipeline is ready for extracting domain related information from text. This comprises six major process steps covering necessary IE tasks. Each task generates a set of hypotheses weighted with confidence values. Confidences are combined by using Dempster-Shafer's belief function. Pipeline and hypotheses are formalized in a process ontology[5].

[5] The pipeline formalization is described at: `http://ontologies.opendfki.de/repos/ontologies/obie/annotation`

Normalization At first, Normalization extracts plain text and existing metadata (e.g., title, author) from documents. Next, it identifies their language.

Segmentation Here, incoming text is partitioned into units of tokens and sentences. Segmentation performs POS tagging also.

Symbolization This step recognizes matches (called symbols) between phrases in text and values of datatype properties. Symbolisation also performs named entity recognition with given gazetteers linked to datatype properties. It performs structured entity recognition given regular expressions linked to datatype properties. Finally, it performs noun phrase chunking.

Instantiation The Instantiation step resolves recognized symbols with candidates for possible instances. Instantiation also disambiguates these instance candidates with existing inter-relation statistics. It recognizes object property candidates in sentences between recognized instances.

Contextualization In Contextualization, recognized instances, recognized object properties, and existing fact knowledge is resolved for creating fact candidates that are valid for generating scenarios for a given template.

Population Extracted fact candidates populate multiple variants of extraction templates called scenarios. Each scenario is weighted with a confidence value.

Resulting scenarios are used as annotations and/or populate ontologies with new facts. Performing the final population step can be set semi-automatically resulting in a recommendation system, or automatically resulting in an automatic annotation system. The latter case requires thresholds for confidence values and the assumption that high confidence values promise high precision values.

4 Progress Plan with Current State and Future Work

My agenda for reaching the goals defined above is organized in a list of work packages. In order to have a demo prototype for presenting my main ideas, these packages are often processed in parallel and thus are not finished in sequence. The following list explains each packages giving details about its current state of work with references to existing publications.

WP1. Feasibility study and prototype: A feasibility study between 2007 and 2008 resulted in a research prototype [14] that covered the OBIE scenario and provided at least basic functionalities or mocks for each of the following work packages. It is called iDocument[6] and was developed based on GATE. It serves demonstration purposes and was presented at the CEBIT 2008 exposition in Hanover, Germany.

WP2. Interchangeable domain ontologies: This package deals with interchanging RDFS ontologies and implements preprocessing algorithms and efficient data structures. Relevant classes and properties for extraction purpose are passed as parameters[7]. This package is finished.

[6] `http://idocument.opendfki.de`

[7] These parameters are formalized in an ontology called Matadata for ontology-based Information Extraction (MOBIE): `http://ontologies.opendfki.de/repos/ontologies/obie/mobie`

WP3. Flexible IE templates: Users should create templates easily, by writing them as SPARQL queries according to the current domain ontology. Using SPARQL as template definition language is implemented prototypically. Currently, the template engine only supports extraction of relations between existing instances. Datatype properties or instances as such cannot be given in template expressions. These topics will be implemented in near future.

WP4. OBIE pipeline architecture: The OBIE pipeline is designed to use preprocessed ontology knowledge in earliest possible stages. The pipeline deals on the one hand with certain knowledge from ontologies and on the other hand with uncertain knowledge extracted from text. For coping with this, Believing Finite State Cascades were developed, combining evidences of hypotheses with Dempster-Shafer's belief functions [15]. This architecture is stable, but is going to be extended for estimating good thresholds and moderation parameters automatically by using statistics about the ontology and current hypotheses' belief distributions. An additional goal is calibrating the system with settings for generating just recommendations with focus on recall or automatically annotating text with focus on high precision values about results.

WP5. Adapting instance knowledge for IE purpose: Existing instance and rule based knowledge (i.e., datatype property values, gazetteer entries, or regular expressions about datatype property values allowed) is used for identifying instance candidates in text. Training extractors just with instance knowledge, leads to problems in identifying labels in text passages as the surrounding text context is unknown. Thus, domain related documents are added for training context sensitive extractors automatically. After preprocessing these documents with Normalization, Segmentation, and parts of Symbolization tasks, matches of datatype property values, gazetteer entries, or regular expressions are used for training.(e.g., domain and language specific conditional random fields for noun phrase chunking). This work package is the scope of my current activities.

WP6. Instance disambiguation and discourse analysis: The identity of recognized instance candidates may be ambiguous. Therefore, analyzing object properties between instances provides metrics for clustering extracted instance candidates in single discourse sets. My current approach uses a Naive Bayes Model that is trained with relations between instances. As result, it ranks a list of ambiguous instances with probabilities. In addition, it even recommends instances that were not recognized in text but are relevant for the current discourse set. This has to be evaluated.

WP7. Populating templates and instance base: Based on the sum of hypotheses generated along the OBIE pipeline, potential scenarios have to be resolved for the given template. In contrast to traditional query evaluation techniques in databases, scenarios may be incomplete or faulty. Thus, they are weighted with confidence values. The population of templates in iDocument for generating scenarios was implemented with a rudimentary graph algorithm [16]. Future work will evaluate this topic by using hypotheses as features and facts as categories in a classification scenario.

Evaluation: In order to prove adaptability, the approaches are evaluated in multiple domains. I already created a corpus about the Olympic summer games 2004 consisting of a domain ontology, instance base, and annotated news articles [17]. In general, when using standard IE corpora, it is necessary to extract domain ontologies with ground truth data. This was done with data of the Pascal Challenge evaluating Machine Learning for IE[8]. It is also planned to enrich other corpus data with domain ontologies for using it in OBIE evaluations (e.g., CoNLL data). Other data sets from Semantic Web sources such as Linked Open Data (e.g., DBPedia) are also planned to be used.

Applications: As this PhD proposal is settled in an application oriented research center (DFKI), it is necessary to show its impacts in a set of applications. The Semantic Desktop has been seen as ideal application for extending with OBIE functionalities. By using the iDocument prototype, a document classification based on instance recognition was implemented in Nepomuk. Comparing results from iDocument and StrucRec, another classification component based on IBM's Galaxy framework, figured out slightly better classification recommendations from iDocument [16].

In another study [18], iDocument was used to create semantic annotations about OCRed documents. Finally, text and annotations were transformed into a semantic wiki article.

During the project Perspecting, iDocument will be implemented as OBIE service. An existing Semantic Wiki (called Kaukolu[9]) will be integrated and used for manually correcting generated annotations about documents. This ground truth data is suitable for retraining existing extraction models. Other small applications are planned to be implemented (e.g., profile tagging on social platforms).

5 Summary

The proposed PhD thesis examines how ontologies and instance knowledge support Information Extraction tasks. Its main contribution is a domain adaptive Information Extraction system. Interchangeable domain ontologies afford the application to multiple domains. Extraction templates may be defined easily in SPARQL. Using existing background knowledge enhances quality of extraction results. My final goal is to narrow the gap between current Information Extraction and the inspiring statement by Grishman (2002): *Our long-term goal is to build systems that automatically find the information you're looking for, pick out the most useful bits, and present it in your preferred language, at the right level of detail.*

This work was financed by the BMBF project Perspecting (Grant 01IW08002).

[8] see `http://nlp.shef.ac.uk/pascal/Corpus.html`
[9] see project page of Kaukolu: `http://kaukolu.opendfki.de`

References

1. Embley, D.W., Campbell, D.M., Smith, R.D., Liddle, S.W.: Ontology-based extraction and structuring of information from data-rich unstructured documents. In: Proc. of CIKM '98, New York, NY, USA, ACM (1998) 52–59
2. Hobbs, J., Israel, D.: Principles of template design. In: Proc. of the Human Language Technology Workshop, Morgan Kaufmann (1994) 172–176
3. Sintek, M., Junker, M., van Elst, L., Abecker, A.: Using information extraction rules for extending domain ontologies. In: IJCAI'2001 Working Notes of the Workshop on Ontology Learning, Seattle, Washington. Volume CEUR-WS 38. (2001)
4. Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V.: GATE: A framework and graphical development environment for robust NLP tools and applications. In: Proc. of the 40th Anniversary Meeting of the ACL. (2002)
5. Freitag, D.: Machine learning for information extraction in informal domains. Machine Learning **39**(2/3) (2000) 169–202
6. Diana Maynard, W.P., Li, Y.: Evaluating evaluation metrics for ontology-based applications: Infinite reflection. In: Proc. of LREC'08. (2008)
7. Maynard, D.: Benchmarking textual annotation tools for the semantic web. In: Proc. of LREC'08. (2008)
8. Maedche, A., Neumann, G., Staab, S.: Bootstrapping an ontology-based information extraction system. In: Studies in Fuzziness and Soft Computing, Intelligent Exploration of the Web, Springer (2002)
9. Handschuh, S., Staab, S., Ciravegna, F.: S-CREAM - Semi-automatic CREAtion of Metadata. In: Proc. of EKAW '02, London, UK, Springer-Verlag (2002) 358–372
10. Buitelaar, P., Cimiano, P., Racioppa, S., Siegel, M.: Ontology-based information extraction with soba. In: Proc. of LREC. Genoa, Italy. (2006)
11. Bontcheva, K., Tablan, V., Maynard, D., Cunningham, H.: Evolving GATE to Meet New Challenges in Language Engineering. Natural Language Engineering **10**(3/4) (2004) 349—373
12. Labsky, M., Svatek, V., Nekvasil, M.: Information Extraction Based on Extraction Ontologies: Design, Deployment and Evaluation. In Adrian, B., Neumann, G., Troussov, A., Popov, B., eds.: Proc. of OBIES 2008. Volume CEUR-WS 400. (2008)
13. Li, Y., Bontcheva, K.: Hierarchical, perceptron-like learning for ontology-based information extraction. In: Proc. of WWW '07:, ACM (2007) 777–786
14. Adrian, B., Maus, H., Dengel, A.: iDocument: Using Ontologies for Extracting Information from Text. 5th Conf. on Professional Knowledge Management (2009)
15. Adrian, B., Dengel, A.: Believing Finite-State cascades in Knowledge-based Information Extraction. In: Proc. of 31st KI 2008. Volume 31. (2008)
16. Adrian, B., Klinkigt, M., Maus, H., Dengel, A.: Using iDocument for Document Categorization in Nepomuk Soical Semantic Desktop. In Pellegrini, T., ed.: Proc. of i-Semantics 2009. (2009)
17. Grothkast, A., Adrian, B., Schumacher, K., Dengel, A.: OCAS: Ontology-Based Corpus and Annotation Scheme. In: Proc. of HLIE 2008. (2008)
18. Adrian, B., Maus, H., Kiesel, M., Dengel, A.: Towards Ontology-based Information Extraction and Annotation of Paper Documents for Personalized Knowledge Acquisition. In: 5th Conference on Professional Knowledge Management. LNI (3 2009)