The Semantic Gap of Formalized Meaning

Sebastian Hellmann

AKSW Research Group, http://aksw.org, Universität Leipzig, Germany
hellmann@informatik.uni-leipzig.de

Abstract. Recent work in Ontology learning and Text mining has mainly focused on engineering methods to solve practical problem. In this thesis, we investigate methods that can substantially improve a wide range of existing approaches by minimizing the underlying problem: The Semantic Gap between formalized meaning and human cognition. We deploy OWL as a Meaning Representation Language and create a unified model, which combines existing NLP methods with Linguistic knowledge and aggregates disambiguated background knowledge from the Web of Data. The presented methodology here allows to study and evaluate the capabilities of such aggregated knowledge to improve the efficiency of methods in NLP and Ontology learning.

1 Problem

Decades of research have been spent on answering the question "How can we teach machines to understand natural language?" and the unanimous response is: It is impossible (for obvious reasons). Any approach that indulges in formalizing the meaning of natural language faces the same problem: the Semantic Gap between formalized meaning and human cognition. So instead, we should modify the question and ask: "How big is the current Semantic Gap?", "How can we reduce it?" and "How can we measure the reduction?". We argue that, if we choose OWL as a unifying Meaning Representation Language (MRL), we are in a position to utilize several advantages (interoperability, expressiveness, available linguistic ontologies, available structured knowledge from the Web of Data, mature tool support) not only to reduce the Semantic Gap, but also to define processes for potential reductions, combined with an evaluation methodology.

2 State of the Art

Currently, work towards standardization of linguistic annotations is in progress (see work by Nancy Ide). Wintner (2009)[8] argues that interest in grounding NLP in Linguistic theories has been declining for two decades with the focus now being on "engineering solutions to practical problems" and corpora "as our source of (implicit) knowledge". A claim that can be verified by looking, for example, at methods included in two recent surveys of Ontology learning from text by Buitelaar et al.(2005)[1] and Zhou (2007)[9]. Parallel to these trends, a

large number of mature linguistic ontologies has emerged formalizing knowledge about linguistic features. A representative example are the Ontologies of Linguistic Annotations ([2], OLiA), an architecture of modular OWL-DL ontologies that formalizes several intermediate steps of the mapping between concrete annotations (such as POS tags) and a Reference Model. While such ontologies exist, the question arises in which ways they can be exploited besides their intended purpose of unifying annotation schemes. One key to knowledge acquisition is clearly a deep linguistic analysis: by using OWL to aggregate knowledge on all levels of Linguistics in a single model and in a formal way, we are able to create a powerful preprocessing step, which can potentially improve a range of current methods.

The advantages become obvious, if we look, for example, at a recent approach by Völker et al. [7], which applies transformational rules to acquire OWL axioms from text. Völker et al. have a three-page long discussion about necessary linguistic features to improve their approach. As an enhancement, the conditional part of these proposed rules can, however, be directly replaced by expressive reasoner queries when employing our approach.

In his keynote talk at the ISWC 2009[1], Tom Mitchell presented an approach to extract facts from the Web with the help of a bootstrap ontology. He mentioned several shortcomings such as missing morphological features (why not more?) and incorporation of existing background knowledge (such as DBpedia[2]).

In this thesis, we will investigate the above-mentioned limitations as a whole, i.e. not just engineering solutions to specific tasks. By combining existing NLP methods with linguistic knowledge in OWL and adding disambiguated background knowledge from the Web of Data, we argue that we can shed light on the underlying problem of knowledge acquisition: The Semantic Gap of Formalized Meaning. Such an integrated approach has several advantages: errors in one layer can be corrected by other layers and external knowledge and also new OWL axioms can be induced based on given positive and negative sentences (see Section 5, where we learned the definition for a passive sentence.)

3 Proposed Approach

We begin with converting natural language text (a character sequence with implicit knowledge) into a more expressive formalism, in this case OWL, to grasp the underlying meaning. This explicated meaning then serves as input for (high-level) algorithms and applications (with a focus on machine learning). The central working thesis is the following:
If features extracted by different NLP approaches (ranging from low-level morphology analysis to higher-level anaphora resolution) are explicated and combined with matching background knowledge (parser-ontology pair) in a model and if, additionally, this model is further enriched by fragments of existing knowledge

[1] `http://videolectures.net/iswc09_mitchell_ptsw`
[2] `http://dbpedia.org`

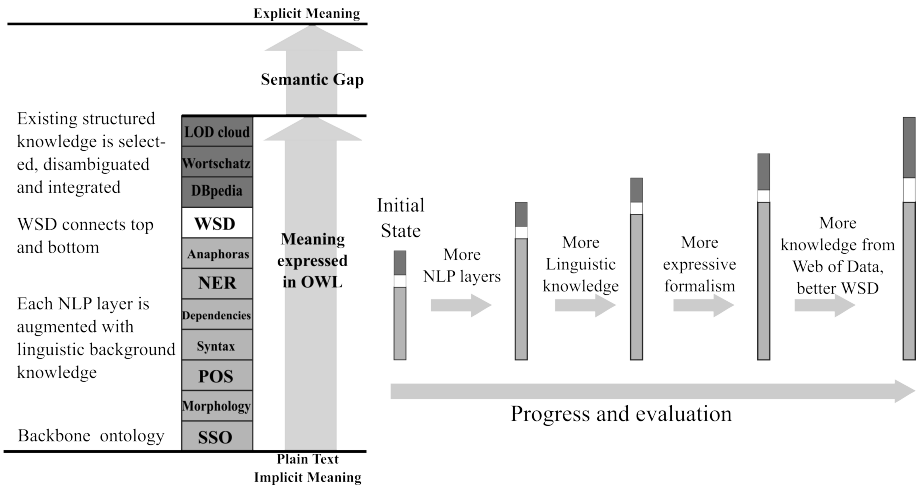

Fig. 1. Left: Display of the integration of NLP approaches and background knowledge connected via a Word Sense Disambiguation (WSD) component. Right: Actions which potentially lead to a measurable increase of the stack (cf. Section 4 for evaluation)

bases from external sources such as DBpedia, it will be possible to reduce the Semantic Gap and improve performance on common knowledge acquisition tasks such as Ontology learning and Text understanding.

Figure 1 gives an overview of the conversion approach in form of a stack (on the left side). In a first step, sentences are tokenized and aggregated in a *Structured Sentence ontology (SSO)*, consisting of a minimal vocabulary to denote the basic structure of the sentence such as tokens and relative position of a token in a sentence. The SSO (bottom) serves as the backbone model, which will be augmented by (1) features from NLP approaches (in light gray), (2) rich linguistic ontologies for these features, (3) background knowledge from the Web of Data (in dark gray) and, finally (4) knowledge which can be derived or inferred and which improves and corrects steps 1-3.

In most cases, output generated by NLP methods can be modeled in RDF in a straightforward way (e.g. POS-Tags are connected to tokens, dependencies are connections between tokens). An increasing number of linguistic ontologies already exist for certain NLP tasks (see [2] and Section 5) and serve as valuable addition to parser output (forming a *parser-ontology pair*). Both types of information aggregated from several parsers and integrated into the backbone model represent the basis for selecting background knowledge from the Web of Data. Fragments[3] of DBpedia, for example, can be retrieved on the basis of disambiguated entities as seed nodes and added to the model.

Preliminarily, we will define the Semantic Gap in a negative way, i.e. meaning that is not covered by the current stack. The stack in Figure 1 can be increased

by the four measures depicted on the right side. We created a reference implementation NLP2RDF[3], which outputs an aggregated OWL ontology.

4 Methodology

Based on the formulation of the working thesis, progress will be driven by rigid evaluation. Although, the reduction of the Semantic Gap can not be conceived directly, the growth of the stack can be measured very well. We will collect a list of suitable machine learning approaches and knowledge acquisition tasks which normally depend on some form of NLP methods for preprocessing. Then, we will use the DL-Learner[5], a supervised concept learner, to repeat experiments and compare existing methods. As the expressiveness and availability of background knowledge directly affects the learning capability, we deem it an ideal evaluation setting because we can study the influence of available knowledge on the performance of certain tasks (e.g. relation extraction). In this way a benchmark suite can be created in which we can vary parsers or deactivate the inclusion of ontologies. This allows to evaluate the conditions under which the aggregated model was created and can thus measure potential improvements of the NLP2RDF stack in Figure 1 (right side).

Initially, we plan to participate in SemEval 2010[4] Tasks 4 (VP Ellipsis - Detection and Resolution) and 7 (Argument Selection and Coercion). Task 4 is similar to already achieved results on learning a concept for passive sentences presented in the next Section. Furthermore, we assume that the key to task 7 is a mixture of DBpedia and syntactical features. Likewise, the Reuters[5] benchmark is interesting, as we argue that text classification methods can be improved the more a machine can grasp the meaning of the text.

As parsers are error-prone, we will deploy several parsers at the same time and experiment with probabilistic logics for correction.

5 Results

The output of Stanford Parser[6] for POS-tagging and dependency parsing was converted to RDF and mapped to the Stuttgart-Tübinger Tagset ontology[7]. Based on 20 passive and 20 active German sentences from the Negra Corpus, the following concept was learned by DL-Learner: "$(Sentence \sqcap \exists hasToken.(VVPP \sqcap \exists previousToken.(APPR \sqcup VAFIN)))$" signifies a sentence with a 'past participle' (VVPP) preceded by 'temporal, causal, modal and local prepositions' (APPR) or 'finite auxiliary verbs' (VAFIN) – a precise description (for German passive), which can be understood as the intentional meaning of a class PassiveSentence and used to classify new sentences. Given a proper background

[3] available as open-source at `http://code.google.com/p/nlp2rdf`
[4] `http://semeval2.fbk.eu/semeval2.php`
[5] `http://www.daviddlewis.com/resources/testcollections/reuters21578/`
[6] `http://nlp.stanford.edu/software/lex-parser.shtml`
[7] `http://nachhalt.sfb632.uni-potsdam.de/owl/stts.owl` by Chiarcos [2]

corpus, it is possible to create expressive concepts just by selecting sentences. This scenario was realized in the TIGER Corpus Navigator[8], which achieved a high F-measure (above 86%) in an Active Learning approach with only few training sentences [4] .

We assume that it is even possible to learn such concepts for semantic relations – just by selecting sentences – after sufficient concepts have been created at the syntactic level.

At the time of writing the work on this project is still in an early phase. Adapters to several other NLP tools have been implemented and evaluation has just begun. Furthermore, the Wortschatz[9] (a source for statistical data about sentences acquired from the web) was converted to RDF and mapped to DBpedia[6]. NER methods, a prerequisite for relation extraction, will be improved by linking to instances from DBpedia and including parts of the hierarchy.

6 Conclusions and Future Work

Planned contributions include: (1) analyzing and pinpointing the underlying problem in Text mining and Ontology learning, (2) theoretical research to acquire a clear definition of the Semantic Gap, (3) leveraging existing (statistical) NLP approaches with an ontology mapping (parser-ontology pair), (4) instrumentalization of background knowledge from the Web of Data (especially DBpedia and Wortschatz), (5) acquisition of additional linguistic knowledge by supervised machine learning, (6) evaluation-driven methodology and (7) creation of an open-source tool that can be used to improve existing solutions.

References

1. P. Buitelaar, P. Cimiano, and B. Magnini. *Ontology Learning from Text: An Overview*, volume 123. IOS Press, 7 2005.
2. C. Chiarcos. An ontology of linguistic annotations. *LDV Forum*, 23(1):1–16, 2008.
3. S. Hellmann, J. Lehmann, and S. Auer. Learning of OWL class descriptions on very large knowledge bases. *IJSWIS*, 5(2):25–48, 2009.
4. S. Hellmann, J. Unbehauen, C. Chiarcos, and A. Ngonga. The TIGER Corpus Navigator. In *Submitted to ACL system demonstrations*, 2010.
5. J. Lehmann. DL-Learner: Learning Concepts in Description Logics. *Journal of Machine Learning Research (JMLR)*, 10:2639–2642, 2009.
6. M. Quasthoff, S. Hellmann, and K. Höffner. Standardized multilingual language resources for the web of data: http://corpora.uni-leipzig.de/rdf. *3rd prize at the LOD Triplification Challenge, Graz*, 2009.
7. J. Völker, P. Hitzler, and P. Cimiano. Acquisition of OWL DL axioms from lexical resources. In *ESWC*, 2007.
8. S. Wintner. What science underlies natural language engineering? *Comput. Linguist.*, 35(4):641–644, 2009.
9. L. Zhou. Ontology learning: state of the art and open issues. *Information Technology and Management*, 8(3):241–252, 2007.

[8] `http://tigernavigator.nlp2rdf.org`
[9] `http://corpora.informatik.uni-leipzig.de`