

ONTOLOGICAL WORD SENSE DISAMBIGUATION FOR DISCOURSE REPRESENTATION

Guntis Bārzdīņš, Normunds Grūzītis, Gunta Nešpore,
Baiba Saulīte, Ilze Auziņa, and Kristīne Levāne-Petrova

Institute of Mathematics and Computer Science, University of Latvia
(Riga, Latvia)

Abstract

Word sense disambiguation (WSD) along with methods for discourse representation of the parsed text, are among the most difficult tasks in computational linguistics today. Without providing a satisfactory solution to these problems, the true automated semantic processing of texts, as envisioned by semantic web, machine translation, or information retrieval communities, remains only an illusory possibility. Today FrameNet (as a state-of-the-art implementation of frame semantics) and computational semantics approaches provide the best insights into these difficult problems — FrameNet formalizes lexical semantics by attaching word senses to the “frames” or idealized situations, while computational semantics provides a method for constructing discourses grounded in the first-order logic (FOL) interpretation of the text. In this paper we argue that the seemingly unrelated FrameNet and computational semantics approaches appear unrelated only due to the lack of proper formalization of their ontologies. We illustrate our approach by forcing FrameNet and computational semantics ideas to interoperate with OWL DL ontology language (a decidable subset of FOL). As a consequence, we present a new approach for addressing the WSD and discourse representation problems. A side-product of the approach is also a new insight into ontology merging — a well-known problem in the semantic web community.

Keywords: word sense disambiguation, ontology merging, reasoning, discourse representation structures, theme-rheme interaction, anaphora resolution

1. Introduction

It is well known that word sense disambiguation (WSD) problem requires access to vast amounts of common-sense world-knowledge. These vast amounts of world-knowledge can be provided either in the form of structured ontologies, e.g. OntoSem (Nirenburg and Raskin 2004) and FrameNet (Fillmore et. al. 2003), defining the conceptualization of the domain of interest, or in the form of statistical probabilities of word sense co-occurrences derived from annotated text corpora (e.g. the WordNet based SemCor¹ corpus that has been extensively used in Senseval² competitions on WSD).

On the other hand, it is well known that WSD depends on the discourse, in which the words appear. The wider the discourse considered, the more precisely the meaning

¹ <http://www.cs.unt.edu/~rada/downloads.html#semcor>

² <http://www.senseval.org>

of the individual words can be pinpointed. Unfortunately, the methodologies for discourse analysis are less mature, with only a few approaches attempted, but without definitive consensus on the optimal methods. Text meaning representation (TMR) of OntoSem and first-order logic (FOL) based discourse representation structures (DRS) of computational semantics (van Eijck 2005) are examples of attempted approaches. Apart from approaches based on FOL, in linguistics the discourse is dealt with by Givenness Hierarchy (Gundel et al. 1993), analyzing the informational status of nominals³. The whole area of discourse representation construction is suffering also from the lack of definitive success in the area of anaphora resolution.

The above mentioned problems of WSD and discourse representation are tightly intertwined and, in our view, the lack of definitive success is largely due to addressing these issues separately — it is impossible to solve WSD problem without solving the discourse representation problem, and vice versa. Therefore we are presenting an original unified approach for addressing both of these problems simultaneously. Our approach can also be viewed as a hybrid of frame semantics and computational semantics with additional elements of anaphora resolution.

The proposed WSD and discourse representation method is enabled by OWL DL (Smith et al. 2004), a standardized ontology language based on description logics (a decidable subset of FOL) for the semantic web community and lately emerging also as “lingua franca” for integration and “cross-pollination” of many previously unrelated areas of research. A rich set of interoperable tools (editors, syntax converters, reasoners, model builders, etc.) available for OWL DL are facilitating this process.

The techniques presented in this paper are inspired by the frame semantics (FrameNet) and computational semantics discourse representation methodology (as implemented in, for example, ACE system (Fuchs et al. 2005)). Meanwhile both of these well-known approaches also have well-known limitations.

The main limitation of FrameNet is the lack of formal semantics of its frame descriptions. Attempts to port FrameNet into formal ontologies, including OWL DL (Scheffczyk et al. 2006) have resulted only in a limited success due to vagueness of FrameNet “semantic type” concept for the fillers of frame slots. FrameNet is primarily intended for a linguist, who can map FrameNet frames to the phrases in a running text.

The main limitation of computational semantics is its disregard for polysemy of words in natural language — they are treated as mono-semantic predicate names⁴, whose “meaning” is defined only by FOL formulas derived from the text being analyzed.

It is clearly tempting to somehow merge these two mutually complementing approaches: computational semantics would benefit from access to the rich “world knowledge” of lexical semantics⁵ captured by FrameNet, while FrameNet usability would benefit from rigorous FOL-based formalism of computational semantics. Unfortunately, naive merger of these vastly different legacy approaches is hardly

³ Other hierarchies of informational status have been discussed by Ariel (1988).

⁴ Minor polysemy is permitted in some computational semantics implementations, if it can be resolved by syntactic means, such as noun/verb distinction or verb sense disambiguation based on predefined valences.

⁵ It should be mentioned that in this paper we are investigating WSD in the context of FrameNet/frame semantics, however, such course alone covers only the syntagmatic aspects of valences, but not the paradigmatic ones. A popular though rather shallow approach to address the latter is semantic networks and WordNet-like taxonomies in particular (Fellbaum 1998).

possible. Therefore we will effectively re-build both approaches from scratch using the formal OWL DL ontology language.

2. Differentiation of ontological and factual sentences

One of the key ideas of our approach is to differentiate two kinds of sentences encountered in natural language: the *ontological sentences* and the *factual sentences*. Ontological sentences are those typically found in dictionaries or school-books defining relationships between categories, for example, “*every dog is an animal*”. Meanwhile factual sentences are those talking about individuals belonging to specific categories, for example, “*a black dog chases a white cat*”. Natural expressions are often a mixture of both kinds of sentences (“*this cat likes every mouse*”), but for the sake of explanation we will consider here only texts where such distinction on sentence level is possible.

Ontological sentences — descriptions of situations permitted in the domain of interest — correspond to formal ontologies⁶ and therefore in this paper instead of ontological sentences we will consider directly OWL DL ontologies. The remainder of this paper therefore focuses on factual sentences — concrete situations that are permitted by the considered ontologies.

DRS	Discourse	'Our DRS'
<code>exists(X, object(X, cat)</code> <code>& property(X, black)</code> <code>& property(X, big)</code> <code>exists(Y, object(Y, cat)</code> <code>& property(Y, black)</code> <code>& property(Y, sleepy)</code> <code>exists(Z, object(Z, cat)</code> <code>& property(Z, sleepy)</code> <code>& property(Z, dreamer))</code>	<p><i>A black cat is big.</i></p> <p><i>The black cat is sleepy.</i></p> <p><i>The cat that is sleepy is a dreamer.</i></p>	<code>b([cat, x1, x2]).</code> <code>b([black, x2]).</code> <code>b([big, x2]).</code> <code>b([sleepy, A]) :-</code> <code> b([black, A]),</code> <code> b([cat, X, A]).</code> <code>b([dreamer, A]) :-</code> <code> b([sleepy, A]),</code> <code> b([cat, X, A]).</code>
Static anaphora resolution: $X=Y \neq Z$		Dynamic anaphora resolution: backtracking



Figure 1. Static versus dynamic construction of DRS. Instead of more traditional paraphrase, we prefer to visualize the discourse structure as a story-board.

The classic computational semantics approach (rooted in the open world semantics) is suboptimal for anaphora resolution in factual sentences⁷, as is illustrated by example in Figure 1. Therefore for factual sentences we introduce a new Prolog

⁶ It has been shown by ACE (Fuchs et. al. 2005) — a state-of-the-art implementation of computational semantics — that it is possible to paraphrase arbitrary OWL DL ontologies into controlled language texts and vice versa — to construct OWL DL ontologies from texts in a controlled language.

⁷ The classic computational semantics results can be considered either wrong (because they do not correspond to the “intuitive” semantics of the text), or only formally correct within the very specific interpretation enforced by computational semantics.

based Horn-clauses interpretation that closely corresponds to the “intuitive” semantics of the factual sentences being processed. The new interpretation is rooted in the closed world assumption and permits dynamic backtracking over available antecedents for anaphora resolution (see Figure 1 for a comparison of these two approaches).

In our approach a factual sentence is split into theme and rheme parts (see Figure 2), where rheme is an assertion about the theme (a theme can be treated as a select statement in a closed world sense, therefore *ASSERT* :- *SELECT*). In the information structure of the text, the anaphora antecedent is chosen from all the possibilities by dynamic backtracking. The antecedent is not necessarily the closest previous compatible unit in the text.

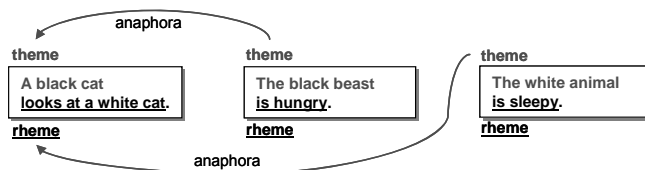


Figure 2. Information structure based anaphora resolution

In contrast to computational-semantics-style translation of sentences into FOL formulas and further application of theorem provers and model builders to interpret the discourse, our method⁸ for factual sentences is faster and scales better, i.e., it is applicable for virtually unlimited discourse scope: paragraph, chapter, book etc.

3. Word meanings and discourse representation

In this section we switch attention to the frame semantics and FrameNet and will try to “fix it” to meet our interoperability needs. Although OWL DL ontologies could come from *ontological sentences* within the text being analyzed (as was described in the previous section), their vast majority shall come from the “background knowledge”. Then why not to use FrameNet frames as a source of the “background knowledge” for OWL DL ontologies? To exploit FrameNet as a source of the background knowledge about lexemes, we first need to formalize the concept of the FrameNet frame.

It is tempting to view a FrameNet frame as an isolated OWL DL domain ontology, which is “invoked” by the characteristic lexemes belonging to the particular frame. The “only” problem is that in the FrameNet approach each sentence is a superposition of a number of FrameNet frames, as shown in Figure 3. One lexical unit may have more than one annotation; its role can change when different target-words are described; and same lexical unit frequently belongs to several frames.

To accommodate the overlapping frames phenomena illustrated in Figure 3, it is necessary to perform on-the-fly domain ontology (frame) merging. In the formal OWL DL based approach (the original FrameNet is non-formal⁹) one would need to ensure that such on-the-fly merged ontologies are consistent.

⁸ Interestingly, in (Blackburn and Bos 2005) possibility for such direct discourse construction for factual sentences is mentioned, but is further discouraged.

⁹ Here we are stepping over a philosophical question: is it possible to define a word meaning by the set-theoretic approach on which OWL DL is based? In general, the answer most likely is “no” due to the tacit knowledge, but for domains with clean conceptualization the answer is “yes”, which thus justifies the described approach.

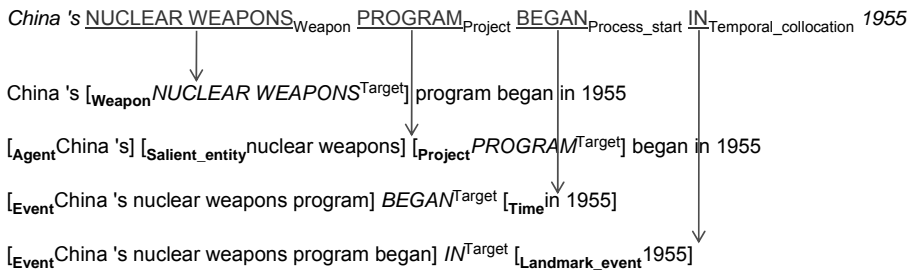


Figure 3. Annotation of running text in FrameNet

Fortunately, if we are sticking to the OWL DL as an ontology description language, consistency of any ontology can be checked in finite time with off-the-shelf OWL DL reasoners, like Pellet¹⁰ or FaCT++¹¹. We are going to use this brilliant property of OWL DL to differentiate the possibly conflicting “word meanings” used in different domain ontologies (frames).

Figure 4 illustrates our proposed approach, which bears a lot of similarity with WSD through semantic mirrors (Dyvik 2005) or OntoSem (Nirenburg and Raskin 2004), but is based on formal and hence fully automated methodology.

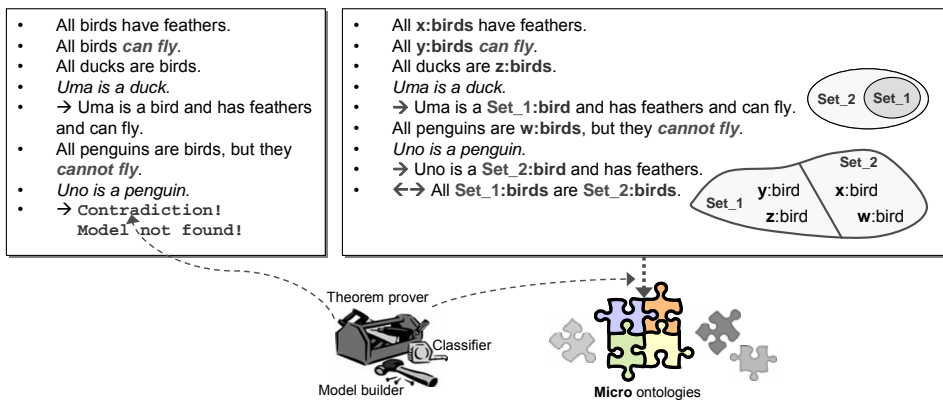


Figure 4. Meaning splits permit use of seemingly “contradictory” micro-ontologies by partitioning the word meanings

In Figure 4 sentences on the left side are the input text. Each sentence in this text can be considered to be either a micro domain ontology¹² (if it is an *ontological sentence*) or an informative statement about individuals (if it is a *factual sentence*). If interpreted straightforward (each word has just one, fixed meaning), the input text is contradictory, what can be easily verified also with formal reasoners. The essence of our approach is that words are permitted (although this is not encouraged) to have different meanings in different ontologies — the only restriction maintained is that within the

¹⁰ <http://pellet.owldl.com>

¹¹ <http://owl.man.ac.uk/factplusplus>

¹² In practice domain ontologies are larger than what a single sentence can cover, but treating individual sentences as individual ontologies is also fine.

same domain ontology words must have the unique meaning. Thus the contradictions in the input text can be avoided by splitting the contradicting meanings in which the same word has been used¹³ in the input text along with the background knowledge acquired from the FrameNet sourced ontologies. If we minimize the number of meaning-splits, then eventually we shall arrive at the correct word meanings by considering a sufficiently large text corpus. Figure 4 illustrates a “correct” minimal meaning split on the right side.

Additionally, we can take into consideration also the possible relationships between the newly split meanings. This is illustrated by the set inclusion in the Figure 4, which corresponds to the last sentence “*All Set_1:birds are (also) Set_2:birds.*” The “correct” set inclusion relationship is determined by maximum set overlap not leading to contradictions.

With this approach the building of the actual discourse representation becomes equivalent to building a minimal model for the FOL theory consisting of the on-the-fly merged and disambiguated ontologies (described in this section) plus the anaphorically resolved individuals from the factual sentences considered (described in Section 2). A model builder can be used for this task¹⁴.

4. Conclusions

To summarize, we have described a framework, within which the rich FrameNet-style lexical semantics (background knowledge) and modified computational semantic-style discourse representation can actually be merged. In our view this merged approach opens a new perspective on WSD and discourse representation.

It should be noted that currently we are only at the proof-of-a-concept stage testing with small examples. There is no evaluation on wide coverage yet.

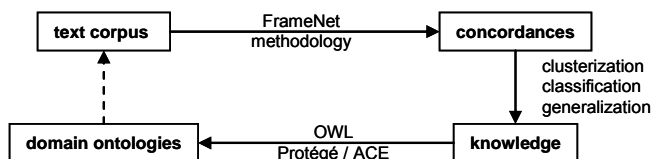


Figure 5. Methodological and tool-supported construction of objective and consistent (domain) dictionaries/ontologies from representative corpora

A huge number of consistent micro domain ontologies are needed for practical applications. A methodological re-engineering of existing ontologies (like FrameNet) would be beneficial, bearing in mind that the interest is in fine-grained, unambiguous concepts that are formally precisely described using “low-level” lexical names. By keeping close to this principle, concepts can be invoked directly from a text via lexeme-name mapping. This is where shallow thesauruses (like WordNet) are welcome back as they can provide additional hypothesis (hints) of potential relatedness between concepts; it is a reasoner’s competence to judge whether to accept or reject these hypotheses.

¹³ Our approach is largely analogous to distributed DL and „bridge axioms” introduced by (Borgida and Luciano 2002); the difference is in the lexical WSD context used here.

¹⁴ We have developed a wrapper (Barzdins 2007a) for Mace4 model builder, which directly accepts OWL DL ontologies and visualises the result in a story-board notation.

As we have made a clear separation between factual and ontological sentences, it might be possible to acquire micro ontologies not only via the formal OWL editors, but also by applying ACE-like text processing techniques. This approach could be developed to learn ontologies from controlled texts or chunks (Bārzdīņš et. al. 2007b) of running texts, which would facilitate the acquisition of a critical mass of micro ontologies. A summary for this observation is given in Figure 5.

Acknowledgements

The research is funded by the National Research Program in Information Technologies and is partially supported by European Social Fund.

References

- Ariel, Mira 1988. Referring and accessibility. In: *Journal of Linguistics*. 24: 65–87.
- Barzdins, Guntis and Barinskis, Martins 2007a. The Minimal Finite Model Visualisation as an Ontology Debugging Tool. In: *Proceedings of the 20th International Workshop on Description Logics*, Brixen. 523–524.
- Bārzdīņš, Guntis; Grūzītis, Normunds; Nešpore, Gunta; Saulīte, Baiba 2007b. Dependency-based hybrid model of syntactic analysis for the languages with a rather free word order. In: *Proceedings of the 16th Nordic Conference of Computational Linguistics*, Tartu. 13–20.
- Blackburn, Patrick and Bos, Johan 2005. *Representation and Inference for Natural Language*. CSLI.
- Borgida, Alex and Serafini, Luciano 2002. Distributed description logics: Directed domain correspondences in federated information sources. In: *Proceedings of the Intentional Conference on Cooperative Information Systems*. 36–53.
- Dyvik, Helge 2005. Translation as a Semantic Knowledge Source. In: *Proceedings of the 2nd Baltic Conference on Human Language Technologies*, Tallinn. 27–38.
- Fellbaum, Christiane 1998 (ed.). *WordNet: an Electronic Lexical Database*. MIT Press.
- Fillmore, Charles J.; Johnson, Christopher R.; Petruck, Miriam R. L. 2003. Background to Framenet. In: *International Journal of Lexicography*. 16: 235–250.
- Fuchs, Norbert E.; Höfler, Stefan; Kaljurand, Kaarel; Rinaldi, Fabio; Schneider, Gerold 2005. Attempto Controlled English: A Knowledge Representation Language Readable by Humans and Machines // In: *Reasoning Web*. Springer. 213–250.
- Gundel, Jeanette K.; Hedberg, Nancy; Zacharski, Ron 1993. Cognitive status and the form of referring expressions in discourse. In: *Language*. 69: 274–307.
- Nirenburg, Sergey and Raskin, Victor 2004. *Ontological Semantics*. MIT Press.
- Scheffczyk, Jan; Baker, Collin F.; Narayanan, Srini 2006. Ontology-based reasoning over lexical resources by means of ontologies. In: *Proceedings of OntoLex 2006*, Genoa. 1–8.
- Smith, Michael K.; Welty, Chris; McGuinness, Deborah L. (eds.) 2004. *OWL Web Ontology Language Guide*. W3C Recommendation.
- van Eijck, Jan 2005. Discourse Representation Theory. In: Brown K. (ed.). *Encyclopedia of Language and Linguistics*. Elsevier.