Bridging the Word Disambiguation Gap with the Help of OWL and Semantic Web Ontologies

Steve Legrand, Pasi Tyrväinen
Department of Computer Science
University of Jyväskylä

stelegra@cc.jyu.fi

Pasi.Tyrvainen@jyu.fi

Harri Saarikoski
Department of Linguistics
University of Helsinki

harri.saarikoski@kolumbus.fi

Abstract

Due to the complexity of natural language, sufficiently reliable Word Sense Disambiguation (WSD) systems are yet to see the daylight in spite of years of work directed towards that goal in Artificial Intelligence, Computational Linguistics and other related disciplines. We describe how the goal could be approached by applying hybrid methods to information sources and knowledge types. The overall aim is to chart the shorfalls of the present WSD systems related to the use of knowledge types and information sources in them. Real world ontologies and other ontologies in the Semantic Web will make a useful contribution towards the WSD knowledge base envisaged here. The inference capabilities inherent in Ontology Web Language (OWL) especially will have an important role to play in natural language disambiguation and knowledge acquisition. The emphasis is on ontologies as one of the important information sources for hybrid WSD.

1 Introduction

WSD methods have undergone changes and increased in number and variety in recent times, reflecting the requirements of many different types of uses WSD is put into. New types of information sources have appeared enabling the utilization of various types of knowledge incorporated in them. Nevertheless, there is no such thing as 100% WSD in any domain, however restricted. This, in fact, smacks of a goal that cannot be realized, taking into account that human WSD cannot reach that goal either.

1.1 Disambiguation Gap

However, we can try and get as close to the level of human WSD as possible. For this we need to specify the gap that currently exists between machine WSD and human WSD. This gap varies and is influenced by:

- *application domain* (machine translation, information retrieval, information extraction, knowledge acquisition, textual data mining, and natural language understanding among them).
- *information sources* available for those domains (machine-readable dictionaries, ontologies, corpora and their combinations among others), and
- *knowledge types* that these information sources incorporate (part-of-speech information, morphology, collocations, semantic associations, syntactic cues, sense frequencies, selectional preferences etc.).

Attempts to systemize the variety of sources and types have been made (Agirre and Martinez 2001, Ide and Véronis 1999) with some success, and attempts to unravel the knowledge types used for particular application domains, for example machine translation (Mowatt 1999),

have also been made. The field now seems ripe for an approach that takes advantage of the potential of hybridisation in the multiplicity of these methods to optimise the effectiveness of WSD.

1.2 Dealing with the Disambiguation Gap

First, it is important to specify the disambiguation gap for various application domains by using our existing knowledge about information sources and knowledge types, and by experimenting with their combinations. The optimum mix varies from application to application: corpus statistical methods can support manual or semi-automatic knowledge methods. The relative weight of each method and knowledge type may be tested on corpora and defined. Hybrid methods (Ng and Lee 1996) and unsupervised methods (Yarowski 1995) have proved their mettle in comparative studies.

Second, one needs to identify those sources and types whose disambiguation potential is currently under-utilised due to their poor availability, costly acquisition, or insufficient appreciation. In particular, the use of ontologies for NLP tasks must be investigated, as ontologies will play a dominant part in the creation of the Semantic Web. OWL seems like a very good choice for disambiguation in which dense ontologies and disambiguation rules are used. There has been a rapid increase in the inference capabilites of Semantic Web languages with each layer added on RDF (RDFS → OIL → Daml-OIL) (Antoniou, 2002). At the time of this writing, the OWL standard is in its Last Call Working Draft phase (McGuinnes and van Harmelen, 2003), and may become a standard before the publication of this paper, and can then be added on top of the stack of the Semantic Web languages. As an indication of OWL's widespread acceptance at this early stage, one can cite some of the tools and technologies that have already been developed to take advantage of it, among them the knOWLer (2003) information management system and the OWL Converter (2003) converting from DAML-OIL format to OWL.

A further advantage of using the OWL together with the Semantic Web ontologies is in the distributed nature of those ontologies. A useable set of domain ontologies will take a considerable time to create: the task will never be completed, because new words and concepts are entering the vocabulary constantly. Hundreds and thousands of experts are needed to make sense of the world. The Open Source community exemplifies the way that the collaborative potential of likeminded people can be effectively harnessed. Already there are Open Source type development groups active in the Semantic Web.

Before Semantic Web came into being, a WSD system for machine translation, Mikrokosmos, based on knowledge-dense ontologies represented by TMRs (Text Meaning Representation) (Mahesh and Nirenburg, 1995) saw daylight: its word sense disambiguation success for all the words in a corpus, 97%, for both training and unseen text, and 90% for ambiguous words (Mahesh et al 1996) is encouraging, but the application lacks portability.

2 Hybrid Multilevel Disambiguation

In hybrid multilevel disambiguation, the idea is to disambiguate word senses using a mix of knowledge types and information sources, including real world knowledge in ontologies and their inference capabilities.

In the ontology related method, the correct ontology that corresponds to the document or document part domain is first detected. The domain itself can be automatically classified by using a hybrid method, for example a committee approach, as outlined in Hammond et al (2002). The domain ontology thus determined and located in the Semantic Web can then be used as the basis for the subsequent hybrid disambiguation.

For example, a paragraph such as:

"*I sat in my old buggy It was very hot, so I turned on the engine, and drove under a tree to get cooler. Then I opened the window.* "

can be disambiguated in several levels:

1. Morpho-syntactic level
2. Semantic level
3. World knowledge level

All of these levels overlap to some extent. As the result of morpho-syntactic level disambiguation, a sentence is annotated with POS (Part-Of-Speech) and morphosyntactic feature tags. From these we derive the syntactic function labels indicating whether the word is the subject, object, predicate, modifier or complement (Figure 1).

The semantic level can use these annotations together with the selectional preferences of the words themselves to clarify the meaning further. For example in the example paragraph above, 'buggy' is more likely to denote a kind of car than a horse-drawn carriage or a baby pram, the noun 'engine' regularly co-occurring with the noun 'car' in the same context. The Part-of ontological hierarchy in Figure 2 confirms their

I	i	@SUBJ
sat	sit	@+FMAINV
in	in	@ADVL
my	i	@A>
old	old	@A>
buggy	**buggy**	**@<P**
....	....	....
I	i	@SUBJ
turned	turn	@+FMAINV
on	on	@ADVL
the	the	@DN>
engine	engine	@<P
and	and	@CC
drove	**drive**	**@+FMAINV**
under	under	@ADVL
the	the	@DN>
tree	tree	@<P
......	......	......
I	i	@SUBJ
opened	open	@+FMAINV
the	the	@DN>
window	window	@OBJ

Figure 1. Syntactic functions (column 3) for the example paragraph according to FDG. Morphosyntactic feature tags are not shown here.

Sumo ontology (in bold):
Entity
  Physical
    Object
      SelfConnectedObject
        ContentBearingObject
          LinquisticExpression
           Word
             Noun
             Verb
           Phrase
             NounPhrase
               Object
               Subject
             VerbPhrase
               Predicate
        CorpuscularObject
          Artifact
            Device
              TransportationDevice
Domain ontology fragments :
               Vehicle
                 MotorVehicleType
                   bus
                   car
                     Jeep
                     sedan
                     buggy
                     ……
                 MotorlessVehicleType
                   bicycle
                   horse carriage
                     buggy
                     ………..
                   baby carriage
                     pram
                     buggy
                     …..
                 …….
                 MotorVehiclePart
                   engine
                   transmission
                   electrical system
                   body
                     roof
                     bumper
                     floor
                     door

Figure 2: SUMO top ontology (bold) subsuming the syntactic function and transportation domain ontologies (lighter color).

close relationship. Naturally, a rule specifying this would need to span over the sentence boundaries.

Even though we are aware now that the protagonist has started the engine of his car, it would be difficult by morpho-syntactic and semantic disambiguation alone to reason that the sentence following: *"Then I opened the window…"* would necessarily mean the car window. In this a real world ontology would be of great help. Apart from confirming that 'buggy' can be a type of 'car', it would confirm that the window, in this case, is a part of a car and not of a house, and that horse-drawn carriages or baby prams do not have engines. One could reason further that the car was a type of a vehicle etc, the selectionally preferred noun for the verb 'to drive' was 'a car' etc. If, however, the context of the paragraph was established to be that of golf, then of course the selectional preference for the verb 'drive' in 'I drove under a tree' would most likely change. These types of inferences could be drawn from ontologies built with the help of OWL, combined to minimize the word sense ambiguity.

The above is to give a rough idea of the way word sense disambiguation can be handled if OWL's inferencing capabilities are combined with traditional means of disambiguation. The example is just to illustrate the idea, and does not provide enough details for a complete disambiguation of the paragraph. It is easy to find fault with it by insisting, for example, that 'buggy', according to a definition found in many dictionaries is a small vehicle without windows and doors and with a roof mounted on the chassis. This would contradict what is said about the window above, unless one classified some off-road vehicles such as converted VW's (with windows) as buggies, which is also quite common. This further illustrates the importance of not relying excessively on any single information source or disambiguation method in trying to reduce the disambiguation gap.

Once the correct domain ontology and the position of the word in it denoting the concept are determined, the word can be matched with the foreign word in the same ontological structure if the purpose is to translate it. It may be that a house 'window' and a car 'window' in another language are denoted by two entirely different words, unlike in English, and therefore it is important to select the correct ontological concept.

3   FDG and Ontological Approach

We use a morpho-syntactic Functional Dependency Grammar (FDG, 2003) analyser as the baseline setter on which to found our research. The FDG analyser is based on ENGCG parser which, when combined with Xerox tagger, reached 98.5% structural disambiguation accuracy, outperforming all the other parsing combinations tested in the study of Tapanainen and Voutilainen (1995). FDG was selected for the present research mainly due to its accuracy. However, although its disambiguation error rate seems very small, it is still significant when considering natural language applications. Using the same formula as Abney (1996) it is easy to show that this word-based disambiguation rate, when applied to sentence-level, still needs some improvement to satisfy the requirements of natural language processing applications. If we assume that a sentence consists of 20 words on an average, the 98.5% word disambiguation accuracy is transformed into 26% error rate on a sentence level ($1-.985^{20} = 26\%$). For the purposes of machine translation this is clearly not yet adequate (1/4 of all the sentences erroneous even in ideal cases where the domain is restricted).

The current morpho-syntactic word sense disambiguation in FDG will soon be augmented with a semantic disambiguation module, which is likely to further improve the parser's accuracy. This is not sufficient, however. In addition to morphological, syntactic and semantic word sense disambiguation, real world knowledge is required for optimal understanding of a natural language. In our approach, the gap remaining in the disambiguation that cannot be bridged using the mix of currently available methods and their modifications is subjected to ontological disambiguation using real world distributed domain ontologies and SUMO upper ontology (Pease et al 2002) in the Semantic Web. Currently, most of these ontologies are in the RDF-based DAML-OIL format, but can be converted to OWL, the standard that is expected to replace DAML-OIL in the near future.

Farrar et al (2002) have suggested the addition of a general ontology for linguistic description (GOLD) to the SUMO upper ontology and published a draft version of it in the OWL format. They see it useful as a part of an expert system reasoning about language data, or as a part of an interlingua for machine translation system. We envisage being able to use their linguistic ontology to hold disambiguated

```
<owl:Class rdf:ID="Car">
  <rdfs:subClassOf
rdf:resource="#MotorVehicleType" />
</owl:Class>

<owl:Class rdf:ID="Buggy">
  <rdfs:subClassOf rdf:resource="#Car" />
</owl:Class>

---------------------------------------------
<owl:Class rdf:ID="Engine">
  <rdfs:subClassOf
rdf:resource="#MotorVehiclePart" />
</owl:Class>

<owl:Class rdf:ID="Body">
  <rdfs:subClassOf
rdf:resource="#MotorVehiclePart" />
</owl:Class>

<owl:Class rdf:ID="Window">
  <rdfs:subClassOf rdf:resource="#Body"/>
</owl:Class>

---------------------------------------------

<owl:Class rdf:ID="Predicate">
  <rdfs:subClassOf
rdf:resource="#VerbPhrase" />
</owl:Class>

<Predicate rdf:ID="Drive">
    <selectionalPreferenceObject
    rdf:resource="#Car" />
</Verb>
```

Figure 3. Owl fragments connecting ontologies with subsumption and relational properties. Namespace declarations, superclass definitions, and property definitions are omitted.

morphological and syntactic data. Syntactic functions for nouns could indirectly be indicated in the case system portion of GOLD. However, these can also be plugged directly to the SUMO upper ontology (Figure 2), although their positioning under the SUMO's Phrase category may prove unsatisfactory in the long run.

The domain ontology holding the real world data and relations for the words can then be aligned to the SUMO ontology (Figure2) mainly through OWL subsumption and to the linguistic ontology using OWL property relations (Figure 3) as the glue: inferences can be drawn from the real world data to increase the disambiguation power of the linguistic ontology. For the purpose of alignment, SUMO also need be formatted to OWL. An agent-based application can then be used to manipulate the structures created for linguistic disambiguation.

The coding and property and class relations in Figure 3 are grossly simplified fragments forming part of an OWL document. The idea here is to show that the verb 'drive' selects a car rather than a baby carriage as its preferred noun phrase object. The word 'buggy' may subsequently be matched with 'car', and 'window' to the car body through their part-of relations. Similarly, both 'body' and 'engine' would be identified as parts of a motor vehicle. A comprehensive OWL statement about the verb 'drive' would have a set of preferentially weighed selectional preference entities to select from and a set of restrictions applied to it. Contextually closest (shortest arc distances in the ontology) selectional preferences would have the greatest preference weighing.

4 WSD Knowledge Base

Ontologies to be tested and designed for our optimal WSD include new and existing ontologies in the Semantic Web suitably modified and contain, for each concept, a dense network of subclass/superclass (eg. car is-a motorVehicle) relationships, property rules (eg. selectional preferences) and associative relations. Essentially, the WSD Knowledge Base will contain the differentiating factors between two senses of a word, which will disambiguate the sense of the target word. Synonym sets may be thought of as a differentiator, the sense's place in

the Knowledge Base hierarchies and categories as another. Selectional preferences of the senses, and, of course, context word statistics - among other differentiators -, can also be used for disambiguation.

It is in this density and multiplicity of knowledge types and "sub-atomicity" (concepts are defined rigorously and adequately from within) that it contrasts the traditional, atomic (concepts are only defined in terms of their few external relations to related terms in network) ontologies. The Mikrokosmos ontology holds 5000 concepts with an average of 16 attributes and relations per concept (Mahesh et al. 1996). Our WSD knowledge base starts from that density and increases/decreases density until WSD is optimised. The result will be referred to as the *WSD Knowledge Base* for which we define each aspect of its construction and functioning. We will rigorously define the principles of designing such knowledge base, both in terms of quality (knowledge types required) and quantity (number of concept-internal definitions = information from knowledge types). As such, this research will also provide a feasible requirements specification for eventual implementation of the WSD system described.

One important application for our optimal WSD system is knowledge acquisition. It is precisely the lack of knowledge, and the high cost of acquiring dense knowledge bases and ontologies, that stands as the bottleneck in the way of knowledge-based NLP systems becoming more useful. WSD Knowledge Base may deliver a solution to both structural and semantic disambiguation tasks, and can as such be utilised in a multitude of NLP applications.

The WSD system could then be tested using corpora and test cases (disambiguable target words) from earlier research. Such starting points, and also points of comparison, could be Ng and Lee (1996) who tested their hybrid system on the senses of a single noun 'interest', Bruce and Wiebe (1994) who worked with the same noun, or Towell and Vorhees (1998) who tested some highly disambiguous words such as 'line' (noun), 'serve' (verb), and 'hard' (adjective). Another possibility, offering an equal amount of comparability to previous research, would be to examine the systems from the SENSEVAL-2 (2001) competition to see what

knowledge types and information sources would most naturally and effectively disambiguate the target words.

5 Conclusion

This paper outlines the three main aspects in bridging the current disambiguation gap in WSD: application domain, information sources and knowledge types. There is a multiplicity of different domains, sources and types. Methods dealing with them have their limitations and can be partially overcome by combining the best of them in hybrid methods. It is important to determine the part of the disambiguation gap for language understanding that is dependent on knowledge aqcuisition.

Ours is an attempt to quantify the disambiguation potential of each information source and their contained knowledge types for each target word type. For example, if we find that what differentiates a word from another is synonym sets, points are added to the knowledge type and information source involved. The idea is to get an overall view on the most useful differentiating and disambiguating factors, knowledge types, and information sources in each particular case. Efforts in the KA and NL communities can then be better directed toward acquiring these information sources and knowledge types and developing more reliable hybrid WSD systems.

The FDG parser that we use in our morpho-syntactic and semantic disambiguation provides a mix of knowledge types (POS, morphology etc) to which we add selectional preferences and other types for the purpose of semantic / world knowledge disambiguation.

Ontologies as information sources are gaining momentum thanks to the emerging Semantic Web language specifications such as RDFS, DAML-OIL, and the most recent arrival, OWL, with its enhanced inference capabilites suitable for knowledge-based NLP. The use of OWL ontologies further reduces the disambiguation gap by allowing word sense disambiguation with the help of real world knowledge contained in Semantic Web domain ontologies.

It is still the early days. However, the OWL will become a standard soon, the SUMO upper ontology will be translated to OWL in a due

course, and linguistic ontologies and ontologies from other domains (knowledge-saturated and knowledge-optimized ontologies) will be added and aligned with it. It is our hope that this paper can offer a glimpse of how Semantic Web, saturated ontologies, and OWL can contribute as one of the disambiguation methods used in hybrid WSD.

References

Abney, S., Part-Of-Speech Tagging and Partial Parsing, In: Church, K., Young, S., Bloothoof, G., Methods in Language and Speech. An ELSENET book, Kluwer Academic Publishers, Dordrecht, 1996.

Antoniou, G., Nonmonotonic Rule Systems on Top of Ontology Layers, *Lecture Notes in Computer Science*, 2342, Online publication: May 29, 2002. Available in: http://link.springer.de/link/service/series/0558/bibs/2342/23420394.htm

Agirre, E., and Martinez, D., Knowledge Sources for Word Sense Disambiguation, *Lecture Notes in Computer Science* 2166, Springer 2001.

Bruce, R., and Wiebe, J., Word-Sense Disambiguation Using Decomposable Models, In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, Las Cruces, New Mexico, 1994.

Farrar, S., Lewis, W.D., Langendoen, D.T., A Common Ontology for Linguistic Concepts, *Proceedings of the Knowledge Technologies Conference*, March 10-13, Seattle, 2002.

FDG. Conexor Functional Dependency Grammar. In: http://www.conexoroy.com/fdg.htm, Last accessed: June 2003

Hammond, B., Amit, S., Kochut, K, Semantic Enhancement Platform for Semantic Applications over Heterogenous Content, To appear in *Real World Semantic Web Applications, V.Kashyap and L.Shklar, Eds, IOS Press, 2002*. Available in: http://lsdis.cs.uga.edu/lib/download/HSK02-SEE.pdf

Ide, N., and Véronis, A., Word Sense Disambiguation: The State of Art, *Computational Linguistics*, Vol.24, No.1, March 1998, p.1-40

knOWLer. Ontology based information management system. In: http://taurus.unine.ch/GroupHome/knowler/wordnet.html. Last accessed: June 2003.

Mahesh, K., Nirenburg, S., Beale, S., Onyshkevych, B., Viegas, E., and Raskin, V., Word Sense Disambiguation: Why Statistics When We Have These Numbers? 1996.

Mahesh, K., and Nirenburg, S., A Situated Ontology for Practical NLP, in *IJCAI-95 Workshop on Basic Ontological Issues in Knowledge Sharing*, Aug. 19-21, Montreal, 1995.

McGuinnes, L.D., van Harmelen, F., Eds., OWL Web Ontology Language Overview, W3C Working Draft 31 March 2003, Available in: http://www.w3.org/TR/owl-features/

Mowatt, D., Types of Semantic Information Necessary. In *Machine Translation Lexicon, Conférence TALN 1999*, Cargèse, 12-17 July 1999

Ng, H. T. and Lee, H.B., Integrating multiple knowledge sources to disambiguate word sense: An exemplar-based approach, 1996. *In Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, Santa Cruz, California, 1996.

OWL Converter. In: http://www.mindswap.org/2002/owl.html. Last accessed: June 2003.

Pease, A., Niles, I., Li, J., (2002) The Suggested Upper Merged Ontology: A Large Ontology for the Semantic Web and its Applications. In *Working Notes of the AAAI-2002 Workshop on Ontologies and the Semantic Web*. Available in: http://reliant.teknowledge.com/AAAI-2002/Pease.ps

Rigau, G., Magnini, B., Agirre, E., Vossen, P., and Carroll, J., MEANING: a Roadmap to Knowledge Technologies, 2002.

SENSEVAL-2. *Second International Workshop on Evaluating Word Sense Disambiguation Systems*. 5-6 July 2001, Toulouse, France. In: http://www.sle.sharp.co.uk/senseval2/

Tapanainen, P. and Voutilainen, A., Tagging accurately: don't guess if you don't know. Technical Report, Xerox Corporation, 1994.

Towell, G., Voorhees E.M., Leacock, C., Disambiguating Highly Ambiguous Words In *Computational Linguistics* Volume 24, Issue 1 / March 1998, p. 125 – 145

Yarowsky, D., Unsupervised word sense disambiguation methods rivaling supervised methods, *ACL95 - 33rd Annual Meeting of the Association for Computational Linguistics* 26-30 June 1995,Massachusetts Institute of Technology, Cambridge, Massachusetts, USA, 1995.