

Dynamic word sense disambiguation with lexical ontologies in computational semantics

TJHSST Senior Research Project

Computer Systems Lab 2009-2010

Sam Zhang

October 30, 2009

Abstract

How do we assign meaning to words? This project investigates semantics from a lexical perspective, using Python, the Natural Language Toolkit, and the WordNet and OpenCyc ontologies to create a semiotic map of our consensus reality. Given a list of words, how can we find the word least like the others? Through a heuristical search across the corpora of lexical relations, computational semantics can discover the contextual meaning of words, even when the only context given is the other words from which it must differentiate itself. This method, which has not been given a name previously, will hitherto be known as dynamic word sense disambiguation. **Keywords:** semantic ontology, computational linguistics, computational semantics, dynamic word sense disambiguation

1 Introduction

What are ontological semantics? Semantics is the study of the meaning of words, and an ontology is a network showing the relationships between words. The relationships most relevant in this scenario are lexical relations: synonyms, antonyms, hypernyms, hyponyms, meronyms, and holonyms (figure 1). This project disambiguates lexical semantics and discovers "errors", or the odd one out, through a heuristical search across these lexical relations.

The ontological approach to semantics is relatively new compared to other linguistic studies, yet is already burgeoning due to its relevancy to postmodern semiotic environments, namely the internet, which is a complex ontology in itself. The use of the internet as a wealth of corpora is not a new idea, marketing research companies and sociologists have been using data mining tech-

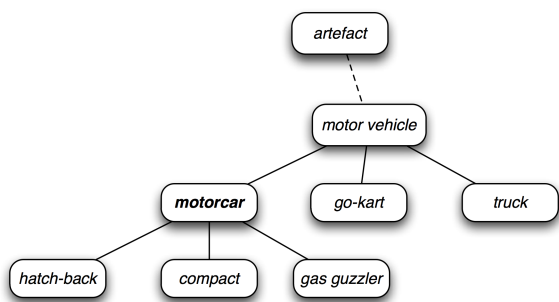


Figure 1: Fig 1: A hypernym/hyponym ontology. Source: <http://nltkbook.googlecode.com>

niques to determine online trends for years. Director of the World Wide Web Consortium Sir Tim Berners-Lee is strongly advocating the Semantic Web as the next step for the internet's evolution, in other words, Web 3.0. In the semantic web, the internet will become fundamentally interlinked with artificial intelligence, providing an automated source of information and help. The W3C endorses a new language for the internet to replace HTML, most notably the Web Ontology Language.

Existing word sense disambiguators have already reached a high degree of accuracy in statistical corpora analysis, for example in part-of-speech tagging. However, this project focuses on disambiguation on a lexical level, rather than sentential, to investigate the extremes to which a word's denotation can still be found. By forcing the computer to discern between subtleties in word meaning using a minimum of context, this author will use a technique philosophically outlined by a book titled, "Ontological Semantics", which

envisioned a network-based approach to semantics to augment or even replace the traditional statistical corpora method.

2 Development

2.1 Theory

Existing word sense disambiguators have already reached a high degree of accuracy in statistical corpora analysis, for example in part-of-speech tagging. However, this project focuses on disambiguation on a lexical level, rather than sentential, to investigate the extremes to which a word's denotation can still be found. By forcing the computer to discern between subtleties in word meaning using a minimum of context, this author will use a technique philosophically outlined by a book titled, "Ontological Semantics", which envisioned a network-based approach to semantics to augment or even replace the traditional statistical corpora method.

2.2 Languages and Modules

The author uses Python as the primary programming language, the natural language toolkit which facilitates corpora manipulation, and the Wordnet corpora, an ontology of common English words. Through a heuristic search across the lexical ontology of the words using such traversions as hyponymy, hypernymy, synonymy, antonymy, melonymy, and holonymy, the computer will be able to discover the average semantic distance from one word to all of the others. Perhaps this ap-

proach will be supplemented with a statistical analysis of corpora, as the Wordnet database of lexical relations could benefit with an update. The OpenCyc project is such a corpora that attempts to update with the current semiotic sphere, although some concerns have been raised by the computational linguistics community on its accuracy and scalability.

2.3 Description of Project

This project employs semantic ontologies to perform dynamic word sense disambiguation on a lexically isolated set of words with only each other as reference. The program will heuristically search through the semantic net, an ontology of lexical relations, to determine the "distance" between the referents of the different words to determine a relevancy index for each word, dynamically based on its environment. The word with the smallest relevancy index is marked as the "odd one out", or the error.

2.4 Tests or Analysis

The strength of the project's ontological traversal will be measured in two aspects: accuracy and time. The accuracy will be the aspect most focused on initially, while optimizations can occur after the accuracy is ensured. As a prototyping step, a graph could be created of the semantic network as the program conducts the bidirectional search.

3 Discussion

Statistical corpora analysis is a computational linguistics tool that could augment the strictly ontological dynamic word-sense disambiguation. This author has experimented with streaming RSS feeds into corpora with moderate success, but has since discovered the OpenCyc ontology. The OpenCyc ontology contains a semiotic map of the current census reality of the English speaking world, but clearly such an ambitious project must have shortcomings. To develop RSS feed data into a lexical relations data, a simple parsing script can be written to find syntactical structures such as "WORD1 is... of WORD2" to determine holonyms.

4 Recommendations

Aside from its advantageous position at the intersection between consciousness studies and artificial intelligence, computational linguistics has a wide arrange of practical applications, from translation to intelligent computing. Specifically, ontological semantics are being researched as the backbone for Web 3.0, the movement to give the internet basic aspects of intelligence. That could produce a resounding impact on our life, like Asimov's MultiVac, so the ethical issues must be carefully analyzed. On a closer timeline, computational semantics is becoming closer related to neuroscience as scientists move to discover the neurological development of linguistic faculties. This project specifically would be crucial to search engine development, com-

puter dictionaries, human-computer interaction, and the back-end of a speech-to-text filter.

5 Sources

Feldman, J.A. 2004, "Computational cognitive linguistics"

Nirenburg, S., Raskin, V. 2004, "Ontological semantics"