Dynamic word sense disambiguation with semantic similarity
TJHSST Senior Research Project
Computer Systems Lab 2009-2010

Sam Zhang

January 27, 2010

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? This question gains relevance when the other words are ambiguous as well, and they must be cross-checked for contextual clues. Through a heuristicial search across the corpora of lexical relations, the most similar senses of words has been discovered to also be most likely the intended meanings, 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, semantic similarity, computational linguistics, computational semantics, dynamic word sense disambiguation

1 Introduction

How can similarity between words be measured? Semantics, which traditionally defines the difference between man and computer in the Turing Test, has become increasingly the domain of machine as technology improves its grasp of language. The database of similarities between words is known as an ontology, a word derived from the philosophical term meaning "the study of beings and their relations." Thus by using hypernym similarity within the WordNet corpora, this project disambiguates lexical semantics and discovers when a word occurs out of context.

The ontological approach to semantics is relatively new compared to other linguistical studies, yet is already burgeoning due to its relevancy to postmodern semiotic environ-

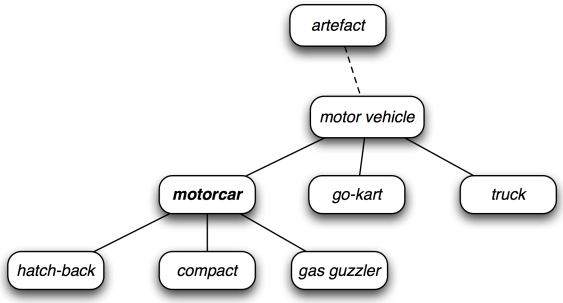

Figure 1: Fig 1: A hypernym/hyponym ontology. Source: http://nltkbook.googlecode.com

ments, 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 techniques 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 investiate the extremes to which a word's denotation can still be found. By forcing the computer to discern between subleties in word meaning using a minimum of context, this author will use a technque 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 investiate the extremes to which a word's denotation can still be found. By forcing the computer to discern between subleties in word meaning using a minimum of context, this author will use a technque 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. Through ontological semantics, the least common hypernym can be determined between different senses of different words, and thus determining the most similar through an algorithm related to the hypernym distance and root distance.

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 heuristical 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 approach 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. The ontological algorithm used traces the words upstream to their least common hypernym, and compares the distance of the least common hypernym to the root. This number is factored together with the distance of the words to each other (i.e. the average distance to the least common hypernym). With an index relating the similarity between two words, a metric now exists for large scale comparison between different sets of different words.

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 intially, 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. Currently, as per the Wu and Palmer algorithm that this project's algorithm is based on, the project has an 88/100 accuracy rating for determing the most similar words out of a pool, as related to human judgment.

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. Yet this author's research has uncovered the importance of the hypernym and hyponym structure to lexical similarity, to the point where the other lexical relations are insignificant. Thus the OpenCyc ontology could be used to determine such facts as whether or not George W. Bush and Barack Obama are more similar than Sarah Palin (which perhaps they are since they are both of the hypernym "US Presidents", although George W. Bush and Sarah Palin are both Republicans).

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, computer dictionaries, human-computer interaction, and the back-end of a speech-to-text filter. The next step for this project is the transition.

5 Sources

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

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