Dynamic word sense disambiguation with lexical ontologies in computational semanticss TJHSST Senior Research Project Proposal Computer Systems Lab 2009-2010

Sam Zhang

October 28, 2009

1 Purpose and scope of the research project

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 heuristicial 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.

2 Background and review of current literature/research in this area

Computational linguistics is a rapidly developing field; existing word sense disambiguators have already reached a high degree of accuracy in stastical 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.

3 Procedure and Methodology

The author uses Python as the primary programming language, the natural language toolkit which facilitates corpora manipulation, and the Wordnet corpora, an ontology that may be improved. 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 linguistics community on its accuracy and scalability.

4 Expected Results and Value to Others (Applications your project may have)

Aside from its appeal as a forefront of 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.