Announcing the Final Examination of Hansen Schwartz for the degree of Doctor of Philosophy

Time & Location: March 29, 2011 at 3:00 PM in HEC 302
Title: The Acquisition of Lexical Knowledge from the Web for Aspects of Semantic Interpretation

This work investigates the effective acquisition of lexical knowledge from the Web to perform semantic interpretation. The Web provides an unprecedented amount of unannotated sentences from which to gain knowledge useful for semantic interpretation. The lexical knowledge acquired is described as common sense knowledge, information one uses in his or her everyday life to understand language and perception. Novel approaches are presented for both the acquisition of this knowledge and its use for semantic interpretation. The goal is to increase accuracy of automatic semantic interpretation systems, and in turn enable stronger real world applications such as machine translation and question answering.

The major contributions of this dissertation consist of two approaches for acquiring common sense knowledge from the Web, namely a database of common sense knowledge and Web selectors. The first approach is a framework for acquiring a database of relationships. Relationships between nouns are found on the Web and analyzed over WordNet using information-theory, producing knowledge about concepts rather than ambiguous words. For the second contribution, words called Web selectors are retrieved which take the place of an instance of a target word in its local context. The selectors serve for the system to learn the types of concepts that the sense of a target word should be similar. Web selectors are acquired dynamically as part of a semantic interpretation algorithm, while the relationships in the database are useful to enhance stand-alone systems. A final contribution of this dissertation is a novel semantic similarity measure that is valuable for knowledge acquisition and semantic interpretation.

Applications to word sense disambiguation, an aspect of semantic interpretation, are used to evaluate the contributions. Disambiguation systems which utilize semantically annotated training data are considered supervised. The algorithms of this dissertation are considered minimally-supervised; they do not require training data created by humans. In the case of evaluating a database of common sense knowledge, integrating the knowledge into an existing minimally-supervised disambiguation system significantly improved results, a 20.5% error reduction. Similarly, the Web selectors disambiguation system, which acquires knowledge directly as part of the algorithm, achieved results comparable with top minimally-supervised systems, an F-score of 80.2% on a standard noun disambiguation task.

This work enables the study of many subsequent related tasks for improving semantic interpretation and its application to real-world technologies. Other aspects of semantic interpretation, such as semantic role labeling, could utilize the methods presented here for word sense disambiguation. As the Web continues to grow, the amount of knowledge and capabilities continue to increase. Although the Web Selectors system achieves great results, this study also shows likely improvements from acquiring more data. Furthermore, the methods for acquiring a database of common sense knowledge were applied to just one type of common sense knowledge. One would expect a more exhaustive database of many types of common sense knowledge would be even more beneficial in applications. Finally, perhaps the greatest benefit from this work is enabling real world technologies that utilize semantic interpretation.

Major: Computer Science

Educational Career:
Bachelor's of Computer Science, BS, 2004, University of Central Florida
Master's of Computer Science, MS, 2006, University of Central Florida

Committee in Charge:
Dr. Fernando Gomez, Chair, Electrical Engineering and Computer Science
Dr. Annie Wu, Electrical Engineering and Computer Science