Natural Language Processing
Fall 2015 Final Project:
Sense Disambiguation with the Naive Bayes Classifier

Max Dumas

Problem

- Unambiguous understanding of word meaning and usage is essential for correct results in tasks requiring semantic understanding of input, i.e. machine translation, question answering, information retrieval, text classification.
- Widespread homonymy and polysemy, especially in English, result in many possible interpretations for word meaning, combinatorially increasing for sentences.
- Many word senses are often very subtly different and often multiple are applicable in a given context. Humans are excellent at identifying contextual clues that indicate which sense of a word is in use, but identifying the features that tell us this and exposing them to computer is often difficult.

Approach

- As a statistical approach to word sense disambiguation, the Naive Bayes Classifier attempts to choose the set of word senses for an input that is maximally probable given the word-sense pairings observed in context in previous inputs.
- To identify context we extract features that help to indicate the appropriate word sense, and maximize the probability of those as well. Examples of features include the k neighboring word parts-of-speech, the previous word, lemma forms, etc.

$$s^*(w) = \underset{s \in S(w)}{\operatorname{argmax}} P(s|w) \prod_{f \in F(s)} P(f|s)$$

Evaluation

- The primary vector for improvement in this project will be identifying indicative features.
- System will be trained on SemCor3.0-all and SemCor3.0-verbs data (660,000+ words, 230,000+ sense tagged) as it is the most comprehensive sense-tagged corpus available
- Testing will be performed on numerous corpora, including portions of SemCor3.0, as well as SenseEval and MASC.
- Performance will be evaluated based on standard metrics of precision and recall.

Selected References

- Jurasky, Daniel, and James H. Martin. "20.2.2: Naive Bayes and Decision List Classifiers." Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. N.p.: Pearson Education, 2008. 738-40. Print.
- Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze. "Naive Bayes Text Classification." Introduction to Information Retrieval. New York: Cambridge UP, 2008. 258-62. Print.