Proposal of Project 2: Word Sense Disambiguation

Shibo Zang(sz428) Hongfei Li(hl963)

We would like to use supervised learning algorithm -- Naive Bayes as the classifier of Word Sense Disambiguation and to use WordNet dictionary to select feature vectors.

We noticed that it's hard to decide what kind of feature vectors to extract from the surronding context. At first we thought about using the nearest word to serve as the feature vector. However, we feel that only making decisions based on the nearest word is not convincing. For example, "electronic device" is a phrase that occurs in the training corpus. And "device" is our target word. Since the word "electronic" and the word "device" are internally relevant, we feel like we are going to miss this kind of bonding word pairs if this phrase occurs very infrequently. Thus, we want to propose a way to utilize the relevance between words to extract valuable feature vectors.

The way we are doing this is to first eliminate punctuations and meaningless words in the corpus such as prepositions, conjunctions, articles, pronouns, interjections, quantifiers, auxillaries, etc., using part-of-speech tagging. Then we will look at previous 10 words and following 10 words around the target word. We will process these 20 words based on their definitions in the dictionary and reward those words (increase their weights in the feature vector) which have consecutive overlap with the definition of target word. And of course we also need to process the content of words in dictionary, which means to keep nouns, adjectives, verbs, and adverbs. Finally we would get the feature vectors we want and apply them to the Naive Bayes classifier.

In addition, to get more accurate classifier, we would like to try different Naive Bayes models in our implementation, such as Gaussian Naive Bayes, Multinomial Naive Bayes, or Bernoulli Naive Bayes. To test which model performs the best, we would partition the training data set into 80% trainin set and 20% validation set.