Word Sense Disambiguation Considering Undefined Senses

Kikuta Atsushi (410036)

School of Information Science,
Japan Advanced Institute of Science and Technology

February 9, 2006

Keywords: Word Sense Disambiguation, Undefined sense, EM algorithm, Naive Bayes.

Word sense disambiguation(WSD) is the process which automatically determine the meaning of a word in a sentence. It is broadly applicable to tasks such as machine translation, information retrieval and so on. However, there is a word have the meaning which is not defined in a dictionary. For example, for the word a "telephone", the two meanings "communication by telephone" and "a telephone set" are defined in the dictionary. However, it may be used in the sense of a "telephone number" in a sentence. The conventional methods of selecting a appropriate sense from the meaning defined in a dictionary always choose wrong meaning, and cause errors for the task of machine translation or information retrieval. Therefore, in order to solve this problem, the system which can distinguish the undefined meaning is built in this research.

As the techniques of WSD, the supervised machine learning using the word sense-tagged corpus works well, but cannot be applied for this research because any corpora with a undefined word sense tag do not exist. Therefore, unsupervised learning using unannotated corpora is used in this research.

In this research, we distinguish a meaning of word by learning the probabilistic Naive Bayes model by unsupervised learning method using the EM algorithm. Furthermore, the statistics derived from a word sense tagged

corpus is used for setting initial values in EM algorithm. First, we estimate $P(f_i|s_k)$ by maximum likelihood estimation, the probability that word sense s_k generates feature f_i. Next, the features of n higher ranks f_i' are extracted for each word sense and $P(f_i|s_k)$ related to f_i' are given r times values than $P(f_i|s_k)$ related to the other features. Then, the features of m higher ranks f_i'' are extracted and $P(f_i|s_k)$ related to f_i'' are given lower value than $P(f_i|s_k)$ related to the other features. The value of r, n, and m were experimentally adjusted so that the rate of selecting correct answers for undefined senses might become the maximum.

In experiments, we disambiguated senses for ten words including undefined senses by the proposed method, and compared it with two methods. One is the method which learns Naive Bayes model by supervised learning and regards senses of words as undefined senses when the probability of the first ranked meaning is lower than a threshold(BL). The other is the method which gives initial values of EM algorithm as a uniform distribution(EM_{uni}). In each method, we compared systems when the accuracy for all words (Cor_{total}) becomes the maximum. The proposed method achieved 48.5% Cor_{total}, and outperformed BL by 1.9% but is 2.7% lower than EM_{uni}. It is 2.8% lower than BL for F-measure for undefined senses, outperformed EM_{uni} by 26%. It outperformed BL by 51.4% for precision for undefined senses, and EM_{uni} by 81.6%. Next, we compared systems when F-measure for undefined senses becomes the maximum. The proposed method outperformed BL by 11.8% Cor_{total}, is 4% lower than EM_{uni}. It achieved 36.8% F-measure, and outperformed BL by 0.9% and EM_{uni} by 36.8%. It outperformed BL by 37.8% for precision, and EM_{uni} by 63.2%.

There were few numbers distinguished from words with undefined senses in almost all words by the proposed method, and the recall was low. However, as for the word "asa" and a "denwa", recall and precision were high. The value of r, n, and m was changed, and when the accuracy was the maximum, for these two words, precision was 100%. We think the reason why these two words are easily disambiguated as follows. One is that the difference between the meanings defined in the dictionary and undefined meanings is clear for cooccuring features. The other is that words with undefined meaning often appear in the training data. The proposed

method is required further improvement, but it is effective as the technique of distinguishing a word with undefined senses.