Title	辞書定義文中の上位概念を用いた頑健な語義曖昧性解消
Author(s)	小川，千隼
Citation	
Issue Date	2005-03
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/1928
Rights	
Description	Supervisor:白井　清昭，情報科学研究科，修士

Robust Word Sense Disambiguation using Hypernyms in Definition Sentences in a Dictionary

Chihaya Ogawa (310024)

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

February 10, 2005

Keywords: Word Sense Disambiguation, Definition Sentences in a Dictionary, Hypernym, Ensemble of Classifiers, Robustness.

Word Sense Disambiguation(WSD) is the processing which decides the meaning of the word which appears in the text. At present, as the technique of WSD, the supervised machine learning using the corpus with a word sense tag is mainstream, but it has a problem which training data cannot always be gotten sufficiently.

As the technique which deals with such a problem, unsupervised learning using the corpus without word sense tags is proposed, however in this study by learning the probabilistic model which reflects hypernym of the meaning of a word extracted from the definition sentences in a dictionary, it improves the right answer percentage of WSD for the low frequency word. For example, a hypernym "the person" extracts from the dictionary definition sentences of one of word sense of the "author", "the person who wrote the sentences and paintings pictures and writings". Like this example, the word except for "author" which has a hypernym "the person", is easy to appear in the corpus, so by using the cooccurrence information of the hypernym and the context than the word sense and the context, there is possibility that it is possible to do learning well in the word which word sense itself doesn't appear too much in the corpus. It uses Naive Bayes model reflecting cooccurrence information of the hypernym and the context as the model for the low frequency word. Features using the model are a surface form and a part of speech of the word which exists just before and after the target word, the basic form of an content word in the sentence including the target word, and the basic form of the word syntactically related with the target word, etc.

Yagi which did the research which is near this research extracted hypernym using the EDR concept dictionary. However, because it specializes to process a machine, the EDR concept dictionary is for which it is simple and difficult to understand definition sentences in the dictionary. For example, the definition sentence of “the dog” of the EDR concept dictionary is “the animal which is called a dog”. On the other hand, the definition sentences of “the dog” of Iwanami Kokugo Jiten is “the brute with the dog department which is fond with the human being breeding as the livestock from the old days” and more information can be gotten about the dog. For the application which the quality of the definition sentence is esteemed, it is more desirable to do the use of the general Japanese dictionary for which it is easy to understand definition sentence in the dictionary. In this research, it uses the general Japanese dictionary where there are many expressions which are more useful for the person, specific Iwanami Kokugo Jiten.

Next, the technique to extract the hypernym of the meaning of a word from the definition sentence in dictionary is described. Generally, it is often that the word which is in the end of the definition sentence is the hypernym of the meaning of a word. Therefore, as the principle, it extracts the word of the end of the definition sentence as the hypernym. However, in the word of the end, it isn’t sometimes suitable as the hypernym. For example, the definition sentence of “拝借 (borrowing)” is “借りることをへりくだって言う語 (the word to say to borrow to humble)”. Because to make “語 (word)” of the end a hypernym from the definition sentence which ends with “NをVて言う語 (the word to say N to V)” isn’t appropriate, it applies the pattern to take out the part of N and it takes out hypernym “借りること (to borrow)” . It created such 116 hypernym extraction patterns by the hand and it extracted hypernyms from the definition sentence of Iwanami Kokugo Jiten. Then, out of the all word sense of Iwanami Kokugo Jiten, it succeeded in the extraction of the hypernyms of 97.25%.

Also, Iwanami Kokugo Jiten is different from the EDR concept dictionary. More than one definition sentences sometimes exists to one word sense. When making to take out hypernyms from each definition sentences, more than one hypernyms are extracted to one word sense. On the other hand, in the model using the cooccurrence information of the hypernym and context, it supposes that the hypernym of the word sense is one. Therefore, the best one must be chosen from more than one which was extracted hypernyms. Specifically, it classifies of after the second definition sentences into the 3 sentence type by the keyword of the hypernym which is extracted in the first definition sentence, and the first and last word of after the second definition sentences, and it chooses a hypernym according to the sentence type. There are 38 kinds of keywords to use for the type classification of above-mentioned process. It takes out random 200 word senses with multi definition sentences for the one word sense, it confirmed whether or not it was the one for which the classification type after the second sentence is appropriate by hand. Then, the classification types of 187(93.5%) word senses were appropriate. Therefore, it is possible to say that the precision of the classification of the definition sentence is high sufficiently.

In this research it finally combines two classifiers which are a classifier by Support Vector Machine(SVM) algorithm of the supervised learning model for the high frequency word and a classifier using hypernym for the low frequency word. The technique to combine is as the following.

- It chooses a classifier in the occurring frequency about the training data every word.

- It chooses a classifier by the right answer content rate about the held-out data every word.

- It chooses a classifier by the technique of the stacking.

  Using the technique of the stacking, it makes the SVM model classifier and the Naive Bayes model classifier which used a hypernym the first classifier, and it does WSD by learning the second classifier which makes their output features. The second classifier uses SVM as the learning algorithm. Also, it did in three ways of changing the training data and test data of the 1st classifier and the features in the second classifier.

Also, it creates the new classifier which uses the cooccurrence information of the word sense and the context and the cooccurrence information of the hypernym and the context at the same time, and it compared the new classifier with combining classifier.

Finally, it did the experiment which evaluates proposition technique. It does comparison of SVM, NB, BL (the baseline model) among the single classifiers, and does comparison of SVM+BL (the combination of SVM and the baseline model), SVM+NB (the proposed method), the simultaneous reflection NB model using the two cooccurrence information among the mixing classifiers. As a result, NB which is the proposed technique in this research didn't reach SVM but more than 2% of rises were seen at the F-measure compared with the BL model. Also, the model which has the highest the F-measure in the combined model of SVM and NB is the stacking using cross-validation, the model gives highest F-measure of all combined models.