Title	辞書の語義立てに基づく語義曖昧性解消に関する研究
Author(s)	玉垣，隆幸
Citation	
Issue Date	2004-03
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/1804
Rights	
Description	Supervisor:白井 清昭，情報科学研究科，修士

Disambiguation of Word Senses Defined by the Dictionary

Takayuki Tamagaki (110076)

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

February 13, 2004

Keywords: Word Sense Disambiguation, Word Sense Tagged corpus,Combination of Classifiers, Machine Readable Dictionary.

Word Sense Disambiguation (WSD) is one of the most important task in Natural Language Processing.

We need to build WSD system for Document Reading Assistant System for learners of Japanese. We make core of this system which classifies word sense.

We use two different language resources because the system needs to achieve high recall and coverage . One is annotated corpus with sense tagged for supervised learning . Annotated corpus is tagged with sense ID associated with dictornary's definitions.

Corpus-based classifier is built by the following step: Choosing the representation as a set of feature for the context of co-occurrence of the target words , and applying a machine learning algorithm to train classifier. In this study, we test several features with Support Vector Machine (SVM). We try several feature to examine to various range in this study. The followings are features used in experiments.

- Global feature

  POSs and base forms of content words within n words around a target word.

- Local feature

  The information of word position and base form for all words within $n$words around target words.

- Semantic class feature

  Semantic classes for content words of within n words around a target word. We attempt to optimize for semantic class feature as follows.

  - Vary figure of thesaurus ID
  - Case of target word having multi semantic classes ,or target word having only one class

Supervised classifier from sense-tagged corpus is generally high performance if enough training data is available. Sense tagged corpus, however, is inexhaustible and very costly. Furthermore, corpus based classifier has the data sparseness problem. The data sparseness problem, which is common for much corpus based work ,is especially severe for WSD. Because enormous amounts of text are required to ensure all senses of polysemous words are represent. To overcome the data sparseness problem, Machine Readable Dictionary (MRD) is also used in addition to sense tagged corpus.

MRD-based classifiers are consisted of two classifiers which refer to information in MRD. We make following two classifiers. Rule-based classifier :extract from grammatical information and implement the situation the which context of an input sentence match the rule . Example-based classifier:extract examples from lexicon usage from a MRD and build example database.

The similarity between an input sentence and the example database is measured with thesaurus .

We suggest the way of combination for three classifiers. The procedure is following. First, each classifier output in heldout data. The definition of accuracy is as follows :

$$\frac{The\,number\,of\,words\,which\,system\,outputs\,correct}{the\,number\,of\,words\,in\,heldout\,data}$$

.

Final result is outputted of the classifier of which the best accuracy in heldout .

Finally, applying the combination method ,accuracy goes down 8 %,F-maesure goes down 3 %,but recall goes up 2 % and coverage goes up 2 % compared with SVM classifier. The aim of this WSD system is to improve recall and coverage,so we achieved our goal.