JAIST Repository
https://dspace.jaist.ac.jp/

Title	Exploring Effective Features for Learning Vietnamese Word Sense Disambiguation Classifiers
Author(s)	グエン, ハイミン
Citation	
Issue Date	2010-09
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/9149
Rights	
Description	Supervisor:Associate Professor Shirai Kiyoaki, 情報科学研究科, 修士

JAIST
JAPAN
ADVANCED INSTITUTE OF
SCIENCE AND TECHNOLOGY

Japan Advanced Institute of Science and Technology

Exploring Effective Features for Learning Vietnamese Word Sense Disambiguation Classifiers

Nguyen Hai Minh (0810204)

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

August 10, 2010

Keywords: Word Sense Disambiguation, Vietnamese, Machine Learning, Feature for WSD, Pseudoword.

Natural language contains various kinds of ambiguity. One of those that attracted the most attention in computational linguistics is lexical ambiguity, because a word sense plays a very important role in natural language understanding. Although human can easily understand a meaning of a polysemous word in an actual text, it is very difficult for a computer to understand which sense of the word is used in the same context. The task of automatically determining senses of words in a certain context is called Word Sense Disambiguation (WSD hereafter). If the disambiguation of words are successfully solved, many natural language applications can take advantages from it, such as question answering, semantic analysis, information retrieval, machine translation, speech processing, etc.

WSD has been noticed since the earliest days of computational language processing in 1950s. So far, this problem has received a large variety of contribution in many languages such as English, Chinese, Basque, etc. The approaches vary according to the main source of knowledge used in sense disambiguation, such as dictionary-based methods, unsupervised corpus-based methods, supervised corpus-based methods and combinations of them. Among them, the supervised method is one of the most successful approaches in WSD though it is often suffer from the knowledge acquisition bottleneck problem.

Copyright © 2010 by Nguyen Hai Minh

Vietnamese is one of languages including highly frequency of ambiguous words. However, there is no published research of WSD in Vietnamese. Thus, it is essential to establish a method for Vietnamese WSD. Our research is the first attempt to study Vietnamese WSD. Since Vietnamese has some distinct characteristics compared to English, we need to know which features are effective for Vietnamese WSD first.

Therefore, the first goal of this research is to explore the effective features for disambiguate Vietnamese ambiguous words. In this study, we applied Support Vector Machines (SVM), which is a powerful machine learning algorithm for solving classification problem. However, there is no sense tagged corpus which is required for training the WSD classifiers. To tackle this problem, we applied 'pseudoword' method, a well-known technique that help us to collect sense tagged corpus automatically. In previous researches, this method is applied to evaluate WSD system when no sense tagged corpus is available. Nevertheless, a pseudoword is not an actual polysemous word. There is no evaluation on the capability of applying pseudoword for real WSD system either. Thus, our second goal is to explore the applicability of 'pseudoword' method for Vietnamese WSD.

In this research, we extracted many types of features which are Bag-of-Words (BOW), Part-of-Speech (POS), Collocation and Syntactic for training Vietnamese WSD classifiers. Bag-of-Words feature is a set of words appearing around the target word (ambiguous word) in a sentence. POS feature contains POS tags of some words surrounding the target word. Collocation feature includes 2-gram, 3-gram and 4-gram in many positions around the target word. As Syntactic feature, we extracted the words that have some syntactic relations to the target word. These features are extracted from Vietnamese TreeBank, a corpus contains around 10.000 sentences manually annotated with syntactic trees. Furthermore, a feature selection algorithm is used to automatically filter out ineffective features to improve the performance of WSD. We also consider some sets of feature combinations, such as 2-feature combinations and 3-feature combinations.

We investigated three tasks for evaluation. The first task is pseudoword task (PW task hereafter), which aims to determine pseudo-sense of a pseudoword in a given sentence. Pseudoword is a combination of two different monosemous words which played as its pseudo-senses. This pseudo-sense

tagged corpus can be automatically obtained without any human intervention. Although it is not a real WSD, pseudoword technique may appropriate for exploring the effectiveness of features for Vietnamese WSD. In many previous researches applying pseudoword technique to evaluate WSD methods, two monosemous words are chosen randomly. However, in this research, they are chosen considering the meaning of a certain word, similar to equivalent pseudoword proposed by Lu et al.

The second task is real word task (RW task hereafter), in which sense tagged corpus is manually constructed to evaluate a real WSD. Since PW task is obviously different with real WSD task, an ordinary WSD is required to investigate effective features more precisely. Furthermore, we can evaluate the applicability of pseudoword technique for WSD by comparing results between PW and RW task.

The final task is PW-RW task, in which WSD classifiers trained from pseudo-sense tagged corpus are applied for real WSD. It is possible since the target word set in RW and PW task is the same and each pseudo-sense in PW task corresponds to a sense in RW task. The attractive advantages of this approach is that no sense tagged corpus is required for supervised learning of WSD systems. PW-RW task can be used to evaluate the validity of pseudoword technique for disambiguation of Vietnamese words.

For each task, we conducted experiments for individual features and their combinations. We achieved the highest accuracies of 89.28% for verb, 91.77% for noun and 89.07% for adjective in PW task when only using individual feature. For RW task, they are 89.55% for verb, 91.34% for noun and 89.61% for adjective. All best results are achieved by classifiers using BOW feature. BOW feature clearly performs well for all three categories of target words (verb, noun and adjective). Its combinations with Collocation or Syntactic can improve the performance of WSD classifiers better than individual ones. Combine BOW with Collocation increase the accuracy to 90.29% for adjectives, BOW+Syntactic increase accuracy to 90.88% for verb, and BOW+Collocation+Syntactic increase accuracy to 92.91% for noun in RW task.

The best feature set in both PW and RW task is BOW for all categories of target words. On the other hand, the best feature combination in PW task is BOW+Collocation+Syntactic for verb, BOW+Syntactic for

noun and BOW+Collocation+Syntactic for adjective, while in RW task it is BOW+Syntactic for verb, BOW+Collocation+Syntactic for noun and BOW+Collocation for adjective. Thus, pseudoword technique can be applicable to find effective features, but not to find the best feature combination. Comparing the best individual feature for each target word, number of target words where the best features are same is 7 of 9 for verb, 5 of 9 for noun and 4 of 5 for adjective. For feature combination, 6 of 9 verbs, 2 of 9 nouns and 4 of 5 adjectives shared the best feature combination in PW and RW task. Therefore, pseudoword technique is also applicable to choose the best individual features and the best combination for each ambiguous word when the word is verb or adjective. However, it is inappropriate for noun.

In PW-RW task, the results are much worse than in RW task. The average of accuracies of the best classifier with individual feature when a target word is a verb is 79.98% in PW-RW task, which is much worse than 89.55% in RW task. For noun, it is 84.6% in PW-RW task, while 91.34% in RW task. For adjective it is 79.98% in PW-RW task, while 90.17% in RW task. It seems that WSD classifiers trained from PW corpus is not good enough for real words although two words of pseudo-senses are not randomly chosen but related with real senses. However, the accuracies for each target word are mostly higher than the most frequency baseline in 7 of 9 verbs, 8 of 9 nouns and 3 of 5 adjectives. It indicates that pseudoword might be a potential technique for WSD task for verb and noun when a sense tagged corpus is not available.

In conclusion, we have found that BOW feature is an effective feature for Vietnamese WSD. The best feature combination varies for individual target word. Maybe we could try choosing the best combination of features automatically when a target word is given. On the other hand, we discovered that although pseudoword technique is still not comparable to SVM classifiers trained from RW corpus, its results are acceptable to be applied in WSD task for verbs and nouns when we have no sense tagged corpus.

It would be interesting to verify that applying pseudoword technique in WSD is better than unsupervised WSD or not. This is one of our future works. Besides, we would like to investigate the effective features

for multi-class WSD classifiers in accompany with increasing the corpus size. Another interesting work is comparing the effective features between Vietnamese WSD and English to explore the differences and similarity between these languages in WSD task.