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

Title	Semantic Class Disambiguation for All Words
Author(s)	Truong Vo, Huu Thien
Citation	
Issue Date	2016-09
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/13744
Rights	
Description	Supervisor:白井　清昭，情報科学研究科，修士

JAIST
JAPAN
ADVANCED INSTITUTE OF
SCIENCE AND TECHNOLOGY

Japan Advanced Institute of Science and Technology

Semantic Class Disambiguation for All Words

Truong Vo Huu Thien (1410219)

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

August 04, 2016

Keywords: Word Sense Disambiguation, Semantic Class, Supervised Learning, Knowledge Acquisition Bottleneck.

Word Sense Disambiguation (WSD) is a task to automatically assign a sense from a predefined list of senses to a word in a particular context. WSD is one of the fundamental and important techniques used for many natural language processing applications such as machine translation, information retrieval, and opinion mining. Recently, supervised machine learning shows the best performance among various approaches of WSD. However, supervised learning still suffers a serious problem, called as "Knowledge Acquisition Bottleneck" or "Data Sparseness". That is, it requires a significant amount of texts annotated with the gold senses of the words as a training data. Since manual annotation takes a lot of cost and time, it is usually difficult to prepare enough data.

In traditional WSD approaches, classifiers are trained for individual target words since sense inventories are different for the target words. It is necessary to train a considerable number of classifiers to disambiguate the sense of all words in a text. To tackle this problem, an approach that pays attention to semantic class was proposed. The semantic class refers to an abstract concept of the words, such as 'plant', 'shape', and 'time'. In this approach, semantic class was used to be a common sense inventory in order to train universal classifiers that can determine the semantic class of any words, including the low frequent words. It can alleviate knowledge acquisition bottleneck problem, since a small number of the classifiers are required to train. Although the semantic class only expresses more general concept than the sense, semantic class disambiguation (a coarse grained WSD in other words) is obviously useful for many applications, such as information retrieval and acquisition of domain knowledge.

This paper proposes a novel method to disambiguate the semantic classes of all words. In the previous method of semantic class disambiguation proposed by Ariyakornwijit et al., the training data often consists of imbalanced positive and negative samples. The system trained from such data tends to classify a new input into the majority class. It causes decline of performance of semantic class disambiguation. Our new architecture enables us to train the classifiers from more balanced training data. In our architecture, a binary classifier is trained for each pair of the semantic class SC_i and SC_j. It chooses one of two semantic classes for the given target word. If the target word has three or more potential semantic classes, the classifiers of all possible pairs are applied. The final semantic class for a given new word is chosen by majority voting.

Copyright © 2016 by Truong Vo Huu Thien

WordNet is a famous lexical database, which is broadly used as a sense repository. The synsets, which are sets of synonyms defined in WordNet, are organized into forty-five lexicographer files. Each file defines a unique beginner of all synstes in it. In this research, the unique beginners, which are considered as the coarsest senses, are used as a set of the semantic classes.

In this method, monosemous words are used as the training data. They are words that have only one semantic class in WordNet. It is a promising way to create a large amount of the training data, since no manual annotation is required. However, it is rather uncertain that the monosemous words are useful for classification of ambiguous words. Because the words in the training data (monosemous word) are totally different from the target word (polysemous or ambiguous word). Such gaps may cause negative influence on semantic class disambiguation.

Support Vector Machine (SVM) is applied for training the semantic class disambiguation classifiers. Sklearn library is used to train SVM classifiers. SVM in Sklearn uses a kernel function to transform data in raw representation to feature vector representation. The kernel we use is the Gaussian radial basis function with two parameters: gamma and C. They are set as the default setting, i.e. gamma = 0.0001 and C=1000.

The features used in this research is exactly the same as Ariyakornwijit's method. They are local context, part-of-speech(POS), collocation, and syntactic feature. The local context features are the content words that appear 5 words before and after the target word. The POS features are 2-gram, 3-gram and 4-gram of POSs including the target word. The collocation features are sequences of 2, 3 and 4 words including the target word. The syntactic features are the dependencies between the words in the sentence. In this study, collapsed typed dependencies extracted from Stanford paper are used as the syntactic feature.

We evaluated our proposed method by two experiments. The Daily Yomiuri corpus, a large collection of newspaper articles, was used as the training data. On the other hand, the sense annotated texts in SENSEVAL-3 data was used as the test data. The gold senses in it were converted to the semantic classes. Seventeen ambiguous verbs were chosen as the target words. Because the dimension of the feature vector was huge, it took too long time to train the classifiers. Hence, we applied two feature selection method in our experiments: frequency based feature selection and Pearson's chi-squared test. In the experiment I, the number of the training samples were reduced to 10,000 per semantic class. These training samples were randomly chosen. Furthermore, frequency based feature selection was applied to select the best features. In the experiment II, the size of the Daily Yomiuri corpus was reduced to 20,000 sentences and we applied Pearson's chi-squared test based feature selection to reduce the dimension of feature vector.

In the experiment I, the proposed method (PW-SCD) achieved 41.2% accuracy on average of 17 target words. It was 1.4% better than Ariyakornwijit's method (OVR-SCD). In the experiment II, the average accuracy of PW-SCD was 46.5%, showing 7.3% improvement comparing to OVR-SCD.

The contribution of this thesis as well as knowledge newly explored by this study can be summarized as follows. First, the effectiveness of our proposed method was confirmed by the experiment. The accuracy of semantic class disambiguation was improved by 7.3% comparing to Ariyakornwijit's method. Second, although the proposed method

was required to train more classifiers, the computational cost and time to train each classifier was much less than the previous method, since the size of the training data was much reduced comparing to the extremely imbalanced the training data. Therefore, the proposed method was more likely to be practical and the reasonable for semantic class disambiguation. Third, the performance of the semantic class disambiguation highly depended on the target word. For some words, our method achieved better accuracy than the previous method, but not for other words. If we can choose an appropriate method for each target word, the overall performance will be much improved. Finally, due to the use of the monosemous words as the training data, the classifiers were trained from the training samples of the words that was different with the target word. Such a gap still remained as an important factor of low performance. In sum, the research showed a promising method to alleviate knowledge acquisition bottleneck in WSD, although it still has much room for improvement.