Word Sense Disambiguation Based on the Japanese Thesaurus Bunruigoihyo

Kento kobayashi (1610226)

School of Information Science, JAIST, s1610226@jaist.ac.jp

Extended Abstract

Word Sense Disambiguation (WSD hereafter) is a task to determine a meaning or sense of an ambiguous word in a context, where "ambiguous word" is a word that has multiple senses. In WSD, a set of senses of a word is usually defined by a dictionary or a thesaurus. One of the well-known Japanese thesauri is Bunruigoihyo. It compiles semantic categories that represent meaning of words. A semantic category in Bunruigoihyo is defined by enumerating words that has a meaning of the category. Since Bunruigoihyo is the most widely used for natural language processing of Japanese, a method to determine a semantic category of Bunruigoihyo for a given target word is valuable. However, only a few researches of this topic have been studied so far. This paper aims at developing a method of WSD that enable us to identify a semantic category of Bunruigoihyo for a given ambiguous word. Since supervised machine learning requires a labeled training data, which needs much human labor and is often difficult to prepare, this study explores a method based on unsupervised machine learning.

Our proposed method is based on a method proposed by Yarowsky. The Yarowsky's method consists of two steps: construction of sub-corpus and acquisition of the feature. At the first step, a sub-corpus is constructed for each semantic category. The sub-corpus is a set of sentences that include words registered the semantic category. A word of a semantic category is searched from a corpus, then 20 words before and after the searched word are extracted as a sentence in a sub-corpus. At the second step, a set of features of each semantic category is obtained from the sub-corpus of that semantic category. Here the feature of the semantic category refers to information representing a meaning of the category. In the Yarowsky's method, content words appearing in the context of the word registered in the semantic category are defined as the feature. That is, content words appearing in the sub-corpus are extracted as the features. Furthermore, a weight of each feature is estimated. It is an index measuring how saliently a feature represents meaning of a semantic category. A classification model for WSD can be trained by the above two steps. A sense (semantic category) of a target word in a test sentence is determined as follows. For each semantic category of the target word, a score of it is calculated by sum of weights of features, which appears in the context of the test sentence, of the semantic category. Then, the semantic category with the highest score is chosen.

This paper proposes three kinds of extension of the Yarowsky's method. The first extension is a use of the collocation feature. The collocation feature is a

sequence of words consisting of a target word and words appearing just before or after it. It is known as an effective feature for WSD. While the Yarowsky's method uses only content words appearing in the context as the feature of the semantic category, our proposed model uses the collocation feature too. It can improve the accuracy of WSD especially for verbs.

The second extension is to use only monosemous words for construction of a sub-corpus of a semantic category. Polysemous or ambiguous words are not disambiguated at the training phase in the Yarowsky's method. It causes a problem that many incorrect features of a semantic category are obtained. This research aims at preventing from obtaining wrong features by using only unambiguous words for training.

The third extension is to increase the amount of the training data by a bootstrapping technique. Its procedures consist of three steps. First, WSD model is trained from the training data of only monosemous words. Second, polysemous words in the training data is disambiguated by the WSD model. If the reliability of WSD is high enough, the disambiguated word and its context is added to the training data as a sentence of a monosemous word. Third, a new WSD model is trained from the increased data. The amount of the training data is gradually increased by repeating the above procedures. We call this method "selecting-reliable-sense model". In addition to this, we also propose a method that does not determine a sense of an ambiguous word but remove unreliable (probably incorrect) senses of an ambiguous word in the second step of the bootstrapping process. It is referred as "removing-unreliable-sense model" hereafter.

Several experiments to evaluate the proposed method was carried out. A part of Balanced Corpus of Contemporary Written Japanese (BCCWJ) annotated with gold semantic categories of Bunruigoihyo was used as a test data. The number of the target words in the test data was 3,912, including 2,467 nouns, 1,175 verbs and 270 words of other parts-of-speech (POSs). The accuracy of the Yarowsky's method on this test data was 51.9%. The proposed WSD model using the collocation feature achieved 51.7% accuracy. Although the accuracy on the overall test data slightly declined by adding the collocation feature, the accuracy of the extended model for verbs was 44.3%, which was much higher than that of the Yarowsky's model, i.e. 40.9%. The accurary of the WSD model trained from monosemous words only was 56.7%, which outperformed the Yarowsky's model by 4.8 points. Next, two bootstrapping methods to increase the training data were evaluated. "Selecting-reliable-sense model" outperformed the model trained from monosemous words by 1.8 points, while "removing-unreliable-sense model" outperformed it by 1.0 points. Since these results indicate that the accuracy of the proposed method is better than that of the Yarowsky's method in most cases, we can conclude that our method is effective to disambiguate semantic classes of Bunruigoihyo of words.