Combining Example and Statistical Based Methods of Word Sense Disambiguation towards Reading Assistant System

Pulkit Kathuria (1010205)

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

February 6, 2003

Keywords: Word Sense Disambiguation, Example Based Method, Examples Expansion, Parallel Corpus, Japanese.

Japanese learners often look up words in dictionaries/internet when they read Japanese documents. One word has several possible translations, although a word has only one meaning when it appears in the document. It is rather hard for non-native readers of Japanese to read definition sentences of all meanings. It would be useful to build a system which can not only show the target word's definition sentence in English and its example usage but also select the correct meaning. In this research we focus on building a reading assistant system that aims to assist Japanese language learners to properly understand the meanings of Japanese words in context.

Currently, ASUNARO is the only reading assistant system for Japanese learners with Word Sense Disambiguation (WSD hereafter) module. However, definition sentences of EDR dictionary produced by ASUNARO are sometimes unnatural and no example sentence is shown for each sense. There exists a more appropriate dictionary for language learners. For example, EDICT, the Japanese-English bilingual dictionary that includes definition sentences in English as well as example sentences in Japanese and English. We believe that example sentences are indispensable for Japanese learners to understand meanings of words.

The goal of this research is to develop a novel reading assistant system for Japanese learners that shows both definition and example sentences for a target word. It also disambiguates the sense of the target word in the context, then only show the information about the correct meaning. In this research, example-based WSD is considered, since it would be suitable for our reading assistant system showing examples to users. It should handle all words, including low frequency words, in a document. Therefore our WSD method does not rely on a sense tagged corpus that requires much human labor to construct. In this thesis, we present the proposed interface of the reading assistant system, example based WSD, machine learning, combination of example and machine learning classifiers and automatic acquisition of example sentences using parallel corpora.

First, we develop an example based WSD classifier that calculates a similarity score $sim(I, E)$ between the input Japanese sentence (I) and example sentences (E) from a dictionary. Then choose example sentence where the measured similarity score is greater or equal to a threshold. Greater the threshold is set, higher the precision is expected with a loss of recall. In order to choose example sentence (E) which is most similar to an input sentence (I), we build an example based classifier which measures overall similarity $sim(I, E)$ as a sum of collocation similarity $coll(I, E)$ and similarity calculated by comparing syntactic relations $syn(I, E)$. $coll(I, E)$ is based on match sequences of n-grams of sizes 4, 5 and 6 between sentences I and E. While $syn(I, E)$ is calculated by comparing syntactic relations extracted from *bunsetsu* dependencies for a target word and common words. Separately, we build two example based classifiers with two kinds of syntactic relations $syn(I, E)$. The first one is RTW, which considers syntactic relations only with respect to target word. The other is RCW, in which syntactic relations with respect to both target word and common words are considered.

The proposed example based classifier focus on achieving high precision rather than recall. Therefore, we also train a robust machine learning classifier based on Robinson's method. Both classifiers compensate each other upon combination. The final WSD classifier is an ensemble of example based and Robinson classifiers. First, precision oriented example based classifier is applied. When it can not choose a sense, a sense from Robinson

classifier is chosen.

Although, EDICT has 150,000 example sentences for 160,000 lexical entries, the number of example sentences might be enough for high frequency words but as we move towards rare senses or low frequency words there are a few or no example sentences available. It is difficult to disambiguate a sense in such cases and moreover no example usage can be shown to the user in the reading assistant system. To tackle this problem, we automatically extract bilingual example pairs for each sense in the dictionary from parallel corpora. The sense of the target word in the example pair is verified by three heuristics. First we check if the target word appears in the Japanese sentence and it sense definition in English, with a constraint that both should be aligned to each other. Further two restrictive constraints are appointed, that content words in the description of a sense definition must also appear in the English sentence and same word in two sense definitions must not match. We used two sentence aligned parallel corpora which consists of 582,005 Japanese-English sentence pairs. It enables us to improve the performance of WSD classifiers as well as prepare more examples to be shown to users.

To evaluate the performance of WSD, we manually prepared two sense tagged corpora, a development and an evaluation data. We first built the development data to design our example based WSD method and optimize a threshold. It consists of 330 input sentences of 17 target words (8 nouns, 8 verbs and 1 adjective). Then, another sense tagged data is built as the evaluation data to measure performance of our proposed method. It consists of 937 input sentences of 49 target words (23 nouns, 24 verbs and 2 adjectives).

On automatic example sentences expansion, number of example sentences are increased by about 3 times and covered 53% of senses for 49 target words. Furthermore, number of senses with no example sentences decreased by 18.5%. Senses with no example are crucial for our WSD method since such a sense is never chosen. And no example sentences can be shown in the reading assistant system either. Accuracy of automatically expanded example sentences is 85% on a manual evaluation on randomly selected samples. The constraints designed for WSD heuristics to extract examples effectively prunes false candidates. Automatic expansion

of example sentences help to improve the WSD accuracy on the evaluation data. It improved the accuracy of example based classifiers by 10% and machine learning method (Robinson classifier) by 9%. The proposed WSD method, combination of example based method with statistical based method (Robinson classifier), achieved 65% accuracy which is 7% greater than the baseline. Furthermore, the accuracy of the combined model was better than single example based classifier and Robinson classifier. It indicates that our approach to combine two types of classifiers is effective.

Any machine learning algorithms can be combined with the proposed example based classifiers. Robinson and other machine learning based classifiers (Naive Bayes, Maximum Entropy, Decision Tree and Support Vector Machine) are empirically compared in terms of performance of both single and combined models. Robinson was one of the best classifiers, although there were no significant differences among these machine learning algorithms.