Unsupervised Word Sense Disambiguation Based on Word Embedding and Collocation

1810410 Han Shangzhuang

Word Sense Disambiguation (WSD) is a long-standing problem in Natural Language Processing, which aims to determine a sense for a target word in a given context. WSD plays a important role in downstream tasks of NLP, such as Machine Translation, Named Entity Recognition (NER), and chatbot. Approaches for WSD can be grouped into two main categories: methods based on supervised machine learning (supervised methods) and knowledge-based methods (unsupervised methods). Supervised methods often train a classifier or neural network model from a sense tagged corpus. Although supervised methods tend to achieve good performance, sense tagged corpora are required for training. Obviously, they are hard to construct due to heavy manual annotation. Knowledge-based WSD methods rely on lexical resources rather than sense tagged corpora. A gloss, which defines a meaning of a word in a dictionary, is first utilized in Lesk algorithm. Given a word and its context, Lesk algorithm calculates a score of each sense by measuring the number of overlapped words in a gloss (definition) of a sense and in a context. Then, the sense with the highest score is chosen. A lot of studies follow it and propose its extended models. One of the advantages of knowledge-based WSD methods is that it can be developed with existing lexical knowledge or database. Since a sense tagged corpus is not required, it can be easily implemented with low costs. However, the performance of the WSD tends to be lower than supervised approaches.

Word embedding is abstract representation of words in a form of an n-dimensional vector. It can be pre-trained from a large amount of corpus, then can be used for solving various NLP tasks. With the development of word embedding, some researchers propose other methods to compute the text similarity using word embedding instead of calculating overlaps in Lesk algorithm. For example, Basile et al. propose an unsupervised WSD algorithm which extends the Lesk's WSD method. In Basile's method, the word similarity in the semantic space is regarded as gloss-context overlap. Three steps are required to determine a sense of a target word: (1) to construct a context vector $\vec{c}$, (2) to construct sense vectors $\vec{s_i}$, and (3) to calculate the cosine similarity of two vectors to choose the sense. The context vector is obtained by averaging vectors of all context words in a context. Pre-trained word embedding is used as word vectors $\vec{w_k}$. Similarly, the sense vector is constructed by averaging the word vectors in a gloss sentence. Finally, the sense whose vector is the most similar to the context vector is chosen.

The goal of this thesis is to propose a novel unsupervised WSD method. We extend the Basile's method in two directions. One is to incorporate a mechanism to determine a sense using a collocation. Rules to determine a sense, which are based on collocations, are automatically acquired from a raw corpus, then these rules are integrated to the Basile's WSD model. Two types of the collocation are considered in this study. The first one is a word collocation that is a sequence of words including the target word. The second collocation is a dependency collocation, which is defined as a pair of words under a certain syntactic dependency. In this study, we call the rule "collocation WSD rule". The other extension is that we investigate the better way to construct the context vector in the Basile's method. Our basic idea is to use not all words but only words that are highly related to the sense for the construction of the context vector. We defined an indicator called *RelevantScore*, which evaluates how a word is strongly related to a particular sense. In this study, *RelevantScore* is obtained by the cosine similarity of the word vector (word embedding) and the sense vector. Then, the new context vector is made by averaging word vectors of words with high *RelevantScore*. Hereafter, we call the WSD method considering *RelevantScore* the Highly Related Word Embedding method (HRWE).

In our method, we use the HRWE and some filtering methods to extract high-quality collocation WSD rules from a raw corpus in advance, and store them in a database. For an ambiguous word in an unlabeled sentence in a row corpus, the sense of it is determined by the HRWE. If the reliability of the chosen sense is high, candidates of the collocation WSD rules are obtained by applying the pre-defined templates. Finally, rules are filtered out if they consist of only function words, do not frequently occur, or are inaccurate for WSD.

In our proposed system, when performing WSD, the sense of a target word is determined by the collocation WSD rule if the collocation in the rule is found in the context of the target word. Otherwise, the HRWE is used to determine the sense of a target word. The several ideas to improve the WSD performance are newly introduced in it: to construct the context vector with only contextual words that are highly related to the sense (HRWE), to acquire and combine the collocation WSD rules, to use two types of collocation WSD rules (word collocation and dependency collocation), and so on. Several experiments are conducted to discuss the contribution of each component in our method through the comparison of different WSD models.

The dataset of Senseval-3 English lexical sample task is used in the experiment. In addition to the test dataset, we also need two external data, one is a lexical resource as a sense inventory, the other is a raw corpus for extraction of the collocation WSD rules. We use WordNet and Leipzig corpus

as our sense inventory and raw corpus respectively.

Comparing the precision of WSD on the dataset, the HRWE outperforms the baseline for nouns and verbs, but not for adjectives. However, the precision is improved by 3.2 point for all POSs. It indicates that our idea to select contextual words strongly associated with senses for construction of the context vector is effective. The system using only the collocation WSD rules tends to achieve the higher precision than other systems with low applicability, especially for nouns and adjectives. It is confirmed that we can obtain the disambiguation rules whose recall is low but precision is high as we aimed. The performance of the systems integrating the HRWE with the word or dependency collocation WSD rules is better than the HRWE only, which means the collocations can contribute to choose the appropriate sense. Our final system, the WSD system with the HRWE and both the word and dependency collocation WSD rules, achieves the best performance for nouns, verbs and all POSs. Its precision is 0.572, which is 4.7 point better than the baseline.