Word Sense Disambiguation

Yifan Bai (260562421)

Introduction

In this project entails an experiment on word sense disambiguation. We compare the performance of a traditional dictionary matching method with the word sense disambiguation model of NLTK, and their advantages and disadvantages.

Datasets

For word sense disambiguation, the corpus we use for training and verification is called *multilingual-all-words.en.xml*, which is provided by the project itself. After the XML file is loaded, training set data set and verification data set will be generated directly. In addition, in the word sense disambiguation method based on WordNet, we need to use WordNet dictionary, using a file called *wordnet.en.key*.

similarity_heuristic_dict: In the heuristic method, we need to use the data file to store some similarity related data.

Methods

(1) most_frequent_word_sense_baseline: The method is to realize word sense disambiguation with the help of thesaurus, and find out the exact word meaning by comparing in the word list.

(2) *from nltk.wsd import lesk*: This method is based on the NLTK's own implementation of the word sense disambiguation method *lesk*, to verify the performance of the methods provided by NLTK. It can be used directly such as from *nltk.wsd import lesk*.

(3) Heuristic: The heuristic method is to determine the word sense disambiguation by specifying a certain calculation rule, and use a variety of external similarity data to conduct heuristic processing.

Experiments

Using the above three methods, several experiments were carried out, in which WordNet dictionary and test data set were introduced. The experimental results are as follows. From the results, we can see that the heuristic method has a certain improvement on the performance of the model, and the built-in function of nltk has better performance than the method implemented by ourselves.

Method Name	Accuracy (%)
Most frequent	0.65
Lesk	0.73
Heuristic	**0.85**
path_similarity	0.82
lch_similarity	0.81
wup_similarity	0.83
Res_similarity	**0.85**
Jcn_similarity	0.76
Lin_similarity	0.75

Table 1 – Accuracy Comparison

Discussions

There are a couple of takeaways. Firstly, the function implemented by ourselves may not be strict in logic and the data set is not large enough, so the performance is lower than that of NLTK. Secondly, through a certain heuristic function design, the performance of the model can be improved to a certain extent, but a good heuristic is needed as a bad one will reduce the performance of the model.