A Feature Extraction Method Based on Word Embedding for Word Similarity Computing

Weitai Zhang, Weiran Xu, Guang Chen, and Jun Guo

Beijing University of Posts and Telecommunications, Beijing 100876, China

Abstract. In this paper, we introduce a new NLP task similar to word expansion task or word similarity task, which can discover words sharing the same semantic components (feature sub-space) with seed words. We also propose a Feature Extraction method based on Word Embeddings for this problem. We train word embeddings using state-of-the-art methods like word2vec and models supplied by Stanford NLP Group. Prior Statistical Knowledge and Negative Sampling are proposed and utilized to help extract the Feature Sub-Space. We evaluate our model on WordNet synonym dictionary dataset and compare it to word2vec on synonymy mining and word similarity computing task, showing that our method outperforms other models or methods and can significantly help improve language understanding.

Keywords: word embeddings, feature sub-space, negative sampling, word similarity, prior statistical knowledge.

1 Introduction

Word similarity in NLP is a task of computing the similarity between two or more words by certain methods. In general, similarity always means the semantic similarity of words, for example, “apple” and “pear” have very close relationship, while “mike” and “class” are not relevant. Researchers improved the performance of word similarity task within methods like brown cluster, topic model, vector space model and so on[1][2]. Recently, word representation, mostly word embedding, is proved to be excellent at word similarity task [3].

In this paper, we focus on a new NLP task similar to word expansion task or word similarity task which could reveal words having the same semantic components with given seed words. The differentia between this task and word similarity task lies in that (1) it reveals words through more than one word and (2) the key point is how to represent the same semantic components of the seed words. We propose a method combining words’ syntactic information gained with state-of-the-art methods and representation of the same semantic components of the seed words based on word embeddings.

Most words have more than one meaning, which leads to that certain aspects of some words may share the same semantic information. One specific example given here is that “Beijing”, “Shanghai” and “Tokyo” have the same facet that they are all

C. Zong et al. (Eds.): NLPCC 2014, CCIS 496, pp. 160–167, 2014.

city names. Another example is that “sad”, “sorrow” and “low” also have a same meaning of sad or upset in mood. On condition of research needs and actual situations, we always need to reveal more words having the same semantic components with the given seed words which is exactly our problem. Like, given “Beijing”, “Shanghai” and “Tokyo”, we can find that “Guangzhou”, “Houston”, “Osaka” are also city names; given “sad”, “sorrow” and “low”, we may find that “upset”, “depressed” also have the meaning of sad or upset.

The structure of this paper is as follows. Section 2 describes Prior Statistic Knowledge proposed. Section 3 describes our method and gives structure of our model. Experimental results are presented in Section 4. Finally, conclusions are made in the last section.

2 Prior Statistical Knowledge

Researchers of deep learning in natural language processing believe word embeddings can represent syntactic information or semantic information [9]. However, current progress shows that word embeddings or neural language model may not outperform state-of-the-art methods in some tasks. For example, Brown Cluster is superior to the word embeddings on NER task [11] and there is only a small difference between Brown clusters and word embeddings on chunking task.

The effectiveness of word representation plays a significant role in our model. To more precisely represent a word, we propose Prior Statistical Knowledge of words to enrich the representation of words and compute using the published open source tools from Stanford NLP Group[10].

Table 1. Samples of labels and features in Prior Statistical Knowledge

Label name	No.	Feature name
POS	29	vb, cc, jjs, prp, in, nnp...
NER	3	person, location, organization
Parsing	94	cc_post, tmod_post, prt_pre, cop_pre ...

As showed in table 1, we assign each word with 3 most important labels: Named Entity Recognition, Part-Of-Speech tagging, Parsing Dependencies. According to Stanford CoreNLP tools and actual situation, we choose 29 features for POS and 3 features for NER. Particularly in Parsing Dependencies part, there are 47 kinds of ternary relation pair in total, so we extract 94 features for Parsing Dependencies, doubled because of the position each word exists. In Parsing Dependencies, features ended with ‘_pre’ indicate that words are in the front of ternary relations and ‘_post’ corresponds to the back of ternary relations. For example, in ternary relation