國立政治大學資訊科學系

Department of Computer Science
National Chengchi University

碩士論文
Master's Thesis

論文中文題目

English Thesis Title

研 究 生：陳ＸＸ

指導教授：蔡銘峰

中華民國 一百零五 年 十一 月

November 2016

105
碩士論文

論文中文題目

政治大學資訊科學系

陳ＸＸ

論文中文題目

English Thesis Title

研 究 生：陳ＸＸ　　　Student：ＸＸ Chen

指導教授：蔡銘峰　　Advisor：Ming-Feng Tsai

國立政治大學

資訊科學系

碩士論文

A Thesis

submitted to Department of Computer Science

National Chengchi University

in partial fulfillment of the Requirements

for the degree of

Master

in

Computer Science

中華民國 一百零五 年 十一 月

November 2016

English Thesis Title

Abstract

The paper demonstrate the popular method in recent years to construct the semantic embedding, and use classification to verify the accuracy of these models on Chinese.

目錄

圖目錄

表目錄

第一章

緒論

1.1 前言

How to make the sentence embedding with its own semantics more precisely is study of interest,since it's beneficial for several NLP tasks like machine translation, sentiment analysis. Since the internet text volume grows so enormously and rapidly, how to make the information can be extracted more efficiently and precisely become more critical for many application. Chinese forums, blogs or microblog expand especially rapidly, and the articles and the posts are produced.

Recently word2vec,m, Mikolov et al.(2013a), is considered to work for evaluated word semantics. Additionally, the character is invariant to the language. Nevertheless, the embedding in sentence level is more complicated, it's related to the sentence structure, intention or context. There is several methods raised in recent years, like Siamese-CBOW, FastText ...etc. Most of them is able to train batch of text and construct the vectors.

1.2 研究目的

So far, most the studies are conducted in English, we are also interested if the feature also works in Chinese or other languages, and if the algorithm is invariant to the language grammar.

第二章

相關文獻探討

2.1　中文測試

第三章

研究方法

The data set we chose is Open WeiboScope, which is collected WeiBo randomly by researchers at the Journalism and Media Center of the University of Hong Kong in 2012. It contains 226 millions posts distributing over the year. We used the tags in post as the indicators of sentiment,and removed some duplicated posts or some posts without any tags, or too many tags. We evaluated the accuracy of the classification for different algorithms. We used the TF-IDF () and SVM (Joachims, 1998). as baseline.

For the data prepossessing and cleansing, it's a Weibo feature to allow the user to use emoticon, and the emoticon in raw data expressed as [笑](smile),[淚](tear).

we removed the posts that contains too many tags, or without any tags. We also removed the duplicated posts by their post id roughly because it is a property of Chinese microblog [cite] for Chinese netizens to post repeatedly, but most algorithms can't resist the duplicated posts. Besides, we only chose the post that over certain length .

The posts meets the criteria is about 7.4 millions. And we removed the tags in the original post, and there are so many tags , we use most-used 6 categories to categorize them as Figure 1.

3.1　模型簡介

1. TF-IDF + SVM

The conventional way to evaluate the semantics based on the occurrence of words and term, and it also takes the occurrence of word in global context into consideration.

It's simple and effective, but it still suffers from some disadvantages like data sparsity and high dimensionality.

2. FastText

The structure of FastText is similiar to CBOW of Mikolov et al. (2013), and it uses the softmax to compute the probabilities for predefined classes. The word representation is looked up through a table and finally averaged into the text representation. Finally it uses the linear classification.

3. Paragraph vector

This is raised in [Quoc Le and Tomas Mikolov. Distributed Representations of Sentences and Documents.]

We use the implementation of Gensim and use SVM with linear kernel to classify.

4. Siamese-Cbow

The method computes a sentence embedding is to average the embeddings of its constituent words, instead of using pre-trained word embedding.

We used the implementation (https://bitbucket.org/TomKenter/siamese-cbow/overview) from the author, and made it compatible with python3 for better compatibility with unicode.

第四章

實驗結果與討論

4.1 實驗設定

第五章

結論

The result shows that FastText can archive better accuracy

5.1 結論

For the baseline, though TF-IDF it can archive the accuracy about 0.44(+-0.04). The most distinguishable features they use are some rarely used terminology. Since we only removed the duplicated post roughly, it may still suffer from the duplicated post from different sources with certain rarely-used words.

Generally, FastText can get the better accuracy , even converting the posts to pinyin, it can also achieve the same accuracy. Though, we tried the different settings for FastText, the accuracy didn't differ so much. And it also took much less time than other algorithms to complete.

The Siamese-CBOW, the performance is below the baseline. We tried evaluate the model it trained, it seemed it's converged enough. The word embedding is not converted correctly.