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
Sentence Reprentation in Chinese

研 究 生：管芸辰
指導教授：蔡銘峰

中華民國 一百零五 年 十一 月
November 2017

105
碩士論文

論文中文題

政治大學資訊科學系

管芸辰

論文中文題

Sentence Reprentation in Chinese

研 究 生：管芸辰　　Student：Yun Chen Kuan
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 2017

Sentence Reprentation in Chinese

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.

Content

Figure Content

Table Content

Chapter 1

Introduction

1.1 Abstract

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[6] is considered to work for evaluating word semantics in general cases. 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 are 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 Purpose

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.

Chapter 2

Related Work

Chapter 3

Methods

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 preprocessing 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 [2] 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.

We used jieba and dictionary to segment to post.

3.1 The model introduction

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 dimension.

2. FastText [3]

The structure of FastText is similar 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 [5]

This is method is purposed in [5]. The idea is obtain the summary of paragraphs,sentences or documents.

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

4. Siamese-Cbow[4]

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.

Chapter 4

Conclusion

4.1 Experiment Settings

We used baseline TD-IDF plus SVM with linear kernel as baseline. We compared it with other models with different settings.

For PVDB, we use 3 different models dm/c,dbow and dm/m. Additionally, we also tried to convert data set to pinyin to evaluate if the pinyin improve the sematic recognition for FastText ,which support vocabulary expansion with subword information [1].

Table 4.1: Results

Tf-IDF	0.44+-0.04
PVDM	
FastText	0.5
FastText(Pinyin)	0.5
Simaese-CBOW	0.41 (+/- 0.04)

Table 4.2: FastText

	8	12	16	32	64
no segmentation	0.369	0.375	0.389	0.372	0.368
segmentation	0.515	0.515	0.514	0.516	0.513
segmentation + pinyin	0.513	0.518	0.516	0.517	0.51

Chapter 5

Discussion

The result shows that FastText can archive better accuracy in general way.

5.1 Discussion

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. In general, the model is not generalized enough, it may not be applicable when the data set changed.

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 is not converged enough. The word embedding is not converted correctly. In the original paper, the dataset they used is Toronto Books, which contains novels, so the sentences may be more coherent with previous sentence and next one.

We demonstrated the various modern methods on the Chinese corpus, and it indicated that some models like FastText are invariant to language property. In general, most models improve the sematic analysis

Reference

[1] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*, 2016.

[2] K.-w. Fu and M. Chau. Reality check for the chinese microblog space: a random sampling approach. *PloS one*, 8(3):e58356, 2013.

[3] A. Joulin, E. Grave, P. Bojanowski, M. Douze, H. Jégou, and T. Mikolov. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*, 2016.

[4] T. Kenter, A. Borisov, and M. de Rijke. Siamese cbow: Optimizing word embeddings for sentence representations. 2016.

[5] Q. V. Le and T. Mikolov. Distributed representations of sentences and documents icml. 2014.

[6] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. pages 3111–3119, 2013.