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 computer can operate the sentence with its own semantics more precisely is study of interest. 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 grow especially rapidly.

Recently word2vec is considered to work for evaluated word semantics. Additionally, the character is invariant to the language. Nevertheless, the problems 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 WeiboScope, which is collected from 2012 WeiBo randomly. 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.

3.1 模型簡介

1. TF-IDF

The conventional way to evaluate the semantics based on the occurrence of words and term, and it also take 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

第四章

實驗結果與討論

4.1 實驗設定

第五章

結論

5.1 結論

中文測試