COMP9318

Data Warehousing and Data Mining

Programing Project Report

Z5004850

Tengyu Ma

Overview

In this project, we are asked to use logistic regression to classify name entity in an article. Generally speaking, a 'TITLE' word could be considered as formal salutation of a person, which could be the occupation title or abbreviation title. Following report will discuss the method for feature extracting and accuracy improvement.

Feature extraction

After reading the original file extracted from training data, some basic features could be easily being found. But some of the features could not be treated as 'law' during classification. For example, most of 'TITLE' words are following a 'DT' type word. In fact, in English language, most of odd noun words may have article words ('a', 'the'). If we take this feature into consideration, the weight of this feature in our classifier will be negative as for most language situation, this feature shows a noun with article word is not a 'TITLE' word.

Characteristic is another significant feature in an English article. Most of the 'TITLE' words are nouns especially 'NNS' which represent proper noun. But same as the feature we discussed in previous paragraph, it shows more like a negative feature in general language environment. Those fact shows that 'TITLE' word cannot be well classified by observed features we should focus more on relationship between word in context. After several attempt, following features had defined in this project:

Occupation List: Some of 'TITLE' words are formal occupation form, for example, Prime Minister and CEO. So an occupation list[1] was added into the program. This is not a feature, but it influences the results a lot, and plays important role in other features. This will be discussed in second part "Improvement".

Head word: This feature works with occupation list, some 'TITLE' word, for example, 'senior lecture' is a combination of head word and a word in occupation list.

Abbreviation: Abbreviation is another important feature for 'TITLE' word. Words like 'Dr.'

---

[1] Career Center, 2016 Department of Training and Workforce Development Western Australia

http://www.careercentre.dtwd.wa.gov.au/Occupations/Pages/OccupationsA-Z.aspx

and ‘Fr.’ are more like ‘TITLE’ words.

Context: This feature was for context judgement. It stores every word’s previous and next words as feature. This feature could classify some set phrases which is always the ‘TITLE’ word in English.

Phase: This is little bit like context feature but it focusses on continuous occupation listed words. It records the number of continuous occupation words. In fact, a single occupation word is not the evidence of ‘TITLE’ words, but multiple occupation words, in my opinion, has a larger probability to be ‘TITLE’ words.

Previous/Next n words: This feature records the previous/next n word and add this into feature vector of each word during extracting Phase feature. It seems the same with context feature, but it focuses more on occupation words.

Characteristic of the word: Characteristic is an important feature for classifying. But we could not simply say noun is ‘TITLE’ or any non-noun word should not be ‘TITLE’. So only way to store this feature is save the feature itself.

Words in training set It is important to save the words in training set as a feature of classifier. It is the basic reference of the classifier. Without this library, classifier could hardly generate a good result.

There are still many features I attempted during experiment, but most of them shows no help in improving F1 score of the classifier. These feature is mostly about the characteristic analysis. This proved that characteristic analysis seems didn’t work for this project. The best way to use this feature was adding characteristic as a feature without preprocessing.

Improvements

Due to my limited English skill, I could not generate very good test cases, the annotation accuracy for my test case was really bad. So I down load test data from the web site and write a program to check F1 score myself. Every single improvement I made was depend on the F1 score of my classifier. Actually, the accuracy gave very small help during my test as it was easy to improve the TN. But the problem of my classifier was always a high rate for FP.

As I use an occupation list as reference, It is important to expand the list as much as we can. The size of the list significantly influences the result. I tried a smaller list in previous experiment,

and the result was not very good. Furthermore, some occupation title consists multiple word. It is important to expand these words into job list because the continuous of the occupation word is an important role in classification.

I also found naming the feature is another very important part. For example, I used to use the characteristic as the feature name, but when I add a prefix ‘type_’, the classifier showed higher F1 score. I think it was because some word in the training data just has the same name with word characteristics, such as NN and all symbols in training data. So I gave every feature an very unique name in order to make sure it is not same with words in training set.

Another thing I found was although some feature actually doing the same thing during feature extraction, you cannot name these features same name. For example, the ‘head’ feature and ‘previous/next word’ feature actually did the same thing in my program, but one of them was a general feature, the other one is for occupation words. Using the same name could lead to mixture of different feature, and influence the result.