A study on sentiment classification with LSTM+GRNN and Sub-tree Mining

CHAU, Ngoc Phuong (1410217)

School of Information Science,
Japan Advanced Institute of Science and Technology

August 4, 2016

Keywords: Sentiment Classification, Natural Language Processing, Data Mining, Deep Learning, Machine Learning.

Recently, with the development of the online social network such as Facebook, Twitter and LinkedIn, sentiment classification (SC) which determines opinions of people is a significant task in natural language processing. This task is a part of computer scientists to make computer "think" as human.

In this thesis, the two main ideas are the representation of the semantic relation among words in a sentence and how to determine and eliminate noises in the document to improve the accuracy in sentiment classification task. The first idea is that we study the structures of the texts according to semantic relations in dependency tree and find a method for eliminating noises in the dataset. Additionally, The FindBestSub-tree was proposed to find out and remove the noises in the dataset to improve the performance in sentiment classification problem.

In our work in sentiment classification, we propose a model which combines deep learning and sub-tree mining to resolve sentiment classification problem. Stanford Parser is used to extract the relation from the beginning to the end of the sentences and a tree represent for a sentence. Afterwards, FindBestSub-tree algorithm with sub-tree mining technique eliminates outliers in the dataset. Then, the order of the words in a sentence changes according to DFS (Depth First Search) from a tree after outlier removal phase. Finally, the association between all words in a sentence and all sentences in a document is captured by LSTM and GRNN, respectively.

We conducted document sentiment classification experiment on multi-domain sentiment

dataset. The elimination of outliers leads to higher performance in this model[2]. In our experiment, the proposed method achieves improvements in term of accuracy in a range of 0.14% - 6.93% over LSTM+GRNN model.

In the current research, the performance of our model was improved with multi-domain sentiment dataset. In the future, we will represent a document by a semantic graph and explore more informative features to improve semantic relation in the document. With the information from the semantic graph, we can extract and select significant features for sentiment classification problem. Another future work is that we can build a deep learning model to capture features from semantic graph automatically.

[2]Code and Datasets are available at
https://github.com/ph0123/sub-tree-mining-and-deep-learning