Graduate School of Science and Technology Master’s Thesis Abstract

Laboratory name (Supervisor)	Natural Language Processing (WATANABE TARO　(Professor))		
Student ID	2111313	Submission date	2023 / 7 / 24
Name	LIU CHANG		
Thesis title	Code-Switching Method for Low-Resource Language Sentiment Analysis on Multilingual Pre-training Models		

Abstract

With the development of the Internet, especially social networks, people have expressed a huge number of opinions on various things. Sentiment analysis is a structured and systematic analysis of these text data, and extract and identify the feelings in it. It is widely used in business, health care, financial markets, government agencies and other fields to complete customer opinion analysis, improve customer satisfaction, improve patients’ mental health status analysis, market trend and public opinion monitoring and other tasks. Because machine learning and deep learning-based approaches have become the mainstream approach to sentiment analysis, surpassing traditional dictionary approaches and rules and pattern matching approaches, while being represented by pre-trained models, achieving amazing performance in a variety of natural language processing tasks. However, the deep learning model contains a large number of parameters, which need to rely on a large number of training data sets to optimize the parameters using gradient descent algorithm. The integrity and representativeness of the training data and whether it contains bias will affect the accuracy of the deep learning model. For sentiment analysis in the English context, it is easier to learn and train on a large amount of labeled data. Because the corpus of natural language processing in the English environment is the richest, there are already a large number of artificially labeled data sets. However, the creation of data sets for low-resource languages is very demanding, such as the low degree of digitization of the language, the lack of large-scale digital text data or the lack of sufficient manpower and funds to organize the data, resulting in the lack of a rich variety of data and a large enough data set. Sentiment analysis on low-resource languages can be limited by the lack of a sufficient corpus to build a high-performance model. Parallel corpus and machine translation approaches offer some effective solutions. However, they still suffer from many limitations and do not provide sufficient access to cross-lingual information for sentiment analysis. In this thesis, We propose a code-switching method based on multilingual embeddings and Multilingual BERT (mBERT) masked language modeling that generates artificial code-switching sentences. This data augmentation is then used to fine-tune the pre-trained multilingual model. Since we only use multilingual embeddings as generation resources, we can improve the performance of low-resource languages in pre-trained multilingual model. We believe that this approach effectively adapts the embedding of high-resource languages to low-resource target languages. The results of the implementation show that our approach is effective for sentiment analysis tasks in low-resource languages and outperforms the machine translation approach.