Etymon-level Convolutional Neural Networks for text classification

1810125 Tie Xinyong

Text classification is a classic topic in natural language processing. In recent years, deep learning has achieved promising results in natural language processing, such as Convolutional Neural Networks. Convolutional Neural Networks is useful in extracting information from row signals based on computer vision, and it is also achieved to natural language processing. Character level Convolutional Neural Networks (Char-CNN) achieved good results for text classification, which quantizes text and learns text classification using letters as features. However, Char-CNN has a problem that the learning feature is too abstract and loses meaning to make accuracy is lower than word-level models. In Char-CNN, learning a character string as features is higher accuracy than single character. We attempt to find a method of character sequence for improving CNN using etymology. It can avoid dimensional curse and learns with a meaningful feature by the etymons. However, research on etymon-level deep learning is still scarce.

We propose a method that uses etymology to make the etymon as features and to clarify the effects of this method.

This paper evaluates the performance of etymon-level text classification by comparing it to that of word-level and character-level text classifications.

We conduct experiments with text classification in three methods to evaluate the effect of etymon-level. The first is to evaluate the performance using 5 machine learning models and evaluate the accuracy and training time. The second is to learn with a Convolutional Neural Networks using various corpora, to evaluate improve speed in accuracy and loss. Accuracy and loss are recorded each epoch. Third, we perform text clustering experiments using word embeddings and discuss the results.

We collected etymon data from the Online Etymology Dictionary (www.etymonline.com) using a web crawler. The dictionary contains approximately 44,000 words and etymon information. Next, a large corpus (word-level) is collected, and a etymon-level and character-level corpus is created. Then, the corpus based on characters, etymons, and words is converted to a vector, and feed to machine learning models.

First, all the words in corpus were converted to prototypes, and then, replaced by the etymons. For words such as SUMITOMO, which are in the corpus that are not included in the etymology dictionary, they were converted [Unknown]. In Char-CNN, string type data is interpreted as a list of character types, so a direct loop is used to separate character strings and add them to the character base corpus. Next, corpus is changed to a vector

using tf-idf or 1-hot. Then, it is divided to 80% for training data and 20% for test data. Finally, fade data into the machine learning model.

The machine learning models are Naïve Bayes, Support Vector Machine, Logistic Regression, K-Nearest Neighbor and Convolutional Neural Networks. We experiment text classification using five machine learning models and discuss the results.

We experiment on text classification of Convolutional Neural Networks using corpora with different characteristics. Since the etymon is a common semantic code of a word, the etymon-level model does not increase so much even if the corpus vocabulary amount increases. And the corpuses used have different characteristics. For example, British Broadcasting Corporation news from large vocabulary British editors.

Moreover, the etymon embedding demonstrated good clustering performance. We train etymon embeddings using Skip-grams. And evaluate it in text clustering using K-means.

The performance of etymon-level text classification is clarified. Results show that in NB, SVM, LR, and CNN, etymon-level is better than character-level, and has a accuracy close to the word-based. Etymon-level clearly increases accuracy and decreases the loss faster in CNN. In text clustering, etymon-level is better than word-based. Etymon-level is a competitive method to traditional methods for text classification task.