学位論文要旨

西暦　2019 年　7 月　24 日

学位申請者
靳展　印

学位論文題目

A Study on Classification Problems in Natural Language Processing through Distributed Representation

学位論文の要旨

The essence of many research topics in the field of natural language processing is to solve specific classification problems. Traditional methods in solving classification problems is to extract feature vectors from text data to train the classifier. With the improvement of computer hardware performance and the emergence of new neural network language models, distributed representations of text data are increasingly used when constructing feature vectors. Existing methods tend to combine the distributed representation of text with rich syntactic and grammatical information to construct feature vectors for classification. Compared with the previous method which using only syntactic and grammatical information, the existing method can improve the accuracy of the classification result, but it will bring a variety of negative effects. For example, the reduction in versatility, the increase in computational cost, and the increase in algorithm complexity.

The motivation for this study is that we want to find out what classification results can be obtained by using only distributed representations of text to solve classification problems. The purpose of this study is to obtain a classification result similar to the existing method by using only the distributed representation of the text data in solving the classification problem of natural language processing.

In the first part of the study, since the vector representation of the word has one-hot representation and distributed representation, we first explored the performance of the one-hot representation of words in solving the classification problem of the semantic relationship between two words. Our research object is the Associative Concept Dictionary (ACD), which is a database that holds any two Japanese words and their semantic relationship. Since the existing ACD needs to be manually created, the efficiency of the existing ACD expansion is low. Therefore, we proposed a new method to automatically extract the association concept from the text data on the network to expand the existing ACD. The most important issue is to correctly classify the semantic relationships between two words. We first combined the data in the existing ACD with the data from Wikipedia into teacher data, then we used this teacher data to train a traditional classifier. Finally, we used this classifier to judge two new words which not in the existing ACD. The purpose of expanding the dictionary is achieved by adding two new words and their semantic relationship to the existing ACD. We used the one-hot representation of the word and the information about the syntax when constructing the feature vector for semantic classification. We used 869 validation data to validate our approach. According to the experimental results, the precision of our proposed method of classification result is relatively high, 80%, the recall is 63.2%, but the accuracy is 48.9% and the F-score 55.1% is relatively low. We believe that this is because the number of negative cases in the candidate data extracted from the Wikipedia data is much larger than the number of positive cases. And the one-hot representation of the word does not carry enough information about the semantic relationship. Although a higher classification precision can be obtained, it is not satisfactory in classification accuracy.

Based on the conclusions of the first part of the above study, we focused on the distributed representation of words in the second part of the study, which is also the main part of this study. we explored the performance of the distributed representation of words when solving the classification problem of the semantic relationship between two words. For problems in existing methods, the purpose of this part of study is to use only distributed representations of words to obtain classification results without using any syntactic and grammatical features and external semantic relational databases. We proposed an approach to build features for relational classification which consisted of only the low-dimensional vectors representing substrings between words called *substring vectors*. For two words in a sentence, if we want to classify the semantic relationship of the two words, we first construct the substring vector of the two words. We extract the partial word sequence between the two words, and then obtain the distributed representation of each word in the partial word sequence from the distributed representation data of words prepared in advance. After weighting and normalization these distributed representations, and we finally calculate the average vector of these distributed representations to obtain the substring vector. We use this substring vector as the feature vector for semantic relationship classification. To be compared with similar studies, we used the same SemEval-2010 Task 8 data in the validation experiment. It contains 8000 training data and 2718 verification data, with 9 kinds of semantic relationships. We used a traditional classifier and a simple three-layer feedforward neural network as a classifier. According to the experimental results, the accuracy of the classification result of our proposed method is 79.73%, and the accuracy is improved to 81.28% after using our weighting method. The accuracy of our classification results is slightly lower than similar studies using deep neural networks as classifiers, but higher than most similar studies using traditional classifiers. Most importantly, unlike similar studies, we obtained such satisfactory classification accuracy without using any syntactic and grammatical features and external semantic relational databases. Because of this, our method has low-dimensional vectors, which makes the calculation cost lower and the versatility is higher. We consider that the distributed representation of words already carries enough information about the semantic relation classification, and this information can be used very well by our proposed substring vector.

In the second part of the study, we found that for distributed representation, deep neural networks as a classifier have better performance than traditional classifiers. In the third part of the study, we explored the performance of the deep neural network as a classifier and the distributed representation of the sentence when solving the classification problem. In the actual project of Computer Supported Collaborative Learning (CSCL), in order to support teachers to find out the problems during the teaching based on the chat data of the students in each group and exploring solutions, an automated method for tagging chat data is needed. we proposed a newly designed coding scheme for a large-scale coordinated education data, and we tried to automate time-consuming coding task by using deep learning technology. The 5 dimensions are Participation, Epistemic, Argument, Coordination and Social dimension. We first constructed feature vectors using the distributed representation of the sentences in the chat data of the students, and used these feature vectors to train the deep neural network to classify the tags of the sentence. Furthermore, we analyzed the relationship between the classification results and the evaluations given by the teachers, and find out the problems during the teaching. In our experiments, the sizes of the data are 8460 for the Epistemic dimension, 7795 for the Augment dimension, 3510 for the Coordination dimension, 2619 for the Social dimension. These data was used for learn the model. The Participation dimension is primarily about the level of participation of students in a group discussion and does not require classification. According to the results of the previous research, we selected the deep neural network based on Seq2Seq with the highest classification accuracy as the classifier, and trained 4 independent classifiers for the 4 dimensions in the experimental data. According to the experimental results, we obtained the classification accuracy similar to the human coder in multiple labels in 4 dimensions. And according to the teacher's evaluation of each group and the actual chat data, we found that when there are more sentences in the Elicitation label of Social dimension , it will cause confusion to other students in the group and have a negative impact on teaching, and it is same with the Quick Consensus of Social dimension , it will make the conversation in the group active and have a positive impact on teaching.

Based on the above, we summarize the following conclusions. In the first part of the study, we found that the one-hot representation of a word is not satisfactory in the classification accuracy of the semantic relationship. We consider that this is because the one-hot representation of the word that it does not carry enough information about the semantic relationship. In the second part of the study, we proposed the substring vectors based on the distributed representation of words to classify semantic relationships. Without any syntactic and grammatical features and external semantic relational databases, we obtain classification accuracy higher than most similar methods. We can conclude that the distributed representation of words carries sufficient information about the semantic relationship, and this information can be used reasonably and efficiently in some way such as substring vector to solve the problem of semantic relationship classification. In the third part of the study, we used a distributed representation of sentences to train a deep neural network classifier in the actual CSCL project, and obtained classification accuracy similar to that of human coder. It shows that distributed representations perform better when combined with deep neural networks as classifiers than traditional classifiers. Based on the conclusions of the three parts of this study, we conclude that The distributed representation of text has better classification results than traditional syntactic and grammatical features when solving classification problems. Through the proposed substring vectors, the potential information related to the semantic classification owned by the distributed representation can be utilized efficiently, and the classification result higher than most similar studies can be obtained, If we use a deep neural network classifier, we can more effectively exploit the advantages of distributed representation in solving classification problems in the field of natural language processing.