Aspect-Oriented Unsupervised Domain Adaptation for Polarity Classification

2110196 LU Binghan

In recent years, there has been a lot of research focused on classifying the sentiment or polarity of reviews written by users. The task of classifying the polarity of a review (polarity classification) is a type of document classification, which is one of multi-class classification problems that predicts the polarity of a text. In general, supervised machine learning methods are frequently used for polarity classification, but there is a well-known problem that the performance of polarity classification tends to be poor when the domains of training data and test data are different.

To tackle the above problem, a technique called "domain adaptation" has been studied to reduce the difference or gap of two different domains as much as possible. Most of current domain adaptation methods on polarity classification of a user review have generally focused on a genre or medium of a text as the domain, but few studies have focused on an aspect. Aspect-based sentiment analysis is a task to classify whether an opinion expressed by a user towards an aspect of a product or service is positive or negative. However, since different labeled data is required for each aspect, a similar problem of ordinary domain adaptation may occur in aspect-based sentiment analysis, i.e., difference between training and test data of different aspects may decrease the accuracy of polarity classification.

The purpose of this study is to propose a domain adaptation method where an aspect is defined as a domain, and trains a model of polarity classification for a certain aspect that is different from one in the training data. Specifically, to automatically construct labeled data for a target domain, we combine two methods: (1) a method that automatically determines labels by using Bidirectional Encoder Representations from Transformers (BERT) trained from source domain data, and (2) a method that automatically generates labeled data for the target domain by replacing sentiment words and keywords in sentences of the source domain with those of the target domain. Here, the source and target domain refer to the domain of the training and test data, respectively. Finally, the automatically constructed labeled data is used to train a classifier that determines the polarity of the aspects in the target domain.

Labeled data of the target domain are constructed by the following two methods. The first method fine-tunes a BERT model with labeled data in the source domain, then uses it to annotate unlabeled reviews in the target domain with the polarity label. In this process, review sentences are

discarded when the probability of the prediction of the label is lower than a pre-defined threshold, because the assigned polarity label may be incorrect. It enables us to construct high quality labeled data.

The second method builds labeled data of the target domain by the following three steps: the extraction of domain specific sentiment words and keywords, the generation of review sentences of the target domain, and the filtering of the generated sentences. In this study, we call this method Cross Aspect Review Generation (CARG). For each source and target domain, we first extract the sentiment words frequently used in the reviews of that domain, i.e., aspect. Sentences in a domain corpus are split into words, then the polarity score for each word is calculated using the part-of-speech tagger and the sentiment dictionary, SentiWordNet, then words whose sentiment scores are greater than or equal to a threshold are extracted. Furthermore, "domain specific keywords", which are defined as words frequently appearing only in a certain domain, are extracted. Given a collection of reviews of different aspects, a set of reviews about one aspect is regarded as a single document, then the TF-IDF of a word is calculated. The words with the highest TF-IDF values are extracted as the domain specific keywords. Next, a new labeled sentence in the target domain is generated by replacing the sentiment words and keywords in the labeled sentence of the source domain with those in the target domain. We use the Masked Language Model (MLM) of BERT for word substitution. Finally, for each generated sentence, we measure the pseudo-log-likelihood (PLL) score, which evaluates the fluency of a sentence, then filter out the sentences whose PLL scores are low.

We fine-tune the BERT model using the dataset constructed by the above two methods to get a polarity classification model for the target domain. For fine-tuning, Focal Loss is used as a loss function to alleviate imbalance of polarity labels in the training data.

To evaluate the proposed method, we conducted experiments of domain adaptation of polarity classification on the restaurant dataset and the laptop dataset. The former consists of reviews of five aspects and the latter of four aspects. The performance of polarity classification was measured when one aspect is chosen as the source domain and another aspect as the target domain (called cross-domain setting). For comparison, we also performed polarity classification where the source and target domains are the same (called in-domain setting). The evaluation criterion is the accuracy for cross-domain, and the micro-average of the accuracy of five trials in five-fold cross-validation for in-domain. We compared the proposed method with the baseline methods and the method of previous work with respect to those evaluation criteria.

The results of the experiments showed that the proposed method outper-

formed the baseline methods in terms of the accuracy for 17 of the total 20 pairs of the aspects in the restaurant dataset. Compared with the previous work, the proposed method achieved the better accuracy in 14 of the total 20 aspect pairs. Comparing the systems with and without filtering of unnatural sentences, the accuracy of the former was better for 12 of the 20 aspect pairs, indicating that the filtering was effective. Comparing the systems using the Focal Loss and the ordinary cross-entropy as the loss function, the accuracy of the former was the same or higher for 15 of the total 20 pairs. The use of Focal Loss was effective in this experiment. The average accuracy of the proposed method using both filtering and Focal Loss was 0.658, which was 0.025 and 0.013 points higher than the two baseline methods, respectively. It was the highest average accuracy among all the methods. Results of the experiments on the laptop dataset also showed that the proposed method outperformed the baseline methods. The best result among the compared methods was achieved by the proposed method that did not use Focal Loss as a loss function; its average accuracy of all combinations of the aspects was 0.801. It was better than the two baseline methods by 0.011 and 0.007 points, respectively. Compared to the restaurant dataset, however, the difference of the accuracy between the baseline and the proposed method was smaller. As for the comparison with the in-domain setting, for some pairs of the aspects, the accuracy of the proposed method in the cross-domain setting was higher than the micro-average of the accuracy in the in-domain setting. Since similar results were observed on two different datasets, the proposed method is robust in the sense that it can be applied to any genres of the review.