Domain Adaptation for Gender Classification of Text

1610401 Sitong Wang

Due to the rapid growth of Internet, nowadays, people can easily release information or contents to the public. Such user-generated contents often include opinions and emotions of users. Opinion mining is a task to analyze texts written by many users and reveal their opinions toward a specific target such as a product, person or service. It is one of the hot research topics in natural language processing research field. Since different trends of opinions are often found for females and males, it is necessary to distinguish texts written by females or males. Therefore, gender classification, which is a task to identify an author's gender of a given text, is a basic and essential research topic in the opinion mining.

Although there are a large number of studies on the gender classification of Web documents at present, most of them are modeled and predicted in a single type of Web document database. A type of Web documents is often called a "domain" of texts. Blog and microblog (e.g. Twitter, Weibo) are examples of the domain. Most past studies of the gender classification rely on supervised machine learning. However, a model trained in one domain rarely works well in other domains, besides, to prepare labeled datasets of many domains needs enormous costs and time. In this thesis, we focus on the domain adaptation of the gender classification. That is, we aims at building a classification model to identify the author's gender in one source domain and apply it to a different target domain without remarkable loss of the performance. Here, "source" and "target" domains refer to types of documents of the training and test data, respectively.

In the proposed method, a classifier of the gender classification, which judges whether a gender of an author of a text is female or male, is trained from a collection of labeled data. It consists of several steps. The first step is preprocessing. Since texts on the Web are classified in this study, there is much information other than texts, such as URL and non-English words. Such noisy information is removed by simple rules based on regular expression. Then, lemmatization is performed to convert words in inflected forms to base forms. The second step is features extraction. We use word unigram and bi-gram as features for machine learning. The third step is feature selection. For each feature, χ^2 value that evaluates correlation between the feature and the gender class is measured. Then the top 5% or 10,000 features that have the highest χ^2 values are chosen. The last step is training of a gender classification model. A classifier is trained by Naive Bayes or Support Vector Machine (SVM).

In addition to the training of the model of the gender classification, this thesis also considers the domain adaptation, which is the main goal of this study. Among several approaches of the domain adaptation, we focus on two existing domain adaptation methods: the cut-off method and fill-up method. The cut-off method improves the classification accuracy by shortening the feature space of source and target domains by retaining only common features in two domains. On the other hand, the fill-up method extends the feature space of two domains. That is, not only the common features but also domain specific features compose the feature space. Both methods only change the feature space without changing weights of the features. Therefore, domain specific features, which appear only in either the source or target domain, are not heavily considered in a trained classifier. It may lead only a little improvement in the gender classification of different domains.

We propose a novel method for the domain adaptation of the gender classification called "fill-up with word similarity". Although our goal is the gender classification, our proposed method is applicable for any kinds of text classification problems. For a given training data in the source domain and test data in the target domain, we make three sets of features: common features, source specific features and target specific features. For each sample in the source domain, we search target specific features that are similar to one of the features appearing in the sample. Although word uni-gram and bi-gram are used as the features, here we consider only the uni-gram, i.e. word itself. The similarity between words is measured by cosine similarity of two word vectors that are derived from word embedding. We use word embedding obtained by fastText, which is pre-trained from a huge amount of English texts. If the word similarity is greater than the threshold T, similar target specific features are added to the feature vector by changing their weights to 1. Similarly, when a feature vector in the target domain is constructed, similar source specific features are added. In this study, the threshold T is set to 0.7 by our intuition. In our method, source specific and target specific features are taken into account in the training of a gender classifier by changing the weights of similar (or related) domain specific features to 1. It enables us to fill a gap of the feature space between the source and target domains.

Several experiments are conducted to evaluate effectiveness of our proposed method. Two datasets are used: one is Twitter dataset consisting of 260,944 tweets, the other is blog dataset consisting of 268,296 blog articles. These datasets are balanced, i.e. they contain the same number of texts written by females and males. Two cases are considered in this experiment. In the first case, the Twitter dataset is used as the source domain and the blog dataset is used as the target domain. In the second case, two datasets

are swapped. First, it is found that SVM outperforms Naive Bayes, and the feature selection of the top 5% features is better than that of the top 10,000 features. Then, we compare two existing and our proposed domain adaptation methods. The accuracy of the gender classification is 54.64% or 53.00% in the first or second case, when no domain adaptation is applied. The accuracy of the cut-off method is 55.67% or 55.34%, while the accuracy of the fill-up method is 57.73% or 58.55%. Thus the performance of the gender classification is improved by the domain adaptation. Finally, the accuracy of our proposed method is 59.97% or 65.55%. It is better than two baseline methods, especially in the second case. These results prove the effectiveness of our new domain adaptation method.

In future, we should refine preprocessing of Web texts to train an accurate gender classifier. In addition, more sophisticated way to calculate word similarity should be explored.