Information Sciences 317 (2015) 67–77

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Combination of active learning and self-training for cross-lingual sentiment classification with density analysis of unlabelled samples

Mohammad Sadegh Hajmohammadi [a], Roliana Ibrahim [a], Ali Selamat [a,*], Hamido Fujita [b]

[a] UTM-IRDA Digital Media Center of Excellence, UTM & Faculty of Computing, Universiti Teknologi Malaysia, 81310 UTM Skudai, Johor, Malaysia
[b] Software and Information Science, Iwate Prefectural University, Takizawa, Japan

ARTICLE INFO

Article history:
Received 1 November 2014
Received in revised form 27 March 2015
Accepted 5 April 2015
Available online 9 April 2015

Keywords:
Cross-lingual
Sentiment classification
Self-training
Active learning
Density measure

ABSTRACT

In recent years, research in sentiment classification has received considerable attention by natural language processing researchers. Annotated sentiment corpora are the most important resources used in sentiment classification. However, since most recent research works in this field have focused on the English language, there are accordingly not enough annotated sentiment resources in other languages. Manual construction of reliable annotated sentiment corpora for a new language is a labour-intensive and time-consuming task. Projection of sentiment corpus from one language into another language is a natural solution used in cross-lingual sentiment classification. Automatic machine translation services are the most commonly tools used to directly project information from one language into another. However, since term distribution across languages may be different due to variations in linguistic terms and writing styles, cross-lingual methods cannot reach the performance of monolingual methods. In this paper, a novel learning model is proposed based on the combination of uncertainty-based active learning and semi-supervised self-training approaches to incorporate unlabelled sentiment documents from the target language in order to improve the performance of cross-lingual methods. Further, in this model, the density measures of unlabelled examples are considered in active learning part in order to avoid outlier selection. The empirical evaluation on book review datasets in three different languages shows that the proposed model can significantly improve the performance of cross-lingual sentiment classification in comparison with other existing and baseline methods.

1. Introduction

Sentiment classification, which aims to classify opinion text documents (e.g., product reviews) into polarity categories (e.g., positive or negative), has received considerable attention in the natural language processing research community due to its many useful applications [8]. These include online product review classification [14] and opinion summarisation [15]. Although traditional supervised classification algorithms can be employed to train sentiment polarity classifiers from labelled text data, manually construction of labelled sentiment data is a labour-intensive and time-consuming task.

* Corresponding author. Tel.: +60 7 5531008; fax: +60 7 5530160.
E-mail address: aselamat@utm.my (A. Selamat).

http://dx.doi.org/10.1016/j.ins.2015.04.003
0020-0255/

Since most recent research studies in sentiment classification have been performed in some limited number of languages (usually English), there are an insufficient number of labelled sentiment data existing in other languages [23]. Therefore, the challenge arises as to how to utilise labelled sentiment resources in one language (the source language) for sentiment classification in another language (the target language). This challenge then leads to an interesting research area called cross-lingual sentiment classification (CLSC). Most existing research works have employed automatic machine translation to directly translate the test data from the target language into the source language [20,25,32,33]. Following this, a trained classifier in the source language has been used to classify the translated test data.

However, term distribution between the original and the translated text document is different due to the variety in writing styles and linguistic expressions in the various languages. It means that a term may be frequently used in one language to express an opinion while the translation of that term is rarely used in the other language. Hence, these methods cannot reach the level of performance of monolingual sentiment classification. To solve this problem, making use of unlabelled data from the target language can be helpful, since this type of data is always easy to obtain and has the same term distribution as the target language. Therefore, employing unlabelled data from the target language in the learning process is expected to result in better classification performance in CLSC.

Semi-supervised learning [24] is a well-known technique that makes use of unlabelled data to improve classification performance. One of the most commonly used semi-supervised learning algorithms is that of self-training. This technique is an iterative process. Semi-supervised self-training tries to automatically label examples from unlabelled data and add them to the initial training set in each learning cycle. The self-training process usually selects high confidence examples to add to the training data. However, if the initial classifier in self-training is not good enough, there will be an increased probability of adding examples having incorrect labels to the training set. Therefore, the addition of "noisy" examples not only cannot increase the accuracy of the learning model, but will also gradually decrease the performance of the classifier. On the other hand, self-training selects most confident examples to add to the training data. But these examples are not necessarily the most informative instances (especially for discriminative classifiers, like SVM) for classifier improvement [16]. To solve these problems, we combine the processes of self-training with active learning in order to enrich the initial training set with some selected examples from unlabelled pool in the learning process. Active learning tries to select as few as possible the most informative examples from unlabelled pool and label them by a human expert in order to add to the training set in an iterative process. These two techniques (self-training and active learning) complement each other in order to increase the performance of CLSC while reduce human labelling efforts.

In this paper, we propose a new model based on the combination of active learning and semi-supervised self-training in order to incorporate unlabelled data from the target language into the learning process. Because active learning tries to select the most informative examples (in most cases, the most uncertain examples), these examples may be outlier, especially in the field of sentiment classification of user's reviews. To avoid outlier selection in the active learning technique, we considered the density of the selected examples in the proposed method so as to choose those informative examples that had maximum average similarity (the more representatives) in the unlabelled data. The proposed method was then applied to book review datasets in three different languages. Results of the experiments showed that our method effectively increased the performance levels while reduced the human labelling effort for CLSC in comparison with some of the existing and baseline methods.

This paper is an extended version of work published in [9]. We extend our previous work in four directions. First, we add more description regarding problem situation and corresponding solutions and also more discussion about experimental results. Some new findings from new experiments are also presented in this version. Secondly, more evaluation datasets in new languages are used in the evaluation section to show the generality of the proposed model in different languages. Thirdly, the comparison scope is extended by adding more baseline methods and one of the best performing previous method in CLSC in order to reveal the effectiveness of the proposed model. Finally, in order to assess whether there are significant performance improvement between the proposed model and other methods, a statistical test is added to the evaluation section.

The remainder of this paper is organised as follows. The next section presents an overview of related works on CLSC. The proposed model is described in Section 3, while the evaluation and experimental results are given in Section 4. Finally, Section 5 concludes this paper.

2. Related works

In recent years, cross-lingual sentiment classification has drawn much research attention. Many research studies have been conducted in this area. These research studies are based on the use of annotated data in the source language (always English) to compensate for the lack of labelled data in the various target languages. Most approaches have focused on resource projection from one language to another with few sentiment resources. For example, Mihalcea et al. [21] generated subjectivity analysis resources into a new language from English sentiment resources by using a bilingual dictionary. In other works [2,3], automatic machine translation engines were used to translate the English resources for subjectivity analysis. In [2], the authors showed that automatic machine translation was a viable alternative for the construction of resources for subjectivity analysis in a new language. In two different experiments, they first translated the training data of subjectivity classification from the source language into the target language. They then utilised this translated data to train a classifier in