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

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 the target language. Subsequently, this trained classifier was applied to classify the test data. Further, in another experiment, machine translation was used to translate the test data from the target language into the source language. A classifier was then instructed based on the training data in the source language. After the training phase, the translated test data were presented to the classifier for sentiment polarity prediction. Wan [32] used unsupervised sentiment polarity classification in Chinese language product reviews. He translated Chinese language reviews into different English reviews using a variety of machine translation engines and then performed sentiment analysis for both the Chinese and English language reviews using a lexicon-based technique. Finally, he used ensemble methods by which to combine the analysis results. This method required a sentiment lexicon in the target language and could not be applied to other languages with no lexicon resource. Pan et al. [25] designed a bi-view non-negative matrix tri-factorization model to solve the problem of cross-lingual sentiment classification. They used machine translation to achieve two representations of the training and test data: one in the source language and another in the target language. This model was then used to combine the information from two views. Although Balahur and Turchi [1] showed that automatic machine translation systems are approaching a reasonable level of maturity to project sentiment information from one language into another, but different term distribution between the original and translated text documents is the main problem faced in the case of using only machine translation.

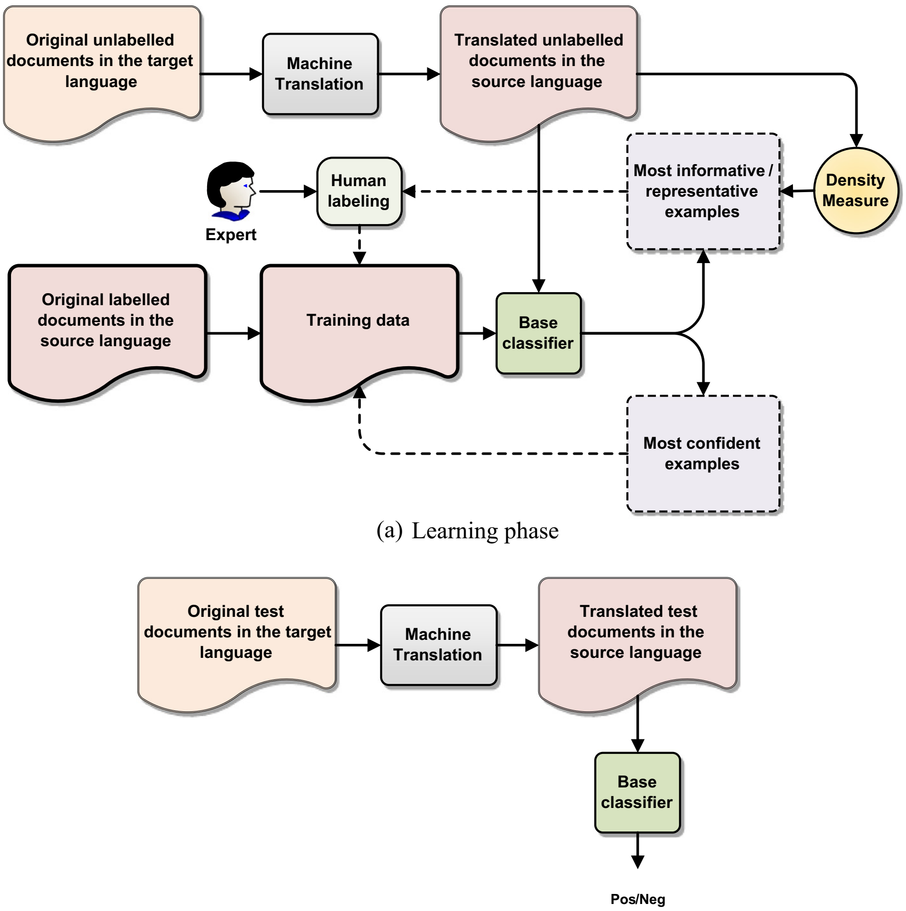
Domain adaptation is one of the transfer learning approaches that were employed in the field of cross-lingual sentiment classification by some researchers. Prettenhofer and Stein [28] generalised structural correspondence learning (SCL), a recently mentioned theory for domain adaptation [4], to use in cross-lingual sentiment classification. Similar to SCL, their method included correspondence between the words from two different languages by using a small set of frequent features in both languages called pivot features. They used a pair of words, (ws, wt), from the source and the target languages as pivot features and tried to model a correlation between pivot features and other words, by a linear classifier and a set of unlabelled data from both languages, which then was used as a language-independent predictor. In another work, Wei and Pal [34] proposed a model that used only key reliable parts from the translated data and applied structural correspondence learning (SCL) to find a low dimensional representation shared by the two languages (the source and the target language). To solve the problem of feature sparseness, the authors employed a search engine in order to find high correlated features to those in the pivot feature set. The new selected features were then used to build extra examples for the training set. The performance of domain adaptation techniques is highly dependent on the selection of pivot features. A good selection of pivot features makes this methods to show a good performance in classification. Unfortunately, there is not a systematic method for pivot feature selection and consequently, these methods cannot guarantee the performance.

Another approach is that of cross-lingual classification, which involves translating the features extracted from labelled documents [22,30]. The features, selected by a feature selection algorithm, are translated into different languages. Subsequently, based on those translated features, a new model is trained for each language. This approach only needs a bilingual dictionary to translate the selected features. It can, however, suffer from inaccuracies in the dictionary translation because words may have different meanings in different contexts. Therefore, selecting the features to be translated can be an intricate process.

To solve the problem of different term distribution between the training and test data in CLSC, some researchers tried to incorporate unlabelled data from the target language by using semi-supervised learning approach [10,11,33]. As an example, Wan [33] used the co-training method to overcome the problem of cross-lingual sentiment classification. He exploited a bilingual co-training approach to leverage annotated English language resources to sentiment classification in Chinese language reviews. In that work, firstly, machine translation services were used to translate English labelled documents (training documents) into Chinese and similarly, to translate Chinese unlabelled documents into English. The author used two different views (English and Chinese) in order to exploit the co-training approach to the classification problem. The selection strategy of semi-supervised learning usually selects high confidence examples to add to the training data. However, if the initial classifiers are not good enough, there will be an increased probability of adding examples having incorrect labels in 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 each classifier. Furthermore, the most confident classified examples are not necessarily the most informative ones in the learning process. Therefore, adding these examples may not be very useful from the classification viewpoint. Considering these problems, the use of active learning with semi-supervised learning is considered in this paper in order to select most informative examples from unlabelled document along with most confident classified unlabelled examples to enrich the training data.

3. Proposed model

As mentioned in the first section, since the training data in cross-lingual sentiment classification has different term distribution with the translated test documents of the target language, the performance of the sentiment classifier in this case is limited. To increase this performance, making use of unlabelled data from the target language can be helpful since these data are always easy to obtain. An additional benefit is that they have the same term distribution as the test documents in the respective target language. However, manually labelling unlabelled data is a labour-intensive and time-consuming task. In an attempt to incorporate unlabelled data and reduce the labelling effort, we propose a new model based on the initial training data from the source language and the translated unlabelled data from the target language. This model attempts to enrich the initial training data through the manual and automatic labelling of some unlabelled data from the target


Original unlabelled documents in the target language
Machine Translation
Translated unlabelled documents in the source language
Expert
Human labeling
Most informative / representative examples
Density Measure
Original labelled documents in the source language
Training data
Base classifier
Most confident examples
Original test documents in the target language
Machine Translation
Translated test documents in the source language
Base classifier
Pos/Neg

(a) Learning phase

(b) Test phase

Fig. 1. Framework of the proposed approach.

language. The framework of the proposed model is illustrated in Fig. 1. In the learning phase, a base classifier is trained based on the initial training data and then applied to the translated unlabelled data. From the newly classified unlabelled data, active learning selects the most informative example (the most useful examples which can help to improve the classifier) for human labelling. In human labelling process, a native speaker in the target language can read the review and evaluate the overall sentiment polarity (e.g. positive sentiment or negative sentiment) of that review. Simultaneously, self-training selects some of the most confident classified examples with the corresponding predicted labels. These selected examples are added to the training set for the next learning cycle. In the next cycle, the model is retrained based on the augmented training data and this process is repeated until a termination condition is satisfied. Further, in this model, the density of the unlabelled data is also considered in the selection of not only more informative, but also more representative, unlabelled instances in order to avoid outlier selection in active learning. In the test phase, the final trained classifier is applied to the translated test data for the classification task. We called this model "density-based active self-training" (DBAST). Active learning and self-training are described in detail in the following sections, respectively.

3.1. Active learning process with density analysis

Active learning is a subcategory of machine learning [7]. The main goal of active learning is to learn a classifier by labelling as few examples as possible when actively selecting the examples to be labelled in the learning process. As reported in previous research [6,12,19,38], active learning is a promising method by which to speed up data labelling while minimising human labelling efforts. An active learner can be modelled as a quintuple $(C,\ Q,\ O,\ L,\ U)$ [18]. Initially, the classifier C is trained based on labelled examples set L and applied to the unlabelled pool U. Following this, a set of the most informative examples is selected from the unlabelled pool U using a query function Q and a true class label is assigned to each of them by an oracle O. After that, these new labelled examples are augmented to L and the classifier C is retrained based on this upgraded training set. This sequential label requesting process continues for some predefined iterations or until a termination condition is satisfied.

The query function is an essential part of the active learning process. The major difference between several previously proposed active learning methods is the difference in their query function. The simplest query function is that of uncertainty sampling [17] in which unlabelled examples with maximum uncertainty are selected for manual labelling in each learning cycle. Maximum uncertainty means that the learner has less confidence in the classification of these unlabelled examples. Entropy is a popular uncertainty measurement widely used in recent research for uncertainty sampling [39]. Formula shows the uncertainty function calculated based on the entropy estimation:

$$H(x) = -\sum_{y \in Y} P(y|x) \log P(y|x) \tag{1}$$

$P(\cdot)$ is the posterior probability of the classifier and $H(\cdot)$ is the uncertainty function. The most informative example, u, is selected as:

$$u = \arg \max_{x \in U} H(x) \tag{2}$$

As reported in [12,29,31], many unlabelled examples selected by the uncertainty sampling function cannot help the learner since these are outliers. This issue is more serious in the field of review sentiment classification because many fake or unrelated reviews may be posted to the review websites. This means that a good selected example for manual labelling in active learning should not only be the most informative example, but also the most representative one. Therefore, adding outlier examples to the training data cannot provide much help to the learner. To solve this problem, Tang et al. [31] proposed a sampling method in which the most uncertain example is selected by a learner from each cluster and weighted using a density measure. In another work Jingbo et al. [12] proposed a density-based re-ranking technique to select the most informative and representative example from unlabelled data so as to solve the outlier problem in selective sampling. They introduced a density concept to determine whether or not an unlabelled example is an outlier. To determine the density degree of an unlabelled example, they used a novel method called the k-nearest neighbour-based density (kNN density). In this measure, the density degree of an unlabelled example is computed by the average similarity between this example and k most similar unlabelled examples in the unlabelled pool. Suppose $S(x) = \{s_1, s_2, \ldots, s_k\}$ is a set of k most similar unlabelled examples to $\mathcal{X}$. Therefore, the average similarity for $\mathcal{X}(A(x))$ can be computed based on the following formula:

$$A(x) = \frac{\sum_{s_i \in S(x)} Similarity(x, s_i)}{k} \tag{3}$$

Cosine distance is used as the similarity function to compute the pair-wise similarity value between the two examples as follow:

$$similarity(x, y) = \frac{x \cdot y}{\|x\| \|y\|} \tag{4}$$

Cosine distance has been successfully used in literature to compute the content similarity of text documents [12,35,38]. Therefore, it can be used as a good criteria to compute the density measure for sentiment text documents. This density measure was used in combination with an uncertainty measure to select the most representative and informative example from the unlabelled pool for human labelling. This measure was also employed to avoid the selection of outlier examples in each cycle of active learning. We referred to this new active learning model as density-based active learning. In the proposed model, an unlabelled example with maximum entropy and density was selected for manual labelling based on the following formula:

$$u = \arg \max_{x \in U} H(x) \times A(x) \tag{5}$$

This selected example is then labelled by an oracle and added to the training data.

3.2. Self-training algorithm

One of the most commonly used semi-supervised algorithms is self-training. In self-training, a base classifier is first trained by using initial labelled data and then used to predict the labels of unlabelled data. After prediction, the most confident unlabelled examples are selected to add to the training data with corresponding predicted labels. The confidence in each newly classified example is computed based on the distance of each example from the current decision boundary. In the proposed model, p positive and n negative (as the most confident examples) were selected as auto-labelled examples for the next iteration. The classifier was retrained and this process was repeated for some predefined iterations or until a stopping criterion was met.

4. Evaluation

In this section, we evaluate our proposed approach in cross-lingual sentiment classification on three different languages in the book review domain and compare it with other existing and baseline methods.

Although the book review domain is used in the evaluation, the proposed model can also be applied to other domain. The book reviews which used in this paper are not written by professional writer. Any person can write his/her opinion about a book in a review website (e.g., Amazon website). Book review domain is a challenging domain in the sentiment classification research community. Because the users always write their opinions on books. Therefore sometimes a mixture of positive and negative attitudes towards a single book is expressed in a book review. In this domain, each review document focuses on a single book and written by a single reviewer. The review documents are about different books with various topics.

4.1. Datasets

Three different evaluation datasets from three different languages were used in the research reported in this paper and are detailed as follows:

1. English–French book review dataset (En–Fr): This dataset contains Amazon book reviewing documents in both English and French languages. This dataset was used by Prettenhofer and Stein [28]. In this dataset, the English language was treated as the source language and the French language was treated as the target language. Documents in the English language containing 2000 (1000 positive and 1000 negative) book reviews were used as the labelled data. A total of 4000 review documents (2000 positive and 2000 negative) were selected from the French dataset and treated as the unlabelled data.
2. English–Chinese book review dataset (En–Ch): This dataset was selected from the Pan reviews dataset [25]. It contains book review documents in the English and Chinese languages. As for the previous dataset, documents in the English language containing 2000 (1000 positive and 1000 negative) book reviews were used as the labelled data. Documents in the Chinese language containing 4000 (2000 positive and 2000 negative) book reviews were treated as the unlabelled data.
3. English–Japanese book review dataset (En–Jp): This dataset contains Amazon book review documents in the English and Japanese languages. This dataset was also used by Prettenhofer and Stein [28]. In this dataset, the English language was treated as the source language and the Japanese language was treated as the target language. Documents in the English language containing 2000 (1000 positive and 1000 negative) book reviews were used as the labelled data. A total of 4000 review documents (2000 positive and 2000 negative) were selected from the Japanese dataset and treated as the unlabelled data.

All review documents are labelled as being either positive or negative based on their sentiment polarities. Each Amazon review has a polarity rating from zero to five stars. Zero star is the most negative review and five stats indicate the most positive review. All reviews with rating greater than three stars are labelled as positives and those with rating less than three stars are labelled as negatives. Reviews with three stars are discarded because their polarities are ambiguous. All the review documents in the target languages were translated into the source language (English) using the Google translate engine.[1] Table 1 shows the properties of these three evaluation datasets.

In the pre-processing step, all the English language reviews were converted into lowercase. Special symbols, words with one character length and other unnecessary characters were eliminated from each review document. In the feature extraction step, unigram and bi-gram patterns were extracted as sentimental patterns. To reduce the computational complexity, especially in density estimation, we performed feature selection using the information gain technique [37]. We selected 5000 high score unigrams and bi-grams as final features. Each document was represented by a feature vector. Each entry of a feature vector contained a feature weight. We used term presence as feature weights since this method has been confirmed as the most effective feature weighting method in sentiment classification [26,36].

4.2. Baseline methods

Comprehensive comparisons to the well-known and best performing methods as well as some baseline methods were performed in this study to evaluate the classification performances of the proposed model. This method of evaluation has been widely used by other researchers to assess the performance of CLSC [1,25,27,33]. Based on the existing literature, the structural correspondence learning (SCL) [27] is one of the most well-known and best performing methods that previously applied to the CLSC. Therefore this method is used as one of the baseline methods. In order to show the actual effects of different parts of the proposed model, some other baseline methods were also designed and implemented to compare with the proposed model. Theses baseline methods were designed to show how the proposed model could overcome the related problems.

The following baseline methods were implemented in order to compare the effectiveness of proposed model:

- Active self-training (AST) model: This model is similar to DBAST but without considering the density measure of uncertain examples. In fact, this model is the combination of uncertainty sampling and self-training. This baseline model was used to determine the effectiveness of the density analysis in the DBAST model.

[1] http://translate.google.com/.

Table 1
Details of the datasets used in our experiments.

Dataset	Domain	Languages		Total documents	Positive documents	Negative documents
En–Fr [28]	Book review	Source Language	English	2000	1000	1000
		Target Language	French	4000	2000	2000
En–Ch [25]	Book review	Source Language	English	2000	1000	1000
		Target Language	Chinese	4000	2000	2000
En–Jp [28]	Book review	Source Language	English	2000	1000	1000
		Target Language	Japanese	4000	2000	2000

- Active learning with uncertainty sampling (AL) model: This model is based on the simple uncertainty sampling approach. This baseline model was used to determine the effect of the automatic labelling of unlabelled examples.
- Self-training (ST) model: In order to compare this model with other active learning-based models, we randomly selected some manually-labelled examples from the unlabelled pool and added them to the initial training set before starting the learning process. The number of manually-labelled examples added to the initial training set is equal to the total number of examples selected by the query function of other active learning-based models. This baseline model was used to determine the effect of active learning part.
- Structural correspondence learning model (SCL): We implemented this model as introduced in [34]. In order to construct the pseudo-example set to use in the training phase of linear pivot predictors, we employed the Google search engine. In our experiment we used 200 pivot features.
- Random sampling (RS) model: In the random sampling approach, in each cycle, t examples were randomly selected from unlabelled data for human labelling and added to the training data with corresponding labels. Since the random sampling model generates high variance results due to the random selection strategy, the performance of random sampling was averaged over 10 runs using the same experimental setting.
- Support vector machine with machine translation (SVM-MT): This is a straightforward approach to cross-lingual sentiment classification. In this baseline model, first, test documents are translated from the target language into the source language. Following this, an SVM classifier is trained based on the training data in the source language and then used to predict the class label of the translated test documents. Note that unlabelled documents are not used in this baseline model.

4.3. Experimental setup

To estimate the average similarity for each unlabelled example in our proposed model, we used the method of Jingbo et al. [12]. At the beginning of the algorithm, we calculated the pair-wise similarity value between any two unlabelled examples and stored these values in a matrix. During the learning process, in order to calculate the average similarity for each unlabelled example, other examples were sorted based on their similarity scores. Therefore, the average similarity for each example can be calculated efficiently by averaging the similarity scores of the top-k examples in its sorted list.

In all the experiments, we used the SVM classifier [13] as the base classifier. SVM$^{\text{light}}$ (http://svmlight.joachims.org/) was used as the SVM classifier in the experiments with all the parameters set to their default values. However, SVM does not directly output the posterior probabilities of predicted labels. Therefore, we used a strategy that was introduced in [5] to compute these probabilities.

4.3.1. Cross-validation on active learning and semi-supervised learning

To generate reliable results, we performed a 5-fold cross-validation to obtain the results in all models (except SVM-MT). For this task, the unlabelled documents in the target language were translated into the source language and randomly divided into five groups of equal size. During each cycle of cross-validation, the text documents from four groups were treated as unlabelled data while the remaining group was treated as the test set. The training documents were fixed in each cross-validation cycle since they were from the source language. The final results were then averaged over five iterations. Configuration of the data splitting is illustrated in Fig. 2. To calculate the average similarity of each of the unlabelled examples, only the pair-wise similarity of those examples that are included in the unlabelled pool were considered.

4.4. Performance measures

Generally, the performance of sentiment classification is evaluated by using four indexes, namely: Accuracy, Precision, Recall and F1-score. Accuracy is the proportion of all true predicted instances against all predicted instances. An accuracy of 100% means that the predicted instances are exactly the same as the actual instances. Precision refers to the portion of true predicted instances against all predicted instances for each class. Recall denotes the portion of true predicted instances against all actual instances for each class. F1 is a harmonic average of precision and recall.

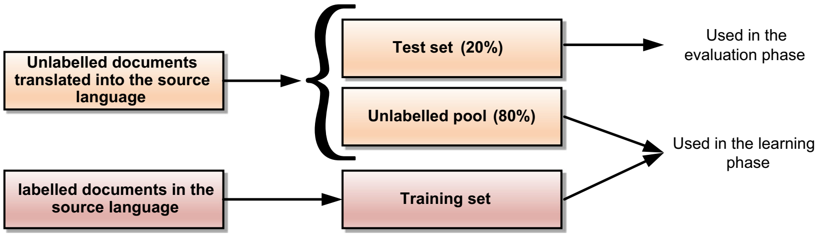

Unlabelled documents translated into the source language
Test set (20%)
Used in the evaluation phase
Unlabelled pool (80%)
Used in the learning phase
labelled documents in the source language
Training set

Fig. 2. Configuration of data splitting for experiment (5-fold cross-validation).

4.5. Results and discussion

In this section, our proposed method is compared with six baseline methods. We set k in the kNN density measure to 20 for all the experiments. We also used $p = n = 5$ for the self-training algorithm and selected one unlabelled example in each cycle of active learning for the manual labelling. The total number of iterations was set to 50 iterations for all iterative algorithms. This setting meant that, in total, 50 unlabelled examples were selected for manual labelling during the learning process and 500 unlabelled examples were labelled automatically. After a full learning process, the test data were presented to the learned classifier for evaluation. Table 2 shows the comparison results after the full learning process. As we can see, our proposed method showed a better performance in all datasets, especially with regard to accuracy.

Because active learning based models benefit from the information of 50 manually-labelled examples during their learning process, we added extra labelled samples (50 samples) from the source language to the training sets of the SCL and SVM-MT models in order to create the same condition for all comparing models. By comparing all semi-supervised and active learning based methods with SVM-MT model in Table 2, we can conclude that the incorporation of unlabelled data from the target language into the learning process can effectively improve the performance of cross-lingual sentiment classification. Also, as we can see in this table, DBAST and AST models demonstrate better performance in comparison to AL and ST after the full learning process. This supports the idea that the combination of active learning and self-training processes can result in a better classification than each individual approach. Moreover, the DBAST model outperforms the AST model in all datasets. This shows that using the density measure of unlabelled examples has a beneficial effect upon selecting the most representative examples for manual labelling.

As we can see, different languages show different accuracies in cross-lingual sentiment classification. This difference can be interpreted from two points of views. First, sentiment classification shows diverse performance in different languages due to the disparity in the structure of languages when expressing the sentimental data even in the same domain (e.g. book review domain). Next, automatic machine translation systems perform translation of varying quality in different languages

Table 2
Performance comparison in three datasets after the first 50 manually labelled examples. The best results are reported in bold-face.

Dataset	Method	Accuracy	Positive			Negative		
			Precision	Recall	F1	Precision	Recall	F1
En–Fr	DBAST	**78.63**	**77.54**	80.89	79.16	79.89	**76.36**	**78.07**
	AST	78.06	76.27	81.84	78.93	80.24	74.24	77.09
	AL	77.45	74.50	**83.89**	78.88	**81.45**	70.97	75.79
	ST	74.59	71.24	82.89	76.61	79.35	66.23	72.17
	SCL	78.41	76.00	82.98	**79.33**	81.34	73.84	77.40
	RS	75.02	71.84	83.24	77.04	79.78	66.73	72.53
	SVM-MT	74.45	70.59	83.80	76.63	80.07	65.10	71.81
En–Ch	DBAST	**71.36**	71.51	**70.67**	**71.07**	**71.25**	72.04	**71.62**
	AST	70.90	71.19	70.01	70.51	70.81	71.79	71.22
	AL	70.73	**71.75**	68.05	69.78	69.95	**73.39**	71.57
	ST	70.30	70.64	69.21	69.87	70.07	71.38	70.68
	SCL	70.58	70.89	69.24	70.06	70.28	71.90	71.08
	RS	70.03	70.59	68.70	69.43	69.93	71.33	70.44
	SVM-MT	69.50	69.51	67.99	68.74	69.41	70.79	70.03
En–Jp	DBAST	**70.04**	69.85	**70.46**	**70.14**	**70.25**	69.62	69.92
	AST	69.79	69.94	69.40	69.66	69.66	70.17	69.91
	AL	69.39	72.20	63.20	67.36	67.26	75.58	71.15
	ST	69.71	70.33	68.30	69.25	69.23	71.12	70.12
	SCL	70.00	73.39	62.89	67.74	67.76	77.10	**72.13**
	RS	69.39	72.75	62.34	66.94	67.17	76.43	71.37
	SVM-MT	68.50	**73.80**	57.13	64.40	65.20	**79.88**	71.79

Table 3
The *p*-value of paired *t*-test that compares the DBAST model with baseline methods for each dataset.

Dataset	Baseline methods					
	AST	AL	ST	SCL	RS	SVM-MT
En–Fr	1.67E−03	1.44E−03	6.22E−05	4.86E−04	4.08E−05	7.56E−08
En–Ch	9.00E−02*	2.24E−02	5.59E−03	1.40E−02	6.87E−04	7.44E−05
En–Jp	3.57E−02	7.72E−02*	2.47E−02	1.53E−02	2.53E−02	8.46E−03

* Not significant difference.

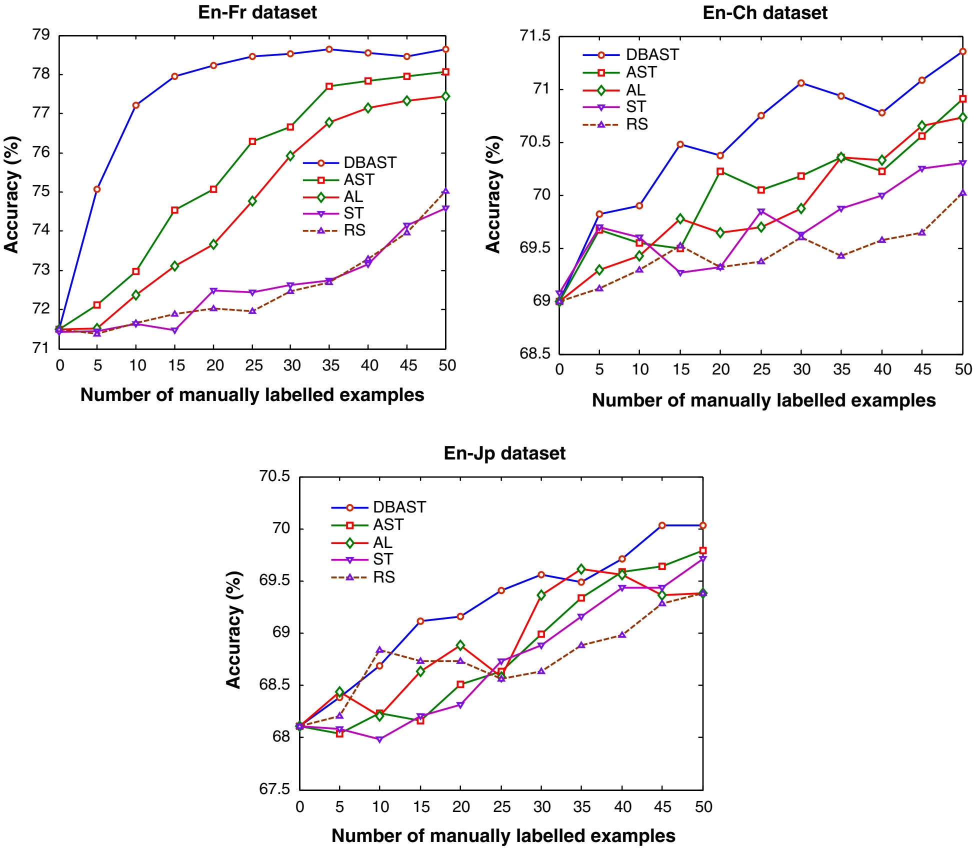


Fig. 3. Classification accuracy over the number of manually labelled examples on three different languages datasets.

(especially when we want to use the translation results to analyse the sentiment polarity of a text). Therefore, the performance of cross-lingual sentiment classification is different in different languages.

In order to assess whether there are significant differences in terms of accuracy between the proposed model and baseline methods, we conducted a statistical test based on accuracy results obtained from a 5-fold cross-validation. We used a paired *t*-test to evaluate whether differences between the two methods are statistically significant. Table 3 shows the numerical results of the statistical test. With the exception of those between the DBAST and AST models in the En–Ch dataset and between DBAST and AL in the En–Jp dataset, all other comparisons showed statistically significant differences, for a significant level of $\alpha = 0.05$.

Fig. 3 shows the classification accuracy of various active learning based methods on the three evaluation datasets. As shown in this figure, by comparing the proposed method (DBAST) with the AST model, the classification accuracy of the proposed model improved very quickly in the first few cycles (especially in the French language). This was because the examples selected based on density and uncertainty were more representative than those examples selected based solely on uncertainty in active learning. The results presented in this figure also showed that the combination of active learning with self-training helped to obtain better performance. This was most likely due to the augmentation of the most confident automatic classified examples, along with the manually labelled examples, into the training data during the learning process.

5. Conclusion

In this paper, we proposed a new model by combining active learning and self-training in order to reduce the human labelling effort and increase the classification performance in cross-lingual sentiment classification. In this model, first, unlabelled data were translated from the target language into the source language. Following this, they were then augmented into initial training data in the source language using active learning and self-training. We also considered a density measure to avoid the selection of outlier examples from unlabelled data and thus increase the representativeness of the selected examples for manual labelling in the active learning algorithm. We applied this method to cross-lingual sentiment classification datasets in three different languages and compared the performance of the proposed model with some baseline methods. The experimental results showed that the proposed model outperformed the baseline methods in almost all datasets. These results also showed that the incorporation of unlabelled data from the target language can effectively improve the performance of CLSC. The experimental results further showed that employing automatic labelling, along with active learning, can increase the speed of the learning process and therefore reduce the manual labelling workload. Experiments also demonstrated that considering the density of unlabelled data in the query function of active learning can be very efficient when selecting the most representative and informative examples for manual labelling.

Acknowledgement

This work is supported by the Ministry of Higher Education and Research Management Centre at the Universiti Teknologi Malaysia under Research University Grant Scheme (Vote No. 07J52, 4F550 and 02G31).

References

[1] A. Balahur, M. Turchi, Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis, Comput. Speech Lang. 28 (2014) 56–75.
[2] C. Banea, R. Mihalcea, J. Wiebe, S. Hassan, Multilingual subjectivity analysis using machine translation, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '08), Association for Computational Linguistics, Stroudsburg, PA, USA, 2008, pp. 127–135.
[3] C. Banea, R. Mihalcea, J. Wiebe, Multilingual subjectivity: are more languages better?, in: Proceedings of the 23rd International Conference on Computational Linguistics, Association for Computational Linguistics, Beijing, China, 2010, pp 28–36.
[4] J. Blitzer, R. McDonald, F. Pereira, Domain adaptation with structural correspondence learning, in: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Sydney, Australia, 2006, pp. 120–128.
[5] U. Brefeld, T. Scheffer, Co-EM support vector learning, in: Proceedings of the Twenty-First International Conference on Machine Learning, ACM, Banff, Alberta, Canada, 2004, pp. 16–23.
[6] J. Cheng, K. Wang, Active learning for image retrieval with Co-SVM, Pattern Recogn. 40 (2007) 330–334.
[7] D. Cohn, L. Atlas, R. Ladner, Improving generalization with active learning, Mach. Learn. 15 (1994) 201–221.
[8] M.S. Hajmohammadi, R. Ibrahim, Z. Ali Othman, Opinion mining and sentiment analysis: a survey, Int. J. Comput. Technol. 2 (2012) 171–178.
[9] M.S. Hajmohammadi, R. Ibrahim, A. Selamat, Density based active self-training for cross-lingual sentiment classification, in: H.Y. Jeong, N.Y. Yen, J.J. Park (Eds.), Advanced in Computer Science and Its Applications, Springer, Berlin, Heidelberg, 2014, pp. 1053–1059.
[10] M.S. Hajmohammadi, R. Ibrahim, A. Selamat, Cross-lingual sentiment classification using multiple source languages in multi-view semi-supervised learning, Eng. Appl. Artif. Intell. 36 (2014) 195–203.
[11] M.S. Hajmohammadi, R. Ibrahim, A. Selamat, Bi-view semi-supervised active learning for cross-lingual sentiment classification, Inform. Process. Manage. 50 (2014) 718–732.
[12] Z. Jingbo, W. Huizhen, B.K. Tsou, M. Ma, Active learning with sampling by uncertainty and density for data annotations, IEEE Trans. Audio Speech Lang. Process. 18 (2010) 1323–1331.
[13] T. Joachims, Making large-scale support vector machine learning practical, in: B. Scholkopf, C.J.C. Burges, A.J. Smola (Eds.), Advances in Kernel Methods, MIT Press, Cambridge, MA, USA, 1999, pp. 169–184.
[14] H. Kang, S.J. Yoo, D. Han, Senti-lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews, Exp. Syst. Appl. 39 (2012) 6000–6010.
[15] L.W. Ku, Y.T. Liang, H.H. Chen, Opinion extraction, summarization and tracking in news and blog corpora, in: Proceedings of AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs, AAAI, Palo Alto, California, 2006, pp. 100–107.
[16] Y. Leng, X. Xu, G. Qi, Combining active learning and semi-supervised learning to construct SVM classifier, Knowl.-Based Syst. 44 (2013) 121–131.
[17] D.D. Lewis, W.A. Gale, A sequential algorithm for training text classifiers, in: Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Springer-Verlag, Dublin, Ireland, 1994, pp. 3–12.
[18] M. Li, I.K. Sethi, Confidence-based active learning, IEEE Trans. Pattern Anal. Mach. Intell. 28 (2006) 1251–1261.
[19] S. Li, S. Ju, G. Zhou, X. Li, Active learning for imbalanced sentiment classification, in: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Association for Computational Linguistics, Jeju Island, Korea, 2012, pp. 139–148.
[20] M.T. Martín-Valdivia, E. Martínez-Cámara, J.M. Perea-Ortega, L.A. Ureña-López, Sentiment polarity detection in Spanish reviews combining supervised and unsupervised approaches, Exp. Syst. Appl. 40 (2013) 3934–3942.
[21] R. Mihalcea, C. Banea, J. Wiebe, Learning multilingual subjective language via cross-lingual projections, in: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, Prague, Czech Republic, 2007, pp. 976–983.
[22] T.S. Moh, Z. Zhang, Cross-lingual text classification with model translation and document translation, in: Proceedings of the 50th Annual Southeast Regional Conference, ACM, Tuscaloosa, Alabama, 2012, pp. 71–76.
[23] A. Montoyo, P. Martínez-Barco, A. Balahur, Subjectivity and sentiment analysis: an overview of the current state of the area and envisaged developments, Decis. Support Syst. 53 (2012) 675–679.
[24] J. Ortigosa-Hernández, J.D. Rodríguez, L. Alzate, M. Lucania, I. Inza, J.A. Lozano, Approaching sentiment analysis by using semi-supervised learning of multi-dimensional classifiers, Neurocomputing 92 (2012) 98–115.
[25] J. Pan, G.R. Xue, Y. Yu, Y. Wang, Cross-lingual sentiment classification via bi-view non-negative matrix tri-factorization, in: J. Huang, L. Cao, J. Srivastava (Eds.), Advances in Knowledge Discovery and Data Mining, Springer, Berlin/Heidelberg, 2011, pp. 289–300.
[26] B. Pang, L. Lee, S. Vaithyanathan, Thumbs up? Sentiment classification using machine learning techniques, in: Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2002, pp. 79–86.

[27] P. Prettenhofer, B. Stein, Cross-language text classification using structural correspondence learning, in: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Uppsala, Sweden, 2010, pp. 1118–1127.

[28] P. Prettenhofer, B. Stein, Cross-lingual adaptation using structural correspondence learning, ACM Trans. Intell. Syst. Technol. 3 (2011) 1–22.

[29] N. Roy, A. McCallum, Toward optimal active learning through sampling estimation of error reduction, in: Proceedings of the Eighteenth International Conference on Machine Learning, Morgan Kaufmann Publishers Inc., Williamstown, MA, USA, 2001, pp. 441–448.

[30] L. Shi, R. Mihalcea, M. Tian, Cross language text classification by model translation and semi-supervised learning, in: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Cambridge, Massachusetts, 2010, pp. 1057–1067.

[31] M. Tang, X. Luo, S. Roukos, Active learning for statistical natural language parsing, in: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics, Philadelphia, Pennsylvania, 2002, pp. 120–127.

[32] X. Wan, Using bilingual knowledge and ensemble techniques for unsupervised Chinese sentiment analysis, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Honolulu, Hawaii, 2008, pp. 553–561.

[33] X. Wan, Bilingual co-training for sentiment classification of Chinese product reviews, Comput. Linguist. 37 (2011) 587–616.

[34] B. Wei, C. Pal, Cross lingual adaptation: an experiment on sentiment classifications, in: Proceedings of the ACL 2010 Conference Short Papers, Association for Computational Linguistics, Uppsala, Sweden, 2010, pp. 258–262.

[35] Q. Wu, S.B. Tan, A two-stage framework for cross-domain sentiment classification, Exp. Syst. Appl. 38 (2011) 14269–14275.

[36] R. Xia, C. Zong, S. Li, Ensemble of feature sets and classification algorithms for sentiment classification, Inform. Sci. 181 (2011) 1138–1152.

[37] Y. Yang, J.O. Pedersen, A comparative study on feature selection in text categorization, in: ICML, 1997, pp. 412–420.

[38] J. Zhu, H. Wang, T. Yao, B.K. Tsou, Active learning with sampling by uncertainty and density for word sense disambiguation and text classification, Proceedings of the 22nd International Conference on Computational Linguistics, vol. 1, Association for Computational Linguistics, Manchester, United Kingdom, 2008, pp. 1137–1144.

[39] J. Zhu, M. Ma, Uncertainty-based active learning with instability estimation for text classification, ACM Trans. Speech Lang. Process. 8 (2012) 1–21.