Cross-lingual Polarity Detection with Machine Translation

Erkin Demirtas
Department of Computer Science
Eindhoven University of Technology
the Netherlands
e.demirtas@student.tue.nl

Mykola Pechenizkiy
Department of Computer Science
Eindhoven University of Technology
the Netherlands
m.pechenizkiy@tue.nl

ABSTRACT

Recent advancements in machine translation foster an interest of its use in sentiment analysis. In this paper, we investigate prospects and limitations of machine translation in sentiment analysis for cross-lingual polarity detection task. We focus on improving classification accuracy in a cross-lingual setting where we have available labeled training instances about particular domain in different languages. We experiment with movie review and product review datasets consisting of polar texts in English and Turkish. The results of the study show that expanding training size with new instances taken from another corpus does not necessarily increase classification accuracy. And this happens primarily not due to (not always accurate) machine translation, but because of the inherent differences in corpora between two subsets written in different languages. Similarly, in case of co-training classification with machine translation we observe from the results that accuracy improvement can be explained by semi-supervised learning with unlabeled data coming from the same domain, but not due to cross-language co-training itself. Our results also show that amount of artificial noise added by machine translation services does not hinder classifiers much in polarity detection task. However, it is important to distinguish the effect of machine translation from the effect of merging different cross-lingual data sources and that like in case of transfer learning we may need to search for ways to account for cross-lingual data distribution differences.

Categories and Subject Descriptors

H.2.8 [**Database Applications**]: Data Mining; I.5.2 [**Pattern Recognition**]: Design Methodology

General Terms

Experimentation, Algorithms, Performance

Keywords

multi-lingual polarity detection, machine translation

1. INTRODUCTION

Sentiment analysis is an emerging research area which already gathered a lot attention from the NLP community. It studies opinions, sentiments, appraisal, and emotions expressed in text. Recently, rapid growth of digital data and widespread information flow stimulate the development of computational methods in this field.

Although the volume of sentiment analysis research is increasing, majority of studies in the field still concentrates on English. There are different motivations for considering multi-lingual sentiment classification. From an analytics perspective we may want to focus on particular language or compare how much and what positive and negative sentiments are expressed per language of interest [13]. From machine learning perspective and NLP perspectives, multi-lingual and cross-lingual sentiment classification allows for using language specific models.

Many advanced tools developed for English are not available for other languages yet, which strains the applicability of sentiment analysis on other languages. The main motivation for multi-lingual sentiment analysis is of researchers from different countries want to build sentiment analysis systems in their own languages, but it is more than what it provides for each language at the individual level as it might contribute to our understanding of the global phenomena. Unfortunately the development of complex NLP tools i.e. parsers, taggers, and linguistic resources for each language is very costly and requires expensive human labor. In this regard, the potential of automated machine translation have been studied to leverage its capability, existing sentiment analysis resources and tools available in English to classify sentiments in other languages [3].

Language specific sentiment analysis mostly depends on the monolingual resources and tools that are available for that language. Previous research focus on improving multi-lingual sentiment analysis resulted in interesting attempts to leverage available resources using machine translation since in most cases they only exist for a limited number of languages.

In this paper, we study whether it is possible to improve classification accuracy in a cross-lingual setting where we have available (labeled) training instances about a particular domain in different languages.

We experiment with movie review and product review datasets consisting of polar texts in English and Turkish. There are already number of publicly available annotated corpora from these domains in English and we also crawled two annotated corpora in Turkish: the first is an annotated movie review corpus from publicly available website[1], and the second is an annotated multi-domain product review corpus from publicy available e-commerce website[2]. We constructed two benchmark datasets for each corpus that we make publicly available[3] for reproducibility of the experimental study and facilitation of further experimentation by other researchers.

The two goals of the experiments are to investigate 1) whether expanding training size with new machine translated instances taken from another corpus improves classification accuracy for the original corpus and 2) whether co-training with machine translation addresses cross-lingual polarity detection.

It is intuitive that quality of machine translation should be high such that instances from the dataset in the other language do not introduce noise to the training data. However, it is also important to realize that even when we consider the same application domain like movie reviews, due to cultural or other differences, there could be different biases in what people with different background (manifested in use of one or the other language) comment on, like or dislike.

Indeed, the results of the study show that expanding training size with new instances taken from another corpus does not necessarily increase classification accuracy. And this happens primarily not due to (not always accurate) machine translation, but because of the inherent differences in corpora between two subsets written in different languages. Similarly, in case of co-training classification with machine translation we observe from the results that accuracy improvement can be explained by semi-supervised learning with unlabeled data coming from the same domain, but not due to cross-language co-training itself.

Experiment results show that the expansion of our training set improves classification accuracy if we add new instances from same source (also when applying sequentially machine translation from English to Turkish and then back to English). However, when we introduce new training instances from another corpus, i.e. machine translated reviews from Turkish datasets, cross-lingual dissimilarities of two corpora overwhelms positive effects of having a larger training set – the classification accuracy for the target language dataset does not increase.

In our co-training with machine translation experiment we observed an improvement in classifying test data constructed from Turkish movie reviews, i.e. when test instances are in the other language. However, there is no improvement (over co-training iterations) for the reviews in English that is set as our target language. Thus, co-training with use of machine translation likely suffers from the same problem of cross-lingual dissimilarities of two corpora.

Thus, from our experimental study with two distinct applications of machine translation for cross-lingual polarity detection we can conclude two important facts. The quality of current machine translation techniques and services is already sufficient for improving cross-lingual sentiment classification (at least with the general-purpose classification techniques like SVMs). However, it is important to distinguish the effect of machine translation from the effect of merging different cross-lingual data sources and that like in case of transfer learning we may need to search for ways to account for cross-lingual data distribution differences.

The rest of the paper is organized as follows. We discuss related work in Section 2. The settings and the motivation behind this experimental study are explained in Section 3. Section 4 presents the details of the used datasets, experiment setup, and main results. Section 5 concludes with a summary of findings and directions for further research.

2. RELATED WORK

Previously authors developed methods to map sentiment analysis on English to other languages. Mihalcea et al. [10] propose a method to learn multilingual subjective language via cross-language projections. They use the Opinion Finder lexicon [16] and use two bilingual English-Romanian dictionaries to translate the words in the lexicon. Since word ambiguity can appear (Opinion Finder does not mark word senses), they filter as correct translations only the most frequent words. The problem of translating multi-word expressions is solved by translating word-by-word and filtering those translations that occur at least three times on the Web.

Another approach in obtaining subjectivity lexicons for other languages than English was explored by Banea et al. [4]. To this aim, the authors perform three different experiments, obtaining promising results. In the first one, they automatically translate the annotations of the MPQA corpus and thus obtain subjectivity annotated sentences in Romanian. In the second approach, they use the automatically translated entries in the Opinion Finder lexicon to annotate a set of sentences in Romanian. In the last experiment, they reverse the direction of translation and verify the assumption that subjective language can be translated and thus new subjectivity lexicons can be obtained for languages with no such resources.

Brooke et al. [6] experimented with translation from the source (English) to the target language (Spanish) and then used a lexicon-based approach or machine learning for target language document sentiment classification.

Steinberger et al. [12] create sentiment dictionaries in other languages using a method called ”triangulation”. They translate the data, in parallel, from English and Spanish to other languages and obtain dictionaries from the intersection of these two translations.

Duh et al. [8] presented their opinions about the research of multilingual sentiment classification, and they claimed that domain mismatch was not caused by machine translation (MT) errors, and accuracy degradation would occur even with perfect MT.

Balahur and Turchi [1] employ fully-formed machine translation systems, also study the influence of the difference in translation performance has on the sentiment classification performance. They report even in the worst cases, when the quality of the translated data is not very high, the drop in performance is of maximum 12%.

Similar to our work, Banea et al. [2] report an improvement in classification accuracy when using out-of-language features, yet our work differs from that in couple of major aspects. Our focus is polarity detection, rather than subjectivity analysis which they investigate. Moreover, their training set is only based on the machine translation of an English corpus, and they do not study how to make use of a new dataset from another language in training set.

In our study we investigate the approach for sentiment classification proposed by Wan [15] who constructs a polarity

[1]`http://www.beyazperde.com`

[2]`http://www.hepsiburada.com`

[3]The datasets are available at `http://www.win.tue.nl/~mpechen/projects/smm/#Datasets`

co-training learning system by using the multi-lingual views obtained through the automatic translation of product-reviews into Chinese and English. While [15] provides empirical evidence that leveraging cross-lingual information improves sentiment analysis in Chinese over what could be achieved using monolingual resources alone, it does not provide any results tested on samples taken from English dataset. Thus, as we show in our experimental study, the conclusions from the reported results in [15] should be interpreted with care.

3. CROSS-LINGUAL SENTIMENT CLASSIFICATION

In general if a text is classified as being subjective, we determine whether it expresses a positive or negative opinion. Structured information available in on-line movie reviews helps us in this regard to eliminate neutrality class as we can rely on user's rating associated on his/her review. We can detect polarity of a subjective review, therefore, based on classified instances on beforehand. However, in the real operational settings we would need to have a subjectivity detection mechanism or three-class polarity detection problem formulation for handling neutral messages. To keep the focus we experiment only with polar messages being either positive or negative.

We can consider cross-lingual sentiment classification as a special case of cross-domain classification settings since even two sources from different languages are from same domain they naturally represent different perspective with respect to cultural biases, hidden sentiments etc. We are tempted to explore how much these differences affect classification performance in a set of movie reviews as it may give hints about applicability of cross-domain classification research on cross-lingual sentiment analysis. We also want to see empirical evidences of introduced machine translation noise in sentiment classification and how much it puts a pressure on potential benefits of having a bigger training set which is expanded with machine translated instances.

We consider two distinct machine translation application scenarios. In the first scenario we simply use machine translation to use labeled instances in Turkish for expanding the training set in English considered as the target language for polarity detection.

In the second scenario we consider the co-training approach as viable alternative to leverage machine translated data as it was proposed in [15]. Although we construct labeled Turkish movie and product reviews during our research, for the co-training approach we regard those reviews as unlabeled to be able to setup the similar experimental settings (yet allowing for expanding the evaluation scenarios) and compare our findings with results reported in [15].

We consider the datasets and experiment setup for two scenarios in the following section.

4. EXPERIMENTAL STUDY

4.1 The benchmark

The following datasets are used in the experiments:

English movie reviews[4]**:** We use the sentence polarity data which was first introduced by [11]. This data consists of 5331 positive and 5331 negative snippets each containing roughly one single sentence. Reviews are gathered from Rotten Tomatoes web pages for movies released in 2002. They classify reviews marked with *fresh* are positive, and those marked with *rotten* are negative.

English multi domain product reviews[5]**:** This dataset was first introduced by [5]. It contains product reviews taken from Amazon.com from many product types. For our experiment we use a benchmark dataset which they constructed from four categories (books, dvd, electronics, and kitchen appliances) each consisting of 1000 positive and 1000 negative reviews.

Turkish movie reviews: We collect Turkish movie review dataset from Beyazperde web pages. In order to reach same size with the English dataset we restrict this dataset with 5331 positive and 5331 negative sentences. In this website, reviews are marked in scale from 0 to 5 by the same users who made the reviews. We consider a review positive if its rating is equal to or above 4, and negative if it is below or equal to 2.

Turkish multi domain product reviews: After building Turkish movie reviews dataset, we also collect Turkish product reviews from Hepsiburada.com (an online retailer operating in Turkey) to conduct our training set expansion experiment with reviews from different domains. We constructed another benchmark dataset also consisting reviews from books, dvd, electronics, and kitchen appliances categories to use them along with English product reviews. In this website, reviews are marked in scale from 1 to 5, and majority class of reviews converges to 5, that's why we have to consider a small amount of reviews marked with 3 stars as bearing a negative sentiment to be able to construct a balanced set of positive and negative reviews. It has 700 positive and 700 negative reviews for each of the four categories in which average rating of negative reviews is 2.27 and of positive reviews is 4.5.

For each experiment, part of these sets is used in training and evaluation phase, while the test set is always blind to the training phase. A small summary of the four dataset described above provided in Table 1. We explain in following sections how we use these datasets in our experiments.

4.2 Expanding training set with machine translated instances

A number of approaches have been proposed for polarity detection, including Prior Polarity classification (also with use of an opinion lexicon such as SentiWordNet[6], WordNet-Affect[7] or SenticNet[8]), statistical methods such as support vector machines, neural networks, and Naive Bayes among others. Aspect-based methods are introduced to spot more accurate sentiments on entities and their aspects. New approaches relying on semantic relationships in natural language concepts are also investigated under the concept-level sentiment analysis [7]. In our study we use three popular general purpose classification techniques; Naive Bayes, Support Vector Machines (Linear SVC), and Maximum Entropy (MaxEnt) classification.

As we have labeled datasets in English and Turkish, we can immediately apply any of the supervised learning approaches to build monolingual sentiment classifiers for both

[4]The dataset is available at `http://www.cs.cornell.edu/people/pabo/movie-review-data/`

[5]The dataset is available at `http://www.cs.jhu.edu/~mdredze/datasets/sentiment/`

[6]`http://sentiwordnet.isti.cnr.it/`

[7]`http://wndomains.fbk.eu/wnaffect.html`

[8]`http://sentic.net/downloads/`

Table 1: The summary of the datasets used in the experimental study.

	English Movie Reviews	Turkish Movie Reviews	English Product Reviews				Turkish Product Reviews			
			Books	DVD	Electronics	Kitchen Appliances	Books	DVD	Electronics	Kitchen Appliances
Positive	5331	5331	1000	1000	1000	1000	700	700	700	700
Negative	5331	5331	1000	1000	1000	1000	700	700	700	700

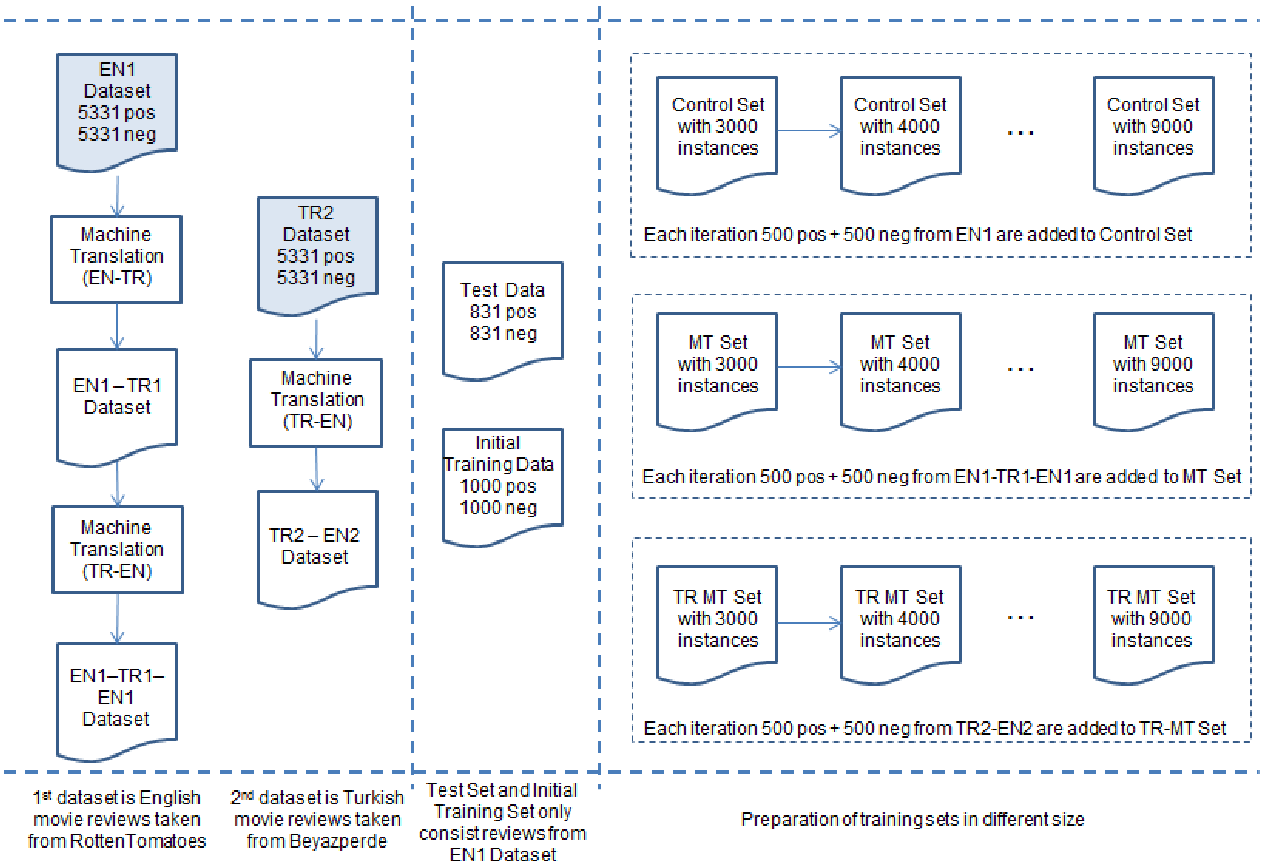


Figure 1: Study of training set expansion with machine machine.

languages. At this point, however, we can also investigate a way of improving classification accuracy of a monolingual classifier for the target language using annotated sources in different languages together. Previously a special case of this question was studied in [2], i.e. a pseudo parallel corpora constructed by machine translation services was used, and the focus was on subjectivity analysis. Their study suggested that the subjectivity classification accuracy can be increased by using features drawn from multiple languages. Our first experiment setting follows the idea of using multiple corpora in different languages but in a more generic way as we do not restrict these corpora to be parallel.

For this experiment, we prepare three types of training sets named as *control*, *machine translated*, and *Turkish machine translated* sets. The control set consists of only reviews from the English dataset. In order to measure the effect of machine translation (quality) we construct machine translated set which consists of reviews from English dataset as well, but then they first translated to Turkish and again back to English just to add artificial translation noise to their original form. Finally, we prepare Turkish machine translated set by compiling reviews from Turkish dataset which are translated to English. For all machine translation processes we use Google Translate service.

As Figure 1 shows, we first sample 1000 (400)[9] positive and 1000 (400) negative reviews from English movie reviews dataset to run the first iteration of the experiment for both training sets. Then, in every next iteration we increase the size of three training sets by adding 500 (100) positive and 500 (100) negative reviews taken from their respective sources. The test set is constructed from 831 (200) positive and 831 (200) negative English reviews that are never used in the training phase.

4.3 Co-training with machine translation

In [15] Wan proposed an application of the co-training method to make use of some amount of unlabeled Chinese product reviews to improve classification accuracy. For our second application scenario while preserving his main idea, we adopt it to our goals. First we use movie reviews instead of product reviews, and we experiment with Turkish-English language setting while Wan uses Chinese-English. These are mostly practical changes in the framework, however, we test combined classifier with reviews taken from both Turkish and English datasets whereas Wan only present results based on test data containing Chinese texts only.

As we can see in Figure 2, training input is the labeled English reviews and some amounts of unlabeled Turkish reviews. The labeled English reviews are translated into labeled Turkish reviews, and the unlabeled Turkish reviews are translated into unlabeled English reviews, by using Google Translate. Therefore, each review is associated with an English version and a Turkish version. The English features and the Turkish features for each review are considered two independent and redundant views of the review.

The co-training Algorithm 1 is then applied to learn two classifiers.

The English and Turkish terms (features) used in our study include unigrams; the feature weight is simply set to term presence following the bag-of-words model. The output value of the Naive Bayes classifier for a review indicates the confidence level of the review's classification. In the training phase, the co-training algorithm learns two separate classifiers: C_{en} and C_{tr}. Therefore, in the classification phase, we can obtain two prediction values for a test review, and the average of these values is used as the overall prediction value of the review.

[9]numbers in parentheses refer to the setting for product review datasets; without parentheses - to the moview review dataset

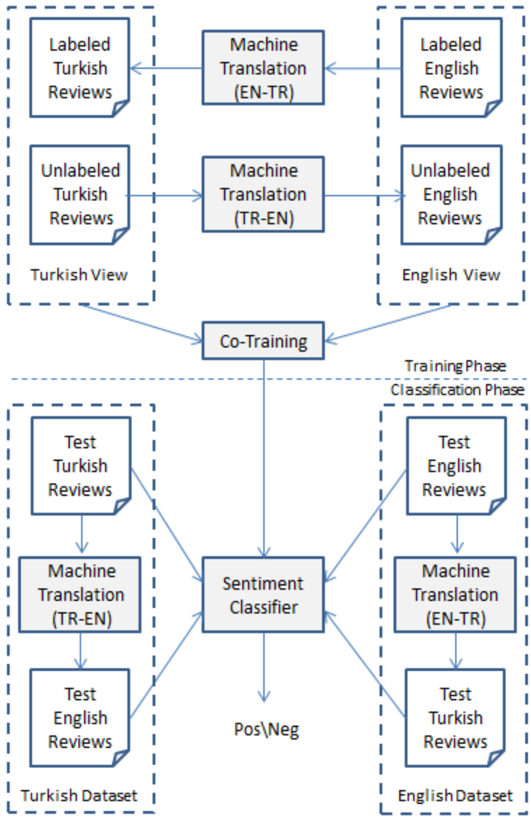


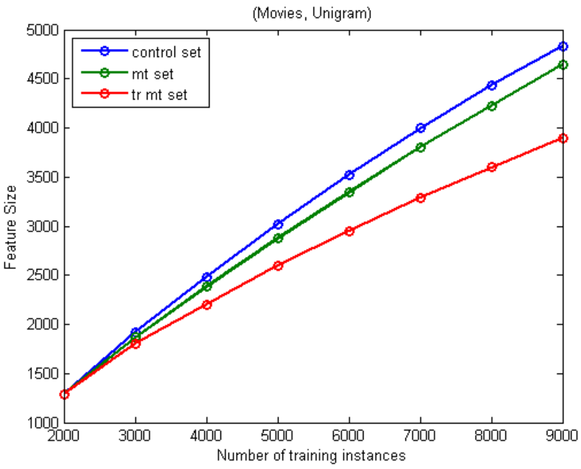
Figure 2: Co-training experiment setup.

4.4 Results and Discussion

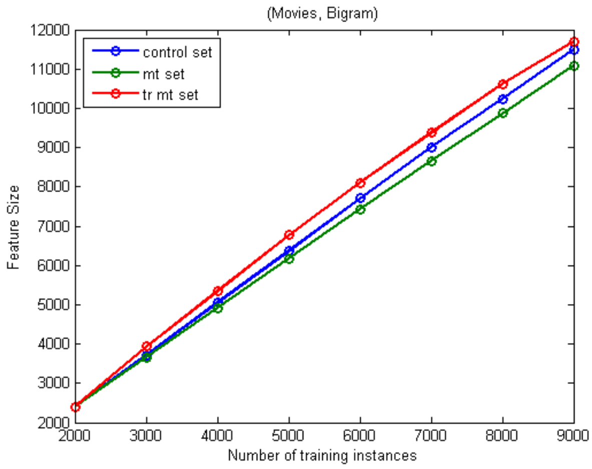
For the training set expansion experiment we present our results in terms of two metrics. First, we measure the feature size increase as we keep adding new instances to the training sets.

The two graphs in Figure 3 show feature size change of movie reviews datasets in which our training sets are represented by unigram and unigram plus bigram features respectively. We observe an interesting behavior of the feature size change in Turkish machine translated set. Despite its slope is smaller in case of unigram feature representation, when we look at bigram representations it produces more features than the any other set does. Relative poor increase in unigram feature size can be explained by the data loss happened during machine translation as such a number of Turkish words could not be translated to English. On the other hand, machine translation introduces some amount of noise as well which portrays itself by producing a vast number of meaningless bigrams.

Accuracy results of the Naive Bayes classifiers on movies reviews datasets are summarized in Figure 4. We can observe some interesting results. First, consistent with our expectation, expanding training size by adding new instances from the same corpus improves the overall accuracy. This



(a) Unigram



(b) Unigram + Bigram

Figure 3: Feature size comparison for the training set expansion experiment.

Algorithm 1 Co-training two classifiers

Input: F_{en} and F_{tr} are redundantly sufficient sets of features, where F_{en} represents the English features, F_{tr} represents the Turkish features, L is a set of labeled training reviews, U is a set of unlabeled reviews
Output: two classifier C_{en} and C_{tr}
for $i \in \{1, 2, \cdots, k\}$ **do**
  Learn the first classifier C_{en} from L based on F_{en}
  Use C_{en} to label reviews from U based on F_{en}
  Choose p positive and n negative the most confidently predicted reviews E_{en} from U
  Learn the second classifier C_{tr} from L based on F_{tr}
  Use C_{tr} to label reviews from U based on F_{tr}
  Choose p positive and n negative the most confidently predicted reviews E_{tr} from U
  Removes reviews $E_{en} \cup F_{tr}$ from U
  Add reviews $E_{en} \cup E_{tr}$ with the corresponding labels to L
end for
return C_{en}, C_{tr}

behavior can be noted following the control set results for both graphs in Figure 4. Machine translation set slightly under-performs than the control set due to the negative effect of machine translation quality, and this difference tends to increase slightly as we add more machine translated sentences to the training set. Nevertheless, the overall effect of machine translation in this case is positive. We can observe 5% increase in accuracy. The results corresponding to the use of Turkish machine translated set (red line fluctuating between 69% and 70%) clearly shows that naive cross-lingual training set expansion does not improve the generalization performance of polarity detection, although we do gather more features from new instances translated from Turkish movie reviews. This problem refers to cross-domain classification as we can regard new features from Turkish reviews as ones from another domain which is not really immediately helpful to classify the test instances taken from the English dataset. These results suggest that an application of resolving cross-corpora dissimilarity may help to utilize labeled instances taken from another language in cross-lingual sentiment analysis.

Table 2: Naïve Bayes classification performance

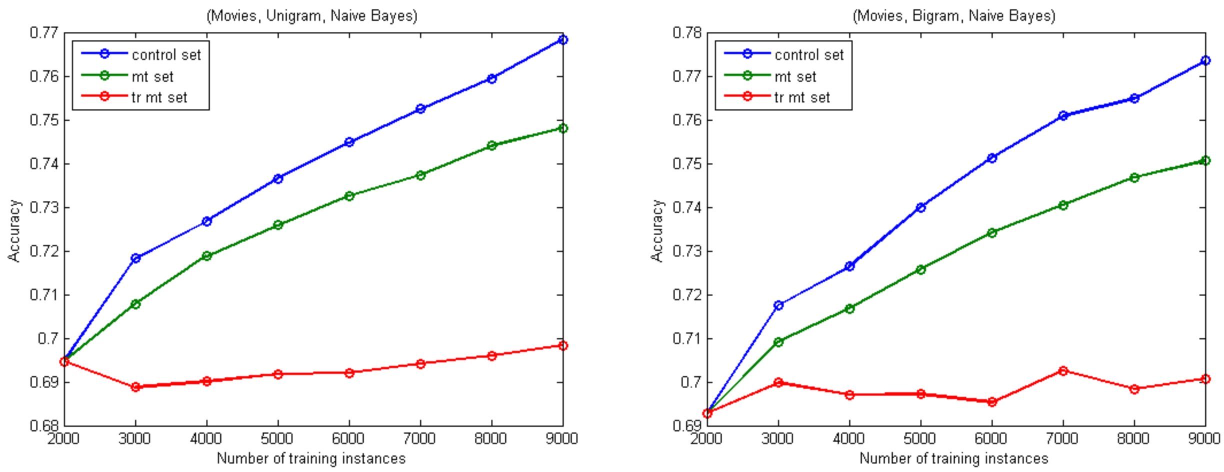
	Initial accuracy	Control set	MT set	TR MT set
Movies	69.5	+10.6	+7.7	+0.5
Books	72.4	+9.2	+8.6	-0.7
DVD	76.0	+4.6	+1.5	-1.1
Electronics	73.0	+8.1	+9.6	-8.6
Kitchen	75.9	+7.2	+8.7	-6.3

Table 3: Linear SVC classification performance

	Initial accuracy	Control set	MT set	TR MT set
Movies	66.0	+11.3	+8.2	+0.5
Books	66.6	+11.1	+14.0	+0.3
DVD	70.3	+7.7	+8.0	-2.7
Electronics	72.4	+7.2	+5.0	-8.0
Kitchen	70.0	+12.3	+11.1	-2.7

This behavior of Naïve Bayes classifier is very similar for Linear SVC and MaxEnt classifiers, and it also generalized to all five datasets we experimented with. The summary of the classification performance is given in Tables 2, 3 and 4. In each table in the first column we give the baseline performance on the initial training data, and the following three columns show the absolute increase (or decrease) in the classification accuracy after the additional training data was added in full according to one of the three setups. We can see from the tables that expanding the training set with additional labeled instances from the same source helps to improve the classification performance and from the different source - does not, and in fact on three datasets even deteriorates the performance.

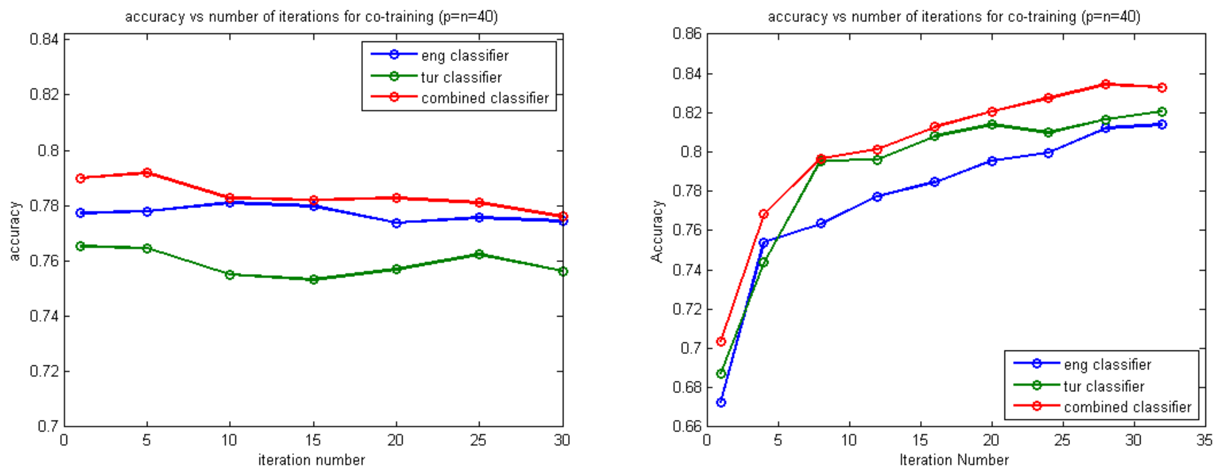
Co-training experiment results give us insightful details to compare our findings with the ones reported by Wan in [15]. In his paper, Wan evaluates the co-training algorithm by classifying labeled Chinese reviews that are taken from same website and which he used in training phase. We present our



(a) Unigram

(b) Unigram + Bigram

Figure 4: Generalization accuracies for the training set expansion experiment.



(a) English dataset

(b) Turkish dataset

Figure 5: Accuracy comparison for the co-training experiment.

Table 4: MaxEnt classification performance

	Initial accuracy	Control set	MT set	TR MT set
Movies	68.2	+11.0	+8.8	+0.4
Books	68.7	+12.8	+12.4	+1.8
DVD	71.8	+9.5	+9.6	+1.1
Electronics	74.0	+9.5	+8.0	-7.7
Kitchen	72.4	+12.7	+12.2	-2.2

results based on labeled Turkish movie reviews corresponding to his labeled Chinese reviews, but also the results based on labeled English movie reviews that are discarded from the training phase. Figure 5b confirms findings reported in [15]: tested on labeled Chinese product reviews the combined classifier performs the best and overall accuracy for all classifiers increases in each iteration of co-training. However, co-training approach fails to improve classification accuracy tested on samples from English dataset as we run the algorithm for multiple iteration. For all classifiers (Turkish, English, and combined) we get the highest accuracies with the first iteration that do get better with more iterations. Since proposed co-training approach leverages only unlabeled Chinese reviews (in our work these are replaced by unlabeled Turkish reviews) it resembles semi-supervised learning that aims to increase the classification performance with the aid of some unlabeled data in a language which is the same as the language of the test set. Therefore most of the performance gain presented in [15] is likely due to semi-supervised learning rather than the aid of the English classifier.

5. CONCLUSION AND FUTURE WORK

In this paper, we examined some of the possible improvements in sentiment classification by leveraging labeled or unlabeled data in different languages.

Our experiments show that naive ways of introducing new sources from other languages causes cross-domain dissimilarity issues. This indicates that existing approaches applicable to cross-domain sentiment classification, e.g. [9] and further advancement in this direction might be fruitful for cross-lingual sentiment analysis too. This is one of the directions of our future work.

In this paper we studied how machine translation affects the performance of the general purpose classification techniques. In the future work we plan to consider also techniques specific to sentiment classification like e.g. a rule-based approach to polarity detection [14].

6. REFERENCES

[1] A. Balahur and M. Turchi. Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis. *Computer Speech and Language*, 2013 (In press).

[2] C. Banea, R. Mihalcea, and J. Wiebe. Multilingual subjectivity: are more languages better? In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING'10, pages 28–36, 2010.

[3] C. Banea, R. Mihalcea, and J. Wiebe. *Multilingual Sentiment and Subjectivity Analysis*. Prentice Hall, 2011.

[4] C. Banea, R. Mihalcea, J. Wiebe, and S. Hassan. Multilingual subjectivity analysis using machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 127–135, 2008.

[5] J. Blitzer, M. Dredze, and F. Pereira. Biographies, bollywood, boomboxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, ACL'07, pages 187–205, 2007.

[6] J. Brooke, M. Tofiloski, and M. Taboada. Cross-linguistic sentiment analysis: From english to spanish. In *Proceedings of RANLP'2009*, pages 50–54.

[7] E. Cambria, B. Schuller, Y. Xia, and C. Havasi. New avenues in opinion mining and sentiment analysis. *Intelligent Systems, IEEE*, 28(2):15–21, 2013.

[8] K. Duh, A. Fujino, and M. Nagata. Is machine translation ripe for cross-lingual sentiment classification? In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2*, HLT '11, pages 429–433, 2011.

[9] X. Glorot, A. Bordes, and Y. Bengio. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *Proceedings of the Twenty-eight International Conference on Machine Learning (ICML'11)*, volume 27, pages 97–110, June 2011.

[10] R. Mihalcea, C. Banea, and J. Wiebe. Learning multilingual subjective language via cross-lingual projections. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 976–983, 2007.

[11] B. Pang and L. Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL'05, pages 115–124, 2005.

[12] J. Steinberger, M. Ebrahim, M. Ehrmann, A. Hurriyetoglu, M. Kabadjov, P. Lenkova, R. Steinberger, H. Tanev, S. Vázquez, and V. Zavarella. Creating sentiment dictionaries via triangulation. *Decis. Support Syst.*, 53(4):689–694, 2012.

[13] E. Tromp and M. Pechenizkiy. Senticorr: Multilingual sentiment analysis of personal correspondence. In *Proceedings of IEEE ICDM 2011 Workshops*, pages 470–479. IEEE, 2011.

[14] E. Tromp and M. Pechenizkiy. Rbem: A rule based approach to polarity detection. In *Proceedings of the Workshop on Issues of Sentiment Discovery and Opinion Mining (WISDOM@KDD'13)*. ACM, 2013.

[15] X. Wan. Co-training for cross-lingual sentiment classification. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1 - Volume 1*, ACL'09, pages 235–243, 2009.

[16] T. Wilson, J. Wiebe, and P. Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 347–354, 2005.