Cross-lingual Subjectivity Detection for Resource Lean Languages

Aida Amini[1], Samane Karimi[2, 3], and Azadeh Shakery[2, 4]

[1]School of Computer Science and Engineering, University of Washington
[2]School of Electrical and Computer Engineering, College of Engineering, University of Tehran
[3]Computer Science Department, University of Houston
[4]School of Computer Science, Institute for Research in Fundamental Sciences (IPM)University of Tehran
amini91@cs.washington.edu
{*samanekarimi, shakery*}*@ut.ac.ir*

Abstract

Wide and universal changes in the web content due to the growth of web 2 applications increase the importance of user-generated content on the web. Therefore, the related research areas such as sentiment analysis, opinion mining and subjectivity detection receives much attention from the research community. Due to the diverse languages that web-users use to express their opinions and sentiments, research areas like subjectivity detection should present methods which are practicable on all languages. An important prerequisite to effectively achieve this aim is considering the limitations in resource-lean languages. In this paper, cross-lingual subjectivity detection on resource lean languages is investigated using two different approaches: a language-model based and a learning-to-rank approach. Experimental results show the impact of different factors on the performance of subjectivity detection methods using English resources to detect the subjectivity score of Persian documents. The experiments demonstrate that the proposed learning-to-rank method outperforms the baseline method in ranking documents based on their subjectivity degree.

1 Introduction

Rapid growth of web 2 applications lead to an increase in textual content generated by users such as comments, reviews and any kind of textual data reflecting peoples opinions. In text mining literature, this kind of data is known as subjective data. Consequently, many automatic methods for detecting this kind of data from objective ones and in the next step, for detecting the polarity of subjective data have been proposed. These methods form one of the main research areas in text mining called subjectivity and sentiment analysis. Most papers in this area propose methods for sentiment analysis on English whereas users of web 2 ap-

plications are from a wide range of languages so there is a serious need for making sentiment analysis possible on other languages. This goal is achieved by different approaches proposed in this research area. Some papers present methods for sentiment analysis on non-English languages such that English resources are translated to the target language using dictionaries, machine translation or other tools, then a sentiment analysis method is proposed for the non-English language using the translated resources.

Some other papers propose a cross-lingual framework for sentiment analysis. In this approach, passing the language boundaries is an internal step which is situated within the whole procedure. Some other papers address sentiment analysis on multi-lingual environments. The main goal in these papers is to employ all useful information from different languages to facilitate subjectivity detection and sentiment analysis on multi-lingual documents.

In this paper, the problem of subjectivity detection in cross-lingual case is studied to see how the sentiment resources in resource-rich languages like English can be used to achieve high performance in subjectivity detection systems of resource-lean languages like Persian. The methods employed as subjectivity detection methods in this paper are a language-model-based method and a method based on learning-to-rank techniques that are implemented in cross-lingual case.

Our experiments show that some factors including translation direction and translation unit have an impact on the performance of cross-lingual subjectivity detection methods. Furthermore, experimental results show that learning to rank approach leads to high performance in subjectivity detection task.

The rest of this paper is organized as follows. In section 2, some of the most relevant studies in

subjectivity detection and sentiment analysis area are reviewed. Section 3 describes two subjectivity detection methods implemented in this paper. Experimental results are explained in section 4. The paper is concluded by future work and conclusion section.

2 Related Work

Most of previous studies in sentiment analysis area has been applied to English data. Sentiment analysis on other languages and on multi-lingual environments and in cross-lingual settings which is using resources of one language like English to do sentiment analysis on another language, has attracted a great deal of attention in recent years so a large number of papers focus on these areas. Previous studies related to these areas are explained in the rest of this section.

2.1 Cross-lingual Sentiment analysis

Ku et al. (Ku et al., 2006) propose algorithms for opinion extraction and summarization for Chinese. To this aim, they provide the necessary Chinese sentiment lexicon by translating the available English one, namely General Inquirer. In an Italian one, Esuli and Sebastiani (Esuli and Sebastiani, 2009) propose an opinion extraction system using lexical resources to improve the performance of the proposed opinion extraction system.

Banea et al. (Banea et al., 2008) have studied the performance of automatic translation in a sentiment analysis system where training sources are in a resource-rich language and test sources are in another language. Authors have done comparative evaluations on Romanian and Spanish through three different experiments. In all three experiments, English is used as a source language which both a manually annotated corpus (MPQA) and a subjectivity analysis tool (OpinionFinder) is available.

Another paper on Spanish sentiment analysis is proposed by Brooke (Brooke et al., 2009). In this paper, a lexicon-based sentiment analysis system is adapted to Spanish such that semantic orientation of each word in Spanish is computed using the translated lexicon.

Steinberger et al. (Steinberger et al., 2011a) propose a method that starts by construction of a sentiment dictionary in two languages (English and Spanish). In the next step, parallel data from English and Spanish are translated to the third language (Arabic, Czech, French, German, Italian and Russian) and the new sentiment dictionary is obtained from the intersection of the translations.

In another paper of Steinberger et al. (Steinberger et al., 2011b), the construction and employment of a parallel corpus with sentiment labels is studied. This paper proposes to detect the polarity of opinions that are about entities like persons, organizations and concepts across different languages. A simple method is selected for polarity detection in this paper, which adds up positive and negative scores in six-word windows around the entities. The sentiment scores of the words are computed using the sentiment dictionaries created by triangulation method (Steinberger et al., 2011a).

Bautin et al. (Bautin et al., 2008) propose a method to detect the sentiment label of news articles gathered from online newspapers in nine languages including Arabic, Chinese, English, French, German, Italian, Japanese, Korean, and Spanish. In this paper, machine translation is used to transfer textual documents into English. Then, a sentiment analysis system entitled Lydia is employed to detect the translated documents labels.

Wei and Pal (Wei and Pal, 2010) propose a cross-lingual sentiment analysis system that uses SCL(Structural Correspondence Learning) technique to find an efficient representation for documents which is shared in both languages.

One of the latest papers in this area address the aspect-based sentiment analysis in cross-lingual setting (Lambert, 2015). In this paper, Lambert propose a method which translates opinionated segments of the source language under some constraints such that their translation in target language would not be reordered.

2.2 Multi-lingual sentiment analysis

Banea et al. (Banea et al., 2014) uses classification methods for subjectivity detection in a multilingual environment using the alignments between word senses in different languages from wordNet, namely English and Romanian.

Balahur and Turchi (Balahur and Turchi, 2014) propose to investigate sentiment detection and classification on different languages other than English. Three languages including German, Spanish and French are selected for this aim. In this article, three machine translation systems such as Google, Bing and Moses is investigated and its

results show that the machine translation method can lead to high performance in sentiment detection and classification similar to the performance in the original language (English).

In Banea et al. (Banea et al., 2010) a labeled English dataset is translated to five other languages including Romanian, Spanish, French, German and Arabic. Then some multilingual versions of the English datasets based on all possible combinations of these six languages are generated and used to train Nave Bayes classifiers with unigram features.

Wan (Wan, 2009) propose a sentiment analysis system whose idea is similar to (Banea et al., 2010). Authors in (Wan, 2009) use the multilingual views to the dataset by automatic translation of reviews. The English reviews are translated to Chinese and the paper illustrates that the proposed cross-lingual sentiment analysis system outperforms the mono-lingual one.

To evaluate the multi-lingual comparability of multi-lingual subjectivity analysis systems, Kim et al. (Kim et al., 2010) have presented an evaluation method. In this method, performance of different subjectivity analysis systems including a corpus-based method, a lexicon-based method and OpinionFinder (a well-known tool for subjectivity analysis) is measured on multi-lingual data on English, Korean, Japanese and Chinese.

3 Methodology

The main aim of this paper is to distinguish subjective text from objective ones. The task of identifying documents containing subjective text is appealing due to the fact that subjectivity detection is a preliminary step before other sentiment analysis tasks such as polarity detection. In many subjectivity detection methods there is a need for a collection of documents with two labels: subjective and objective. The main aim of this paper is to explore and investigate different methods for making the full use of labeled datasets from resource rich languages like English (as train data) to improve quality of subjectivity detection in resource lean languages like Persian (as test data).

Having both test and train datasets in the same language, would cause less ambiguities in subjectivity detection results in comparison with having them in different languages. Consequently, better subjectivity detection performance is expected when both train and test datasets are in the same language.

In cross-lingual domains, the similarity between the source and target language and the quality of translation tool are critical in the performance. Each language has its own ambiguities. For example, both words "milk" and "lion" translate to the same word in Persian or the word "pretty" has two meanings in English. Each of these ambiguities affect both monolingual and cross-lingual results. However, these effects may be more catastrophic in cross-lingual cases since we are dealing with ambiguities in both languages and the errors caused by incorrect translations. In this paper, two different approaches for crossing the language borders are investigated:

In the first approach, resources of the source language are translated to the target language, then a model is built using the translated data and finally the model is applied on the test data (source translation).

The second approach uses an inverse translation direction. In this approach, first, a model is built using the data in the source language. Simultaneously, the data in the target language is translated to the source language. Finally, the model is applied on the translated test data (target translation).

These two approaches are studied by means of two different subjectivity detection methods including a language-model based method and a learning-to-rank method. In the following sections, the details related to each of the mentioned methods are further discussed.

3.1 Cross lingual Subjectivity Detection

In cross lingual subjectivity detection, the test data is form the resource lean language(Persian) and the train data is from the resource rich languages(English). Due to the different languages of the train and test data, there is a need to cross the language borders in one of the steps of cross lingual subjectivity detection process. Translation phase adds some more challenges to the problem in terms of translation direction and unit.

Translation direction is one of the factors that affects the translation quality due to the different ambiguities in each language.

Translation unit Translation can be done in word or sentence level that leads to different results. For example, the context available in sentence level translation can help the translator solve

the the ambiguity of words' different meanings. Therefore, the translation unit can be another factor to be considered. These two factors are considered in each of the proposed methods, i.e. the language-model based method and the learning-to-rank method.

3.1.1 Language-Model based method

Language modeling is an approach which has been widely used in Information retrieval recently. In this paper, we employed the language-model based subjectivity detection method proposed in (Karimi and Shakery, 2017). In (Karimi and Shakery, 2017), each test document is assigned a subjectivity score based on its similarity to the language models of subjective and objective train datasets. This score is computed from the difference of the similarity between the test document language model and the language model of subjective train dataset (subjective model), $sim_{subj}(d)$, and the similarity between the test document language model and the language model of objective train dataset (objective model), $sim_{obj}(d)$.

$$score(d) = sim_{subj}(d) - sim_{obj}(d) \quad (1)$$

The subjective and objective models are built over the unigrams of subjective and objective documents in the train dataset. Each unigram in each model is assigned a value representing the its occurrence probability in the subjective or objective documents. In this approach, the effect of each word on the subjectiveness and objectiveness of the document is measured. For example the word "opinion" would appear more frequently in subjective documents. In addition to that, the effect of neutral words(propositions or modal verbs) can be ignored with the usage of an appropriate scoring formula since these words are present almost equally in both categories. In our method, we utilize this approach to compute the subjectivity score of a document in a cross-lingual setting. The details of how the similarity values (i.e. $sim_{subj}(d)$ and $sim_{obj}(d)$) are computed in our method are explained in the rest of this section.

As mentioned before, the translation unit and translation direction are two important factors affecting cross-lingual subjectivity detection performance. The proposed language-model based method is explained considering these two factors. Words and documents are two translation units considered in this paper:

- Word level translation: In this case, the translation phase follows the training phase. The translation can be done in two directions.

From English to Persian (source translation): In this case, the translation should be applied to the reference language models, namely subjective and objective models. These models consist of unigrams and their occurrence probability, so the translation can be easily applied to the models. The translated reference language models are computed as follows.

$$p(f|\theta_{subj}) = \sum_w p(f|w, \theta_{subj})p(w|\theta_{subj})$$
$$\approx \sum_w p(f|w)p(w|\theta_{subj}) \quad (2)$$

$$p(f|\theta_{obj}) = \sum_w p(f|w, \theta_{obj})p(w|\theta_{obj})$$
$$\approx \sum_w p(f|w)p(w|\theta_{obj}) \quad (3)$$

$$p(w|\theta_{subj}) = \frac{\sum_{d \in D_{subj}} c(w, d)}{\sum_{d \in D_{subj}} |d|} \quad (4)$$

$$p(w|\theta_{obj}) = \frac{\sum_{d \in D_{obj}} c(w, d)}{\sum_{d \in D_{obj}} |d|} \quad (5)$$

where $c(w, d)$ represents the frequency of word w in document d and $|d|$ is the length of document d. D_{Subj} and D_{obj} represent the set of subjective and objective documents in the train dataset respectively. f, w and $p(f|w)$ represent a Persian word, an English word and the probability that f translates to w using the employed translation tool (i.e. translation probability) respectively. $p(w|\theta_{subj})$ and $p(w|\theta_{obj})$ stands for the occurrence probability of word w in the primal (before translation) subjective and objective language models that are calculated using the Eq. (4) and the Eq. (5). Accordingly, $p(f|\theta_{subj})$ and $p(f|\theta_{obj})$ represent the word probabilities in the translated subjective and objective language models respectively.

The outputs of the translation phase are the reference language models in the target language,
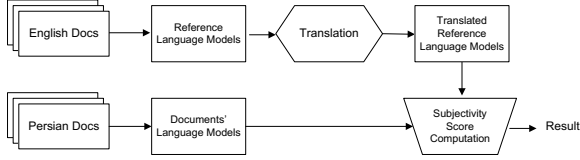
Figure 1: Language model-based subjectivity detection from English to Persian

namely, the translated subjective and objective models which can be used for computing the subjectivity score of a test document d. To this aim, first, we need to calculate the similarity between the language model of d and the subjective model, i.e. $sim_{subj}(d)$, and also the similarity between the document's language model and the objective model, i.e. $sim_{obj}(d)$. The similarity between the language models are measured using kl-divergence formula according to the Eq. (6) and the Eq. (7):

$$sim_{subj}(d) = \sum_{f \in d} -p(f|\theta_d) . \log \frac{p(f|\theta_d)}{p(f|\theta_{subj})} \quad (6)$$

$$sim_{obj}(d) = \sum_{f \in d} -p(f|\theta_d) . \log \frac{p(f|\theta_d)}{p(f|\theta_{obj})} \quad (7)$$

Figure 1 illustrates this procedure in more details.

Persian to English (target translation): In this case, the Persian unigrams are mapped to the model. For each of the Persian documents, the unigrams are derived and translated to English. Therefore, the model in the source language can be used over the translated unigram language model of the test document. In this case, the reference language models are computed the same as mono-lingual case using the Eq. (4) and the Eq. (5) and the documents language model is translated according to the Eq. (8).

$$p(w|\theta_d) = \sum_f p(w|f, \theta_d)p(f|\theta d))$$
$$\approx \sum_f p(w|f)p(f|\theta d) \quad (8)$$

where $p(w|\theta_d)$ represents the word probabilities in the translated document language model and $p(f|\theta_d)$ represents the word probabilities in the primal (before translation) document language model. Final subjectivity score is computed according to the Eq. (1).

• Document level translation: This level of translation is independent of the direction of the translation. So, the translation is applied as a preprocessing step. Then, the training and testing steps would be similar to monolingual situation.

3.1.2 Learning to rank

Learning to rank approach in information retrieval refers to a method using machine learning techniques to rank documents based on their relevance to a query. Therefore, to use this approach in subjectivity detection task a query list is needed. In this paper, for each language a list of subjective terms is specified as the query list. These lists are the subjective portion of a sentiment lexicon in each language which is specified by selecting terms with higher subjectivity weights. Then, the relevance of documents to the queries in the list is used as a measure of their subjectivity level. In other words, the learning to rank method computes the relevance degree of each test document to the query and the final result is a ranked list of documents for each query. The query set used in this paper is a sentiment lexicon constructed in (Dehdarbehbahani et al., 2014) which is employed in two manners:

• List query level: In this manner, all of the words in the sentiment lexicon are assumed to be one enormous query.

• Term query level: The assumption of having a query with the length of more than 1000 words can be inane. Therefore, in this section, each term of the lexicon is used as a separate query so the number of feature vectors would be the multiplication of the number of queries and documents. The final output should be the scores assigned to the test documents. As mentioned above, there would be q*d feature vectors, hence, q*d individual scores. To compute a single score for each test document, as the desired output of the subjectivity detection method, a weighted average over scores obtained from each query term q_i of the query q can be computed in this manner as below:

$$score(d, q) = \sum_{q_i \in q} \frac{w(q_i).score(d, q_i)}{|q|} \quad (9)$$

where $w(q_i)$ is the weight of each query term q_i in the sentiment lexicon. $score(d, q_i)$ is

the score of each test document d with query term $, q_i$ obtained from the learning to rank method.

In this approach, the translation can be done via two procedures. These procedures are figured based on the feature set used in the learning to rank method as explained below:

- Translated Unigrams: In the first procedure which unigrams are used as query independent features, the translation phase is performed simply on the unigrams of the train dataset.

- Query dependent features: These features are computed using the probabilistic translations for query terms as described below:

 a) Term Frequency: This feature is calculated according to the following relations in cross lingual situation:

$$TF(f_i, d_f) = c(f_i, d_f) \tag{10}$$

$$CLTF(e_j, d_f) = \sum_{f_i} p(f_i|e_j).TF(f_i, d_f) \tag{11}$$

$$CL\_Feat1(d, e) = \sum_{e_j \in e} \log(CLTF(e_j, d)) \tag{12}$$

 where d_f represents the document in Persian, e_j represents the query term in English and e represents the whole query, f_i represents the ith translation of e_j, $p(f_i|e_j)$ shows the translation probability of e_j to f_i and $c(f_i, d_f)$ represents the frequency of the query word f_i in document d_f.

 b) Inverse Document Frequency: This feature is calculated as below:

$$IDF(e_j) = \frac{N}{|d \in D| e_j \in D|} \tag{13}$$

$$CLIDF(e_j) = \sum_{f_i} p(f_i|e_j).IDF(f_i) \tag{14}$$

$$CL\_Feat2(e) = \sum_{e_j \in e} \log(CLIDF(e_j)) \tag{15}$$

 where N represents the number of documents in the collection, d represents the test document, D represents the collection of documents, e_j represents the query term and e represents the whole query.

 c) BM25: This feature provides a measurement of the relevance degree of the document to the query. This feature is calculated as below:

$$CL\_Feat3(d, q) = \sum_{q_i \in q} CLIDF(q_i) \times \\ \frac{c(q_i, d).(k_1 + 1)}{c(q_i, d) + k_1.(1 + b - b.\frac{|d|}{avg(dl)})} \tag{16}$$

 Where $|d|$ represents the document length and $avg(dl)$ is the average length of the documents in the collection and $c(q_i, d)$ is the frequency of query term q_i in document and d. k_1 and b are constant parameters used for determining the effect of query term frequency and normalizing the document length respectively.

 d) KL-Divergence: This feature that shows the relevance degree of the document to the query is computed according to the Eq. (8) where the document language model is computed based on the probabilistic translation (Azarbonyad et al., 2012).

 e) Document length: this feature shows the length of the document.

4 Experimental Results and Discussion

4.1 Datasets

The datasets used in this paper consist of documents about movies. Movie is a difficult domain to analyze since it contains a great variety of words based on the movie story while comments regarding other products can mostly be narrowed to a set of technical words that are used in the products domain. The datasets used in this paper are in two languages:

English Dataset : This dataset (Ku et al., 2006) contains 5000 movie reviews from Rotten Tomatoes that form subjective documents and 5000 movie summaries from the Internet Movie Database that form objective documents of the dataset. The average document length in this dataset is 11 words.

Persian Dataset : As there was no Persian dataset in movie domain with subjective and objective labels, it is constructed to be employed in this paper. Subjective documents are gathered from the websites containing movie critics such as: www.naghdefarsi.com, myturn.blogsky.com,

yasserbayani.persianblog.ir. To ensure that all documents gathered from these websites are opinionated, only some specific URL patterns are followed. More than 7500 subjective documents are accumulated using these URLs. For the objective part of this dataset, Wikipedia is used as the main information resource. This website poses information regarding the movies that are produced over the past century so it contains objective data about movies. To ensure that none of documents contains any opinions, only the textual content under specific titles are collected. These titles are Actors/Actresses, Awards and Movie Story. All these data, if existed, are considered as one document. In addition to Wikipedia content, the paragraphs that start with movie story title from naghdefarsi website are also added to the objective dataset. In total, 3500 objective documents are gathered. The average document length in this dataset is 83.

Language Differences There are major distinctions between Persian and English languages including the difference in the sentence structure, negative and modal verb formations, and the use of adjectives. In contrast to English language, the verb in Persian appears at the end of the sentence. Negative verbs are constructed by adding *ne* to the beginning of the verb in contrast of having separate *not*. The usage of adjectives, which are also critical in subjectivity detection, are different in a way that they usually appear after the word they are describing.

4.2 Translators

Translators accuracy has a great impact on the final results. The more accurate the translator, the closer to mono lingual results would be. In addition, based on the methodology employed, different kinds of information regarding the translation are needed. For the language-model based method, Google translate is used as the translation tool. The source text is sent to this tool in packages via 100 separate threads. In word level translation, the unigrams of the dataset are submitted as input and the outcome is the list of possible translations. In this case, the first translation is considered to be the most probable translation. In the document level translation, the whole documents are extracted and sent for translation. The result would be the whole document in the target language that does not include different possible translations so the returned document is

Runs	MAP
LM-EToP-W	0.693
LM-PToE-W	0.849

Table 1: The comparison of LM-EToP-W and LM-PToE-W in terms of MAP.

used as the translated document. In the learning-to-rank method, the probabilistic translations are needed. The Moses translator is used for this purpose (Koehn et al., 2007). This translator is built over the Wikipedia corpus in Intelligent Information Systems Lab in University of Tehran. All possible translations with their probabilities are considered in our learning to rank method.

4.3 Experimental results

In this section, two different methods for cross-lingual subjectivity detection are evaluated through different experiments investigating their different characteristics. These methods are language-model based method and learning-to-rank method which are explained in section 3. In the rest of this section, experiments related to each of the methods are presented.

4.3.1 Language-model based method

In this section, experimental results of the language-model based method are represented. In this experiment, the English dataset is used to build the reference language models and the Persian dataset is used as the test data. Translation is applied on both directions, in other words on both English and Persian datasets and top translation is chosen for each word. Furthermore, translation is done on both document level and word level. In word level, based on the direction of translation, unigrams of the language models in the source language are translated to the target language and used to be compared with the other language models according to the methods formulas explained in section 3.1.1. Table 1 shows the results of this experiment which contains MAP values of the cross-lingual word level runs namely LM-EToP-W and LM-PToE-W which only differ in the direction of translation.

As shown in table 1, the Language-model based method using the word level translation performs better when the translation is from Persian to English. One of the reasons for this result is that in LM-EToP-W the unigrams in the reference language models (English dataset) are translated and

Runs	MAP
LM-EToP-D	0.800
LM-PToE-D	0.483

Table 2: The comparison of LM-EToP-D and LM-PToE-D in terms of MAP.

subsequently are subject to translation errors. As the methods performance basically depends on the quality of subjective and objective reference language models, the results would be reasonable.

In the next experiment, translation is applied in document level. The documents are translated to the target language in a preprocessing step and the rest of the experiment is similar to a monolingual problem. LM-EToP-D and LM-PToE-D are two runs executed in this section with document level translation from English to Persian and Persian to English respectively.

The results in Table 2 show that in document level translation, translation from English to Persian is more accurate than from Persian to English, contrary to word level translation. The reason is that the translation tool failed in translating most of the Persian documents, while in translating from English to Persian more documents are translated correctly by the machine translation tool so the MAP value of LM-EToP-D is higher than LM-PToE-D

4.3.2 Learning-to-Rank method

To obtain the results of our learning-to-rank method, a query list is needed. In this paper, the Persian query list is selected from the lexicon constructed in (Dehdarbehbahani et al., 2014). The lexicon contains 7491 terms and one tenth of it which has higher weights are selected as the Persian query list. The English query list which is selected from the Sentiwordnet (Baccianella et al., 2010) contains 156581 terms.

As explained in section 3.1.2, the queries are used via two approaches. In the first approach, the whole list is considered as one big query. In the second approach, each term in the query list is considered as an individual query and final subjectivity detection result is obtained by aggregating the results of each query weighted by the querys subjectivity weight. We employed SVM-rank (Joachims, 2006) and RankLib (Dang) to implement our learning-to-rank based method. In computations of the second feature set, BM25 feature has two parameters, including k and b, which

Runs	MAP
AD-EToP-LQ	0.675
AD-PToE-LQ	0.708
AD-EToP-TQ	0.929
AD-PToE-TQ	0.764

Table 3: The comparison of AD-EToP-LQ, AD-PToE-LQ, AD-EToP-TQ and AD-PToE-TQ in terms of MAP. are set to 1.5 and 0.8 respectively.

In the following experiments, the English dataset is used as the train set and the Persian dataset as the test set, MAP value is reported and the translation is applied in word level.

In the next experiment, ADARank algorithm is executed using the RankLib tool by four runs: 1) Using the whole query list of subjective words as one query while the translation direction is from English to Persian (AD-EToP-LQ). 2) Using the whole query list as one query while the translation direction is from Persian to English (AD-PToE-LQ). 3) Using each word of the query list as an individual query while the translation direction is from English to Persian (AD-EToP-TQ) 4) Using each word of the query list as an individual query while the translation direction is from Persian to English (AD-PToE-TQ). The results of this experiment are shown in table 3.

According to Table 3, using each word of the list as a separate query leads to better results than using one big query containing all subjective words as AD-EToP-TQ and AD-PToE-TQ outperform AD-EToP-LQ and AD-PToE-LQ respectively. Since words may have different translations and the selected translation may be incorrect, in AD-PToE-TQ and AD-EToP-TQ runs, the translation error only affects the single search corresponding to that query term but in AD-PToE-LQ and AD-EToP-LQ runs, the translation error of query terms would affect the whole query and it leads to lower results in the search corresponding to the list of query terms.

In the next experiment, we compare the results of the learning-to-rank based method with a baseline method. As a baseline method for ranking documents according to their subjectivity score, the language-model based method explained in section 3.1.1 can be a good choice since:

- It provides quantitative values as subjectivity scores of documents which facilitates ranking them similar to the output of the learning to rank approach.

88

Runs	MAP
LM-PToE-W	0.849
AD-EToP-TQ	0.929

Table 4: The comparison of AD-EToP-TQ and LM-PToE-W in terms of MAP.

Runs	MAP
RF-EToP-TQ-rev	0.811
AD-EToP-TQ-rev	0.809
CA-EToP-TQ-rev	0.860

Table 5: The comparison of RF-EToP-TQ-rev, AD-EToP-TQ-rev and CA-EToP-TQ-rev in terms of MAP.

- In previous papers, this method has been used for detecting positive documents from negative ones (Hu et al., 2007) and also for detecting subjective documents from objective ones (Karimi and Shakery, 2017).

To do the comparison, the best results of each method obtained in the experiments is selected. According to the results reported in previous tables, the best result of the language-model based method is achieved when translation is from Persian to English and translation units are words (namely LM-PToE-W). The best result of the learning-to-rank based method is obtained when each word of the list is used as an individual query while the translation direction is from English to Persian and the ADARank algorithm is employed (namely AD-EToP-TQ). These results are shown in table 4.

According to the results in table 4, in case translation tools are available, subjectivity detection using the learning-to-rank based method outperforms the language-model based method. The next experiment is designed to check if the results are biased to the dataset. Therefore, the Persian and English datasets are used interchangeably. In other words, in this experiment, the Persian dataset is used as the train set and the English dataset is used as the test set. Hence, the direction of translation is from Persian to English and term query level is used in this experiment. In this experiment, three learning to rank algorithms using Ranklib tool including Random Forests (RF-EToP-TQ-rev), ADARank (AD-EToP-TQ-rev) and Coordinate Ascent(CA-EToP-TQ-rev) are used and the MAP values are measured. Table 5 shows the results of these three runs.

As table 5 shows, Coordinate Ascent outperforms both other algorithms while Random Forests and ADARank performs similarly.

5 Conclusion

In this paper, we propose an extensive investigation on the cross-lingual subjectivity detection problem. Our main focus is to employ of English resources to rank Persian documents based on their subjectivity degree. In this study, two methods are employed as subjectivity detection systems. The first method is a language-model based method which computes the subjectivity score of each test document based on the similarity between the statistical language model of the test document and a reference subjective model and a reference objective model. The reference subjective and objective models are built using the labeled English data. Moreover, a cross-lingual subjectivity detection method is proposed which employs learning-to-rank techniques to rank documents according to their subjectivity score. In this method, the terms of a sentiment lexicon are used as query terms and the documents of the train data with subjective labels are considered as relevant documents to the query terms. Based on these definitions, the learning-to-rank framework is employed to rank test documents in resource-lean languages benefiting from resources including sentiment lexicon or labeled data in resource-rich languages. These two methods are evaluated using various translation directions and different translation units. Experimental results show how different parameters impact on the methods performance. Experiments also demonstrate that the proposed learning-to-rank based method outperforms the language-model based approach as a baseline method of ranking document according to their subjectivity degree. One of the future works for this research is studying the impact of translation on the performance of subjectivity detection in other resource lean languages.

References

Hosein Azarbonyad, Azadeh Shakery, and Heshaam Faili. 2012. Using learning to rank approach for par-

allel corpora based cross language information retrieval. In *proceedings of 20th European Conference on Artificial Intelligence (ECAI)*, pages 79–84.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. volume 10.

Alexandra Balahur and Marco Turchi. 2014. Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis. *Computer Speech & Language*, 28(1):56–75.

Carmen Banea, Rada Mihalcea, and Janyce Wiebe. 2010. Multilingual subjectivity: Are more languages better? In *Proceedings of the 23rd international conference on computational linguistics*, pages 28–36. Association for Computational Linguistics.

Carmen Banea, Rada Mihalcea, and Janyce Wiebe. 2014. Sense-level subjectivity in a multilingual setting. *Computer Speech & Language*, 28(1):7–19.

Carmen Banea, Rada Mihalcea, Janyce Wiebe, and Samer Hassan. 2008. Multilingual subjectivity analysis using machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 127–135. Association for Computational Linguistics.

Mikhail Bautin, Lohit Vijayarenu, and Steven Skiena. 2008. International sentiment analysis for news and blogs. In *ICWSM*.

Julian Brooke, Milan Tofiloski, and Maite Taboada. 2009. Cross-linguistic sentiment analysis: From english to spanish. In *RANLP*.

Van Dang. The lemur project-wiki-ranklib.

Iman Dehdarbehbahani, Azadeh Shakery, and Heshaam Faili. 2014. Semi-supervised word polarity identification in resource-lean languages. *Neural networks : the official journal of the International Neural Network Society*, 58:50–59.

Andrea Esuli and Fabrizio Sebastiani. 2009. Enhancing opinion extraction by automatically annotated lexical resources. In *Language and Technology Conference*, pages 500–511. Springer.

Yi Hu, Ruzhan Lu, Xuening Li, Yuquan Chen, and Jianyong Duan. 2007. A language modeling approach to sentiment analysis. In *Computational Science – ICCS 2007*, pages 1186–1193, Berlin, Heidelberg. Springer Berlin Heidelberg.

Thorsten Joachims. 2006. Training linear svms in linear time. volume 2006, pages 217–226.

Samaneh Karimi and Azadeh Shakery. 2017. A language-model-based approach for subjectivity detection. *Journal of Information Science*, 43(3):356–377.

Jungi Kim, Jinji Li, and Jong-Hyeok Lee. 2010. Evaluating multilanguage-comparability of subjectivity analysis systems. In *ACL*, page 595.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondej Bojar, Alex Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation.

Lun-Wei Ku, Yu-Ting Liang, and Hsin-Hsi Chen. 2006. Opinion extraction, summarization and tracking in news and blog corpora. In *Proceedings of AAAI*, pages 100–107.

Patrik Lambert. 2015. Aspect-level cross-lingual sentiment classification with constrained smt. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 781–787.

Josef Steinberger, Polina Lenkova, Mohamed Ebrahim, Maud Ehrmann, Ali Hurriyetoglu, Mijail Kabadjov, Ralf Steinberger, Hristo Tanev, Vanni Zavarella, and Silvia Vázquez. 2011a. Creating sentiment dictionaries via triangulation. In *Proceedings of the 2nd workshop on computational approaches to subjectivity and sentiment analysis*, pages 28–36. Association for Computational Linguistics.

Josef Steinberger, Polina Lenkova, Mijail A. Kabadjov, Ralf Steinberger, and Erik Van der Goot. 2011b. Multilingual entity-centered sentiment analysis evaluated by parallel corpora. In *RANLP*.

Xiaojun Wan. 2009. 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*, pages 235–243. Association for Computational Linguistics.

Bin Wei and Christopher Joseph Pal. 2010. Cross lingual adaptation: An experiment on sentiment classifications. In *ACL*, pages 258–262.