Comparative Experiments for Multilingual Sentiment Analysis Using Machine Translation

Alexandra Balahur and Marco Turchi
alexandra.balahur@jrc.ec.europa.eu
marco.turchi@jrc.ec.europa.eu

European Commission Joint Research Centre
IPSC, GlobeSec, OPTIMA
Via E. Fermi 2749, Ispra, Italy

Abstract. Sentiment analysis is the Natural Language Processing (NLP) task dealing with sentiment detection and classification from text. Given the importance of user-generated contents on the recent Social Web, this task has received much attention from the NLP research community in the past years. Sentiment analysis has been studied in different types of texts and in the context of distinct domains. However, only a small part of the research concentrated on dealing with sentiment analysis for languages other than English, which most of the times lack or have few lexical resources In this context, the present article proposes and evaluates the use of machine translation and supervised methods to deal with sentiment analysis in a multilingual context. Our extensive evaluation scenarios, for German, Spanish and French, using three different machine translation systems and various supervised algorithms show that SMT systems can start to be employed to obtain good quality data for other languages. Subsequently, this data can be employed to train classifiers for sentiment analysis in these languages, reaching performances close to the one obtained for English.

1 Introduction

During the past years, the contents that are generated by users on the Web, in the form of comments and statements of opinions in fora, blogs, reviewing sites, microblogs, have become more and more important. Their high volume and unbiased nature, as well as the fact that they are written by people from all social categories, all over the world, make such information useful to many domains, such as Economics, Social Science, Political Science, Marketing, to mention just a few. Nevertheless, the high quantity of such data and the high rate in which it is produced requires that automatic mechanisms are employed in order to extract valuable knowledge from it. In the case of opinionated data, this issue motivated the rapid and steady growth in interest from the Natural Language Processing (NLP) community to develop computational methods to analyze subjectivity and sentiment in text. These tasks received many names, from which "subjectivity analysis", "sentiment analysis" and "opinion mining" are the most frequently employed ones. The body of research conducted within these tasks has proposed different methods to deal with subjectivity and sentiment classification in different texts and domains, reaching satisfactory levels of performance for English. However, for certain

applications, such as news monitoring, the information in languages other than English is also highly relevant and cannot be disregarded, as it represents a high percentage of relevant data. In this type of systems, additionally, sentiment analysis tools must be reliable and perform at similar levels as the ones implemented for English.

In order to overcome the above-mentioned issue, the work presented herein aims to propose and evaluate different methods for multilingual sentiment analysis using machine translation and supervised methods. In particular, we will study this issue in three languages - French, German and Spanish - using three different Machine Translation systems - Google Translate, Bing Translator[1] and Moses [11] and different machine learning models. To have a more precise measure of the impact of quality translation on this task, we create Gold Standard sets for each of the three languages.

Our experiments show that machine translation systems are reaching a reasonable level of maturity so as to be employed for multilingual sentiment analysis and that for some languages (for which the translation quality is high enough) the performance that can be attained is similar to that of systems implemented for English, in terms of weighted F-measure.

2 Related Work

Most of the research in subjectivity and sentiment analysis was done for English. However, there were some authors who developed methods for the mapping of subjectivity lexicons to other languages. To this aim, [9] use a machine translation system and subsequently use a subjectivity analysis system that was developed for English to create subjectivity analysis resources in other languages. [12] propose a method to learn multilingual subjective language via cross-language projections. They use the Opinion Finder lexicon [22] and use two bilingual English-Romanian dictionaries to translate the words in the lexicon. Another approach was proposed by Banea et al. [3]. To this aim, the authors perform three different experiments - translating the annotations of the MPQA corpus, using the automatically translated entries in the Opinion Finder lexicon and the third, validating the data by reversing the direction of translation. In a further approach, Banea et al. [2] apply bootstrapping to build a subjectivity lexicon for Romanian, starting with a set of 60 words which they translate and subsequently filter using a measure of similarity to the original words, based on Latent Semantic Analysis (LSA) [8] scores. Yet another approach to mapping subjectivity lexica to other languages is proposed by Wan (2009), who uses co-training to classify un-annotated Chinese reviews using a corpus of annotated English reviews. [10] create a number of systems consisting of different subsystems, each classifying the subjectivity of texts in a different language. They translate a corpus annotated for subjectivity analysis (MPQA), the subjectivity clues (Opinion Finder) lexicon and re-train a Naive Bayes classifier that is implemented in the Opinion Finder system using the newly generated resources for all the languages considered. [4] translate the MPQA corpus into five other languages (some with a similar ethimology, others with a very different structure). Subsequently, they expand the feature space used in a Naive Bayes classifier using the same data translated to 2 or 3 other languages. Finally, [18, 19] create sentiment dictionaries in other

[1] http://translate.google.it/ and http://www.microsofttranslator.com/

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.

Attempts to use machine translation in different natural language processing tasks have not been widely used due to poor quality of translated texts, but recent advances in Machine Translation have motivated such attempts. In Information Retrieval, [17] proposed a comparison between Web searches using monolingual and translated queries. On average, the results show a drop in performance when translated queries are used, but it is quite limited, around 15%. For some language pairs, the average result obtained is around 10% lower than that of a monolingual search while for other pairs, the retrieval performance is clearly lower. In cross-language document summarization, [21, 5] combined the MT quality score with the informativeness score of each sentence in a set of documents to automatically produce summary in a target language using a source language texts. In [21], each sentence of the source document is ranked according both the scores, the summary is extracted and then the selected sentences translated to the target language. Differently, in [5], sentences are first translated, then ranked and selected. Both approaches enhance the readability of the generated summaries without degrading their content.

3 Motivation and Contribution

The work presented herein is mainly motivated by the need to develop sentiment analysis tools for a high number of languages, while minimizing the effort to create linguistic resources for each of these languages in part. Unlike approaches we presented in Related Work section, we employ fully-formed machine translation systems. In this context, another novelty in our approach is that we also study the influence of the difference in translation performance has on the sentiment classification performance.

Additionally, whereas the distinct characteristics of translated data (when compared to the original data) may imply that other features could be more appropriate. Moreover, such approaches have usually employed only simple machine learning algorithms. No attempt has been made to study the use of meta-classifiers to enhance the performance of the classification through the removal of noise in the data.

More specifically, we employ three MT systems - Bing Translator, Google Translate and Moses to translate data from English to three languages - French, German and Spanish. We create a Gold Standard for all the languages, used, on the one hand, to measure the translation quality and to test the performance of sentiment classification on translated (noisy) versus correct data. These correct translations allow us to have a more precise measure of the impact of translation quality on the sentiment classification task. Another contribution this article brings is the study of different types of features that can be employed to build machine learning models for the sentiment task. Further on, apart from studying different features that can be used to represent the training data, we also study the use of meta-classifiers to minimize the effect of noise in the data.

Our comparative results show, on the one hand, that machine translation can be reliably used for multilingual sentiment analysis and, on the other hand, which are the main characteristics of the data for such approaches to be successfully employed.

4 Dataset Presentation and Analysis

For our experiments, we employed the data provided for English in the NTCIR 8 Multilingual Opinion Analysis Task (MOAT)[2]. In this task, the organizers provided the participants with a set of 20 topics (questions) and a set of documents in which sentences relevant to these questions could be found, taken from the New York Times Text (2002-2005) corpus. The documents were given in two different forms, which had to be used correspondingly, depending on the task to which they participated. The first variant contained the documents split into sentences (6165 in total) and had to be used for the task of opinionatedness, relevance and answerness. In the second form, the sentences were also split into opinion units (6223 in total) for the opinion polarity and the opinion holder and target tasks. For each of the sentences, the participants had to provide judgements on the opinionatedness (whether they contained opinions), relevance (whether they are relevant to the topic). For the task of polarity classification, the participants had to employ the dataset containing the sentences that were also split into opinion units (i.e. one sentences could contain two/more opinions, on two/more different targets or from two/more different opinion holders).

For our experiments, we employed the latter representation. From this set, we randomly chose 600 opinion units, to serve as test set. The rest of opinion units will be employed as training set. Subsequently, we employed the Google Translate, Bing Translator and Moses systems to translate, on the one hand, the training set and on the other hand the test set, to French, German and Spanish. Additionally, we employed the Yahoo system (whose performance was the lowest in our initial experiments) to translate only the test set into these three languages. Further on, this translation has been corrected manually by a person, for all the languages. This corrected data serves as Gold Standard[3]. Most of these sentences, however, contained no opinion (were neutral). Due to the fact that the neutral examples are majoritary and can produce a large bias when classifying the polarity of the sentences, we eliminated these examples and employed only the positive and negative sentences in both the training, as well as the test sets. After this elimination, the training set contains 943 examples (333 positive and 610 negative) and the test set and Gold Standard contain 357 examples (107 positive and 250 negative). Although the upper bound for each of the systems would be possible to estimate using Gold Standard for each of the training sets, as well, at this point we considered the scenario that is closer to real situations, in which the issue is related to the inexistence of training data for a specific language.

5 Using Machine Translation for Multilingual Sentiment Analysis

The issue of extracting and classifying sentiment in text has been approached using different methods, depending on the type of text, the domain and the language considered. Broadly speaking, the methods employed can be classified into unsupervised

[2] http://research.nii.ac.jp/ntcir/ntcir-ws8/permission/ntcir8xinhua-nyt-moat.html

[3] We translated the whole sentences, not opinion units separately, so sentences containing multiple opinion units were translated twice. After duplicate elimination, we remained with 400 sentences in the test and Gold Standard sets and 5700 sentences in the training set.

(knowledge-based), supervised and semi-supervised methods. The first usually employ lexica or dictionaries of words with associated polarities (and values - e.g. 1, -1) and a set of rules to compute the final result. The second category of approaches employ statistical methods to learn classification models from training data, based on which the test data is then classified. Finally, semi-supervised methods employ knowledge-based approaches to classify an initial set of examples, after which they use different machine learning methods to bootstrap new training examples, which they subsequently use with supervised methods.

The main issue with the first approach is that obtaining large-enough lexica to deal with the variability of language is very expensive (if it is done manually) and generally not reliable (if it is done automatically). Additionally, the main problem of such approaches is that words outside contexts are highly ambiguous. Semi-supervised approaches, on the other hand, highly depend on the performance of the initial set of examples that is classified. If we are to employ machine translation, the errors in translating this small initial set would have a high negative impact on the subsequently learned examples. The challenge of using statistical methods is that they require training data (e.g. annotated corpora) and that this data must be reliable (i.e. not contain mistakes or "noise"). The lower the performance in classifying, the more sparse will be the feature vectors employed in the machine learning models. However, the larger this dataset is, the less influence the translation errors have.

Since we want to study whether machine translation can be employed to perform sentiment analysis for different languages, we employed statistical methods in our experiments. More specifically, we used Support Vector Machines Sequential Minimal Optimization (SVM SMO), with different types of features (n-grams, presence of sentiment words), since the literature in the field has confirmed it as the best-performing machine learning algorithm for this task [16].

For the purpose of our experiments, three different SMT systems were used to translate the human annotated sentences: two existing online services such as *Google Translate* and *Bing Translator*[4] and an instance of the open source phrase-based statistical machine translation toolkit Moses [11], trained on freely available corpora.This results in 2.7 million sentence pairs for English-French, 3.8 for German and 4.1 for Spanish. All the modes are optimized running the MERT algorithm [13] on the development part of the training data. The translated sentences are recased and detokonized (for more details on the system, please see [20].

6 Experiments

In order to test the performance of sentiment classification when using translated data, we employed supervised learning using Support Vector Machines Sequential Minimal Optimization [14] - SVM SMO - with different features:

– In the first approach, we represented, for each of the languages and translation systems, the sentences as vectors, whose features marked the presence/absence

[4] http://translate.google.com/ and http://www.microsofttranslator.com/

(boolean) of the unigrams contained in the corresponding training set (e.g. we obtained the unigrams in all the sentences in the training set obtained by translating the English training data to Spanish using Google and subsequently represented each sentence in this training set, as well as the test set obtained by translating the test data in English to Spanish using Google marking the presence of the unigram features).
- In the second approach, we represented the training and test sets as in the previous representation, with the difference that the features were computed not as the presence of the unigrams, but the tf-idf score of that unigram.
- In the third approach, we represented, for each of the languages and translation systems, the sentences as vectors, whose features marked the presence/absence of the unigrams and bigrams contained in the corresponding training set.

In our experiments, we also studied the possibility to employ sentiment-bearing words in the sentences to be classified as features for the machine learning algorithm. In order to do this, we employed the SentiWordNet, General Inquirer and WordNet Affect dictionaries for English and the multilingual dictionaries created by (Steinberger et al., 2012). The main problem of this approach was, however, that very few features were found, for a small number of the sentences to be classified, on the one hand because affect is not expressed in these sentences using lexical clues and, on the other hand, because the dictionaries we had at our disposal for languages other than English were not very large (around 1500 words). For this reason, we will not report these results.

Table 1 presents the number of unigram and bigram features employed in each of the cases.

Language	SMT system	Nr. of unigrams	Nr. of bigrams
	—	5498	15981
English French	Bing	7441	17870
	Google	7540	18448
	Moses	6938	18814
	Bing+Google+Moses	9082	40977
German	Bing	7817	16216
	Google	7900	16078
	Moses	7429	16078
	Bing+Google+Moses	9371	36556
Spanish	Bing	7388	17579
	Google	7803	18895
	Moses	7528	18354
	Bing+Google+Moses	8993	39034

Table 1. Features employed for representing the sentences in the training and test sets.

Subsequently, we performed two sets of experiments:

– In the first set of experiments, we trained an SVM SMO classifier on the training data obtained for each language, with each of the three machine translations, separately (i.e. we generated a model for each of the languages considered, for each of the machine translation systems employed), using the three types of aforementioned features. Subsequently, we tested the models thus obtained on the corresponding test set (e.g. training on the Spanish training set obtained using Google Translate and testing on the Spanish test set obtained using Google Translate) and on the Gold Standard for the corresponding language (e.g. training on the Spanish training set obtained using Google Translate and testing on the Spanish Gold Standard). Additionally, in order to study the manner in which the noise in the training data can be removed, we employed one meta-classifier - Bagging [6] (with varying sizes of the bag and SMO as classifier). In related experiments, we also employed other meta-classifiers, such as AdaBoost[1]), but the best results were obtained using Bagging.
– In the second set of experiments, we combined the translated data from all three machine translation systems for the same language and created separate models based on the three types of features we extracted from this data (e.g. we created a Spanish training model using the unigrams and bigrams present in the training sets generated by the translation of the training set to Spanish by Google Translate, Bing Translator and Moses). We subsequently tested the performance of the sentiment classification using the Gold Standard for the corresponding language, represented using the corresponding set of features of this model.

The results of the experiments (in terms of weighted F-score, per language) are presented in Tables 2, 3, 4 and 5, and for the second set of experiments are presented in Table 6.

Feature Representation	Test Set	SMO	Bagging
Unigram	GS	0.683	**0.687**
Unigram tf-idf	GS	0.651	**0.681**
Unigram+Bigram	GS	0.685	**0.686**

Table 2. Results obtained for English using the different representations.

7 Results and Discussion

Generally speaking, from our experiments using SVM, we could see that incorrect translations imply an increment of the features, sparseness and more difficulties in identifying a hyperplane which separates the positive and negative examples in the training phase. Therefore, a low quality of the translation leads to a drop in performance, as the features extracted are not informative enough to allow for the classifier to learn. For German, an agglutinative language, wrong translation also leads to an explosion of features, of which many are irrelevant for the learning process.

Feature Representation	SMT	Test Set	SMO	AdaBoost M1	Bagging	BLEU Score
Unigram	Bing	GS	0.655	0.62	**0.658**	
		Tr	0.655	0.625	**0.666**	0.227
Unigram	Google T.	GS	0.64	0.622	**0.655**	
		Tr	**0.695**	0.645	0.693	0.209
Unigram	Moses	GS	0.649	0.641	**0.675**	
		Tr	**0.666**	0.654	0.661	0.17
Unigram tf-idf	Bing	GS	0.627	0.628	**0.64**	
		Tr	0.654	0.625	**0.673**	0.227
Unigram tf-idf	Google T.	GS	0.626	0.598	**0.643**	
		Tr	0.667	0.627	**0.693**	0.209
Unigram tf-idf	Moses	GS	0.654	0.646	**0.659**	
		Tr	0.664	0.66	**0.673**	0.17
Unigram+Bigram	Bing	GS	0.641	0.631	**0.648**	
		Tr	0.658	0.636	**0.662**	0.227
Unigram+Bigram	Google T.	GS	0.646	0.623	**0.674**	
		Tr	**0.687**	0.645	0.661	0.209
Unigram+Bigram	Moses	GS	0.644	0.644	**0.676**	
		Tr	0.667	0.667	**0.674**	0.17

Table 3. Results obtained for German using the different feature representations.

From Tables 2,3, 4 and 5, we can see that there is a small difference between performances of the sentiment analysis system using the English and translated data, respectively. In the worst case, there is a maximum drop of 12 percentages using SMO and 8 percentages using Bagging. Ideally, to better measure this drop we would have had to use gold standard training data for each language. As mentioned in Section 4, the creation of the gold standard is a very difficult and time consuming task. We are considering the manual translation of the training data into French, German and Spanish for the future work. Nonetheless, the scenario considered was aimed at studying the use of MT for SA in the real-life scenario, in which there is no annotated data for the language on which SA is done.

The noise in the data appears from two sources - namely the incorrect translations or the features that are not appropriate. Manual inspection of the results has shown that in case of German, the tf-idf obtains the best results because it removes irrelevant features (words that are mentioned very few times). On the other hand, for languages for which the translation quality is higher - i.e. Spanish and French in our case - we obtained better results when using a combination of unigrams and bigrams. After manually inspecting the data, we noticed that cleaner are the data the most useful is the unigram and bigram representation, as this representation increases the quantity of useful features for training. This is not the case for German, where this representation increases to a higher degree the noise (the number of noisy features).

In the line of the previous consideration, Bagging, by reducing the variance in the estimated models, produces a positive effect on the performance increasing the F-score, as compared to the learning process and features without Bagging. These improve-

Feature Representation	SMT	Test Set	SMO	AdaBoost M1	Bagging	BLEU Score
Unigram	Bing	GS	0.627	0.62	**0.633**	
		Tr	**0.634**	0.629	0.618	0.316
Unigram	Google T.	GS	0.635	0.635	**0.659**	
		Tr	0.63	0.63	**0.665**	0.341
Unigram	Moses	GS	**0.644**	0.644	0.639	
		Tr	0.675	0.675	**0.676**	0.298
Unigram tf-idf	Bing	GS	**0.659**	0.649	0.655	
		Tr	0.622	0.637	**0.646**	0.316
Unigram tf-idf	Google T.	GS	0.652	0.652	**0.673**	
		Tr	0.624	0.624	**0.637**	0.341
Unigram tf-idf	Moses	GS	0.646	0.646	**0.66**	
		Tr	**0.677**	0.677	0.676	0.298
Unigram+Bigram	Bing	GS	**0.656**	0.658	0.646	
		Tr	0.633	0.633	**0.633**	0.316
Unigram+Bigram	Google T.	GS	0.653	0.653	**0.665**	
		Tr	0.636	0.667	**0.665**	0.341
Unigram+Bigram	Moses	GS	0.664	0.664	**0.671**	
		Tr	0.649	0.649	**0.663**	0.298

Table 4. Results obtained for Spanish using the different feature representations.

ments are larger using the German data, because the poor quality of the its translations increases the variance in the data. For the same reason, Bagging is quite effective when unigrams and bigrams are used to represent low quality translated data. In this work we pair Bagging with SMO, but we are interested in running experiments using weak classifiers such as Naive Bayes or neural networks.

Finally, as expected, the performance of the classification is much higher for data obtained using the same translator than on the Gold Standard. This is true, as the same incorrect translations are repeated in both sets and therefore the learning is not influenced by these mistakes.

Looking at the results in Table 6, we can see that adding all the translated training data together makes the features in the representation more sparse and increases the noise level in the training data, creating harmful effects in terms of classification performance: each classifier loses its discriminative capability. This is not the case when using tf-idf on unigrams, in which case the combination of the data improves the classification, as this type of features deter sparsity in data.

At language level, clearly the results depend on the translation performance. Only for Spanish (for which we have the highest Bleu score), each classifies is able to properly learn from the training data and try to properly assign the test samples. For the other languages, translated data are so noisy that or the classifier is not able to properly learn the correct information for the positive and the negative classes, and this results in the assignment of most of the test points to one class and zero to the other, or there is significant drop in performance, e.g. for the French language, but the classifier is still able to assign the test points to both the classes.

Feature Representation	SMT	Test Set	SMO	AdaBoost M1	Bagging	Bleu Score
Unigram	Bing	GS	0.604	0.634	**0.644**	
		Tr	0.649	0.654	**0.657**	0.243
Unigram	Google T.	GS	0.628	0.628	**0.638**	
		Tr	0.652	0.652	**0.679**	0.274
Unigram	Moses	GS	0.646	**0.666**	0.642	
		Tr	**0.663**	0.657	0.66	0.227
Unigram tf-idf	Bing	GS	**0.646**	0.641	0.645	
		Tr	0.652	0.661	**0.664**	0.243
Unigram tf-idf	Google T.	GS	0.635	0.635	**0.645**	
		Tr	0.672	0.672	**0.68**	0.274
Unigram tf-idf	Moses	GS	**0.656**	0.635	0.653	
		Tr	**0.686**	0.646	0.671	0.227
Unigram+Bigram	Bing	GS	0.644	0.645	**0.664**	
		Tr	0.644	0.649	**0.652**	0.243
Unigram+Bigram	Google T.	GS	0.64	0.64	**0.659**	
		Tr	0.652	0.652	**0.678**	0.274
Unigram+Bigram	Moses	GS	0.633	0.633	**0.645**	
		Tr	0.666	0.666	**0.674**	0.227

Table 5. Results obtained for French using the different feature representations.

The results confirm the capability of Bagging to reduce the model variance and increase the performance in classification, in particular for the ungrams plus tfidf representation or for the Spanish language. In both the cases, performances are really close (for some configurations even better) to what we obtained using each dataset independently.

8 Conclusions and Future Work

The main objective of this work was to study the manner in which sentiment analysis can be done for languages other than English by employing MT systems and supervised learning. Overall, we could see that MT systems have reached a reasonable level of maturity to produce sufficiently reliable training data for languages other than English. Additionally, for some languages, the quality of the translated data is high enough to obtain performances similar to that for the original data using supervised learning without any subsequent meta-classification for noise reduction. Finally, even in the worst cases, when the quality of the translated data is not very high, the drop in performance is of maximum 12% and it can be improved on using meta-classifiers. From the different feature representations, we could see that wrong translations lead to a large number of features, sparseness and noise in the data points in the classification task. This is especially visible in the boolean representation, which is also more sensitive to noise. Through the different types of features and classifiers, we used showing that using unigrams or tf-idf on unigrams as features, and/or Bagging as a meta-classifier, has a

	Unigrams			Unigrams + tfidf			Unigrams+Bigrams		
Language	SMO	AdaBoost M1	Bagging	SMO	AdaBoost M1	Bagging	SMO	AdaBoost M1	Bagging
To German	0.565*	0.563	0.563*	0.658	0.64	**0.665**	0.565*	0.563*	0.565*
To Spanish	0.587	0.599	0.598	0.657	0.646	**0.666**	0.419	0.494	0.511
To French	0.609	0.575	0.578	0.626	0.634	**0.635**	0.25	0.255	0.23

Table 6. For each language, each classifier has been trained merging the translated data coming form different SMT systems, and tested using the Gold Standard. *Classifier is not able to discriminate between positive and negative classes, and assigns most of the test points to one class, and zero to the other.

positive impact in the results. Furthermore, in case of good translation quality, we noticed that the union of the same training data translated with various systems can help the classifiers to learn different linguistic aspects from the same data.

In future work, we plan to further study methods to improve the classification performance, both by enriching the features employed, as well as extending the use of meta-classifiers to enhance noise reduction. In particular, the first step will be to adding specialized features corresponding to words belonging to sentiment lexica (in conjunction to the types of features we have already employed) and include high level syntax information can reduce the impact of the translation errors. Finally, we plan to employ confidence estimation mechanisms to filter the best translations, which can subsequently be employed more reliably for system training.

Acknowledgements

The authors would like to thank Ivano Azzini, from the BriLeMa Artificial Intelligence Studies, for the advice and support on using meta-classifiers. We would also like to thank the reviewers for their useful comments and suggestions on the paper.

References

1. Balahur, A. and Turchi, M. 2012. Multilingual Sentiment Analysis using Machine Translation?. Proceedings of the 3rd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis Workshop, 52 Jeju, Republic of Korea.
2. Banea, C., Mihalcea, R., and Wiebe, J. 2008. A bootstrapping method for building subjectivity lexicons for languages with scarce resources. Proceedings of the Conference on Language Resources and Evaluations (LREC 2008), Maraakesh, Marocco.
3. Banea, C., Mihalcea, R., Wiebe, J., and Hassan, S. 2008. Multilingual subjectivity analysis using machine translation. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2008), 127-135, Honolulu, Hawaii.
4. Banea, C., Mihalcea, R. and Wiebe, J. 2010. Multilingual subjectivity: are more languages better?. Proceedings of the International Conference on Computational Linguistics (COLING 2010), p. 28-36, Beijing, China.
5. Boudin, F. and Huet, S. and Torres-Moreno, J.M. and Torres-Moreno, J.M. 2010. A Graph-based Approach to Cross-language Multi-document Summarization. Research journal on Computer science and computer engineering with applications (Polibits), 43:113–118.

6. Breiman, L 1996. Bagging predictors. Machine learning, 24(2):123–140.
7. P. F. Brown, S. Della Pietra, V. J. Della Pietra and R. L. Mercer. 1994. The Mathematics of Statistical Machine Translation: Parameter Estimation. Computational Linguistics 19:263–311.
8. Deerwester, S., Dumais, S., Furnas, G. W., Landauer, T. K., and Harshman, R. 1990. Indexing by latent semantic analysis. Journal of the American Society for Information Science, 3(41).
9. Kim, S.-M. and Hovy, E. 2006. Automatic identification of pro and con reasons in online reviews. Proceedings of the COLING/ACL Main Conference Poster Sessions, pages 483.
10. Kim, J., Li, J.-J. and Lee, J.-H. 2006. Evaluating Multilanguage-Comparability of Subjectivity Analysis Systems. Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pages 595 Uppsala, Sweden, 11-16 July 2010.
11. P. Koehn and H. Hoang and A. Birch and C. Callison-Burch and M. Federico and N. Bertoldi and B. Cowan and W. Shen and C. Moran and R. Zens and C. Dyer and O. Bojar and A. Constantin and E. Herbst 2007. Moses: Open source toolkit for statistical machine translation. Proceedings of the Annual Meeting of the Association for Computational Linguistics, demonstration session, pages 177–180. Columbus, Oh, USA.
12. Mihalcea, R., Banea, C., and Wiebe, J. 2009. Learning multilingual subjective language via cross-lingual projections. Proceedings of the Conference of the Annual Meeting of the Association for Computational Linguistics 2007, pp.976-983, Prague, Czech Republic.
13. F. J. Och 2003. Minimum error rate training in statistical machine translation. Proceedings of the 41st Annual Meeting on Association for Computational Linguistics, pages 160–167. Sapporo, Japan.
14. Platt, J. C. 1999. Fast training of support vector machines using sequential minimal optimization. Advances in kernel methods, isbn 0-262-19416-3, pages 185–208.
15. K. Papineni and S. Roukos and T. Ward and W. J. Zhu 2001. BLEU: a method for automatic evaluation of machine translation. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, pages 311–318. Philadelphia, Pennsylvania.
16. Pang, B. and Lee, L. 2008. Opinion Mining and Sentiment Analysis. Found. Trends Inf. Retr., vol. 1, nr. 1–2, 2008.
17. J. Savoy, and L. Dolamic. 2009. How effective is Google's translation service in search?. Communications of the ACM, 52(10):139–143.
18. Steinberger, J. and Lenkova, P. and Ebrahim, M. and Ehrman, M. and Hurriyetoglu, A. and Kabadjov, M. and Steinberger, R. and Tanev, H. and Zavarella, V. and Vazquez, S. 2011. Creating Sentiment Dictionaries via Triangulation. Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis, Portland, Oregon.
19. Steinberger, J. and Lenkova, P. and Kabadjov, M. and Steinberger, R. and van der Goot, E. 2011. Multilingual Entity-Centered Sentiment Analysis Evaluated by Parallel Corpora. Proceedings of the Conference on Recent Advancements in Natural Language Processing (RANLP), Hissar, Bulgaria.
20. Turchi, M. and Atkinson, M. and Wilcox, A. and Crawley, B. and Bucci, S. and Steinberger, R. and Van der Goot, E. 2012. ONTS:"Optima" News Translation System. Proceedings of EACL 2012, pages 25–31.
21. Wan, X. and Li, H. and Xiao, J. 2010. Cross-language document summarization based on machine translation quality prediction. Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pages 917–926.
22. Wilson, T., Wiebe, J., and Hoffmann, P. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. Proceedings of HLT-EMNLP 2005, pp.347-354, Vancouver, Canada.