Multilingual Sentiment Analysis using Machine Translation?

Alexandra Balahur and Marco Turchi
European Commission Joint Research Centre
Institute for the Protection and Security of the Citizen
Via E. Fermi 2749, Ispra, Italy
alexandra.balahur, marco.turchi@jrc.ec.europa.eu

Abstract

The past years have shown a steady growth in interest in the Natural Language Processing task of sentiment analysis. The research community in this field has actively proposed and improved methods to detect and classify the opinions and sentiments expressed in different types of text - from traditional press articles, to blogs, reviews, fora or tweets. A less explored aspect has remained, however, the issue of dealing with sentiment expressed in texts in languages other than English. To this aim, the present article deals with the problem of sentiment detection in three different languages - French, German and Spanish - using three distinct Machine Translation (MT) systems - Bing, Google and Moses. Our extensive evaluation scenarios show that SMT systems are mature enough to be reliably employed to obtain training data for languages other than English and that sentiment analysis systems can obtain comparable performances to the one obtained for English.

1 Introduction

Together with the increase in the access to technology and the Internet, the past years have shown a steady growth of the volume of user-generated contents on the Web. The diversity of topics covered by this data (mostly containing subjective and opinionated content) in the new textual types such as blogs, fora, microblogs, has been proven to be of tremendous value to a whole range of applications, in Economics, Social Science, Political Science, Marketing, to mention just a few. Notwithstanding these proven advantages, the high quantity of user-generated contents makes this information hard to access and employ without the use of automatic mechanisms. 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. Different methods have been proposed to deal with these phenomena for the distinct types of text and domains, reaching satisfactory levels of performance for English. Nevertheless, for certain applications, such as news monitoring, the information in languages other than English is also highly relevant and cannot be disregarded. Additionally, systems dealing with sentiment analysis in the context of monitoring must be reliable and perform at similar levels as the ones implemented for English.

Although the most obvious solution to these issues of multilingual sentiment analysis would be to use machine translation systems, researchers in sentiment analysis have been reluctant to using such technologies due to the low performance they used to have. However, in the past years, the performance of Machine Translation systems has steadily improved. Open access solutions (e.g. Google Translate[1], Bing Translator[2]) offer more and more accurate translations for frequently used languages.

Bearing these thoughts in mind, in this article we study the manner in which sentiment analysis can be done for languages other than English, using Machine Translation. In particular, we will study

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

[2] http://www.microsofttranslator.com/

Proceedings of the 3rd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis, pages 52–60,
Jeju, Republic of Korea, 12 July 2012.

this issue in three languages - French, German and Spanish - using three different Machine Translation systems - Google Translate, Bing Translator and Moses (Koehn et al., 2007).

We employ these systems to obtain training and test data for these three languages and subsequently extract features that we employ to build machine learning models using Support Vector Machines Sequential Minimal Optimization. We additionally employ meta-classifiers to test the possibility to minimize the impact of noise (incorrect translations) in the obtained data.

Our experiments show that machine translation systems are mature enough 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.

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, (Kim and Hovy, 2006) 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. (Mihalcea et al., 2009) propose a method to learn multilingual subjective language via cross-language projections. They use the Opinion Finder lexicon (Wilson et al., 2005) 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. (Banea et al., 2008b). 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. Further on, another approach to building lexicons for languages with scarce resources is presented by Banea et al. (Banea et al., 2008a). In this research, the authors apply bootstrapping to build a subjectivity lexicon for Romanian, starting with a set of seed subjective entries, using electronic bilingual dictionaries and a training set of words. They start with a set of 60 words pertaining to the categories of noun, verb, adjective and adverb from the translations of words in the Opinion Finder lexicon. Translations are filtered using a measure of similarity to the original words, based on Latent Semantic Analysis (LSA) (Deerwester et al., 1990) 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. He first translates the English reviews into Chinese and subsequently back to English. He then performs co-training using all generated corpora. (Kim et al., 2010) 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 retrain a Nave Bayes classifier that is implemented in the Opinion Finder system using the newly generated resources for all the languages considered. Finally, (Banea et al., 2010) 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 Nave Bayes classifier using the same data translated to 2 or 3 other languages. Their conclusion is that by expanding the feature space with data from other languages performs almost as well as training a classifier for just one language on a large set of training data.

Attempts of using 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, (Savoy and Dolamic, 2009) 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, (Wan et al., 2010; Boudin et al., 2010) 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 (Wan et al., 2010), 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 (Boudin et al., 2010), 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 main motivation for the experiments we present in this article is the known lack of resources and approaches for sentiment analysos in languages other than English. Although, as we have seen in the Related Work section, a few attempts were made to build systems that deal with sentiment analysis in other languages, they mostly employed bilingual dictionaries and used unsupervised approaches. The very few that employed supervised learning using translated data have, in change, concentrated only on the issue of sentiment classification and have disregarded the impact of the translation quality and the difference that the use of distinct translation systems can make in this settings. 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.

Our main contribution in this article is the comparative study of multilingual sentiment analysis performance using distinct machine translation systems, with varying levels of translation quality. In this sense, we employ three different systems - Bing Translator, Google Translate and Moses to translate data from English to three languages - French, German and Spanish. We subsequently study the performance of classifying sentiment from the translated data and different methods 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)[3]. 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 judgments 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

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

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 to translate only the test set into these three languages. Further on, this translation of the test set by the Yahoo service has been corrected by a person for all the languages. This corrected data serves as Gold Standard[4]. 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, we decided to eliminate these examples and employ 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).

5 Machine Translation

During the 1990's the research community on Machine Translation proposed a new approach that made use of statistical tools based on a noisy channel model originally developed for speech recognition (Brown et al., 1994). In the simplest form, Statistical Machine Translation (SMT) can be formulated as follows. Given a source sentence written in a foreign language f, the Bayes rule is applied to reformulate the probability of translating f into a sentence e written in a target language:

$$e_{best} = arg\max_{e} p(e|f) = arg\max_{e} p(f|e)p_{LM}(e)$$

where $p(f|e)$ is the probability of translating e to f and $p_{LM}(e)$ is the probability of producing a fluent sentence e. For a full description of the model see (Koehn, 2010).

The noisy channel model was extended in different directions. In this work, we analyse the most popular class of SMT systems: PBSMT. It is an extension of the noisy channel model using phrases rather than words. A source sentence f is segmented into a sequence of I phrases $f^I = \{f_1, f_2, \dots f_I\}$ and the same is done for the target sentence e, where the notion of phrase is not related to any grammatical assumption; a phrase is an n-gram. The best translation e_{best} of f is obtained by:

$$\begin{aligned} e_{best} &= arg\max_{e} p(e|f) = arg\max_{e} p(f|e)p_{LM}(e) \\ &= arg\max_{e} \prod_{i=1}^{I} \phi(f_i|e_i)^{\lambda_\phi} d(a_i - b_{i-1})^{\lambda_d} \\ &\quad \prod_{i=1}^{|e|} p_{LM}(e_i|e_1 \dots e_{i-1})^{\lambda_{LM}} \end{aligned}$$

where $\phi(f_i|e_i)$ is the probability of translating a phrase e_i into a phrase f_i. $d(a_i - b_{i-1})$ is the distance-based reordering model that drives the system to penalise significant reorderings of words during translation, while allowing some flexibility. In the reordering model, a_i denotes the start position of the source phrase that is translated into the ith target phrase, and b_{i-1} denotes the end position of the source phrase translated into the $(i-1)$th target phrase. $p_{LM}(e_i|e_1 \dots e_{i-1})$ is the language model probability that is based on the Markov's chain assumption. It assigns a higher probability to fluent/grammatical sentences. λ_ϕ, λ_{LM} and λ_d are used to give a different weight to each element. For more details see (Koehn et al., 2003).

Three different SMT systems were used to translate the human annotated sentences: two existing online services such as *Google Translate* and *Bing Translator*[5] and an instance of the open source phrase-based statistical machine translation toolkit Moses (Koehn et al., 2007).

To train our models based on Moses we used the freely available corpora: Europarl (Koehn, 2005), JRC-Acquis (Steinberger et al., 2006), Opus (Tiedemann, 2009), News Corpus (Callison-Burch et al., 2009). 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 (Och, 2003) on the development part of the News Corpus. The translated sentences are recased and detokonized (for more details on the system, please see (Turchi et al., 2012).

[4]Please note that each sentence may contain more than one opinion unit. In order to ensure a contextual translation, we translated the whole sentences, not the opinion units separately. In the end, we eliminate duplicates of sentences (due to the fact that they contained multiple opinion units), resulting in around 400 sentences in the test and Gold Standard sets and 5700 sentences in the training set

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

Performances of a SMT system are automatically evaluated comparing the output of the system against human produced translations. Bleu score (Papineni et al., 2001) is the most used metric and it is based on averaging n-gram precisions, combined with a length penalty which penalizes short translations containing only sure words. It ranges between 0 and 1, and larger value identifies better translation.

6 Sentiment Analysis

In the field of sentiment analysis, most work has concentrated on creating and evaluating methods, tools and resources to discover whether a specific "target"or "object" (person, product, organization, event, etc.) is "regarded" in a positive or negative manner by a specific "holder" or "source" (i.e. a person, an organization, a community, people in general, etc.). This task has been given many names, from opinion mining, to sentiment analysis, review mining, attitude analysis, appraisal extraction and many others.

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 (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"). 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) since the literature in the field has confirmed it as the most appropriate machine learning algorithm for this task.

In the case of statistical methods, the most important aspect to take into consideration is the manner in which texts are represented - i.e. the features that are extracted from it. For our experiments, we represented the sentences based on the unigrams and the bigrams that were found in the training data. Although there is an ongoing debate on whether bigrams are useful in the context of sentiment classification, we considered that the quality of the translation can also be best quantified in the process by using these features (because they give us a measure of the translation correctness, both regarding words, as well as word order). Higher level n-grams, on the other hand, would only produce more sparse feature vectors, due to the high language variability and the mistakes in the traslation.

7 Experiments

In order to test the performance of sentiment classification when using translated data, we performed a series 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). Subsequently, we tested the models thus obtained on the corresponding test set (e.g. training on the Spanish train-

ing 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 two meta-classifiers - AdaBoost and Bagging (with varying sizes of the bag).

- In the second set of experiments, we combined the translated data from all three machine translation systems for the same language and created a model based on the unigram and bigram features 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 features of this model.

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

In the following subsections, we present the results of these experiments.

7.1 Individual Training with Translated Data

In the first experiment, we translated the training and test data from English to all the three other languages considered, using each of the three machine translation systems. Subsequently, we represented, for each of the languages and translation systems, the sentences as vectors, whose features marked the presence/absence (1 or 0) of the unigrams and bigrams contained in the corresponding trainig set (e.g. we obtained the unigrams and bigrams 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 and bigram features). In order to test the approach on the Gold Standard (for each language), we represented this set using the corresponding unigram and bigram features extracted from the corresponding training set (for the example given, we represented each sentence in the Gold Standard by marking the presence/absence of the unigrams and bigrams from the training data for Spanish using Google Translate).

The results of these experiments are presented in Table 2, in terms of weighted F1 measure.

7.2 Joint Training with Translated Data

In the second set of experiments, we added together all the translations of the training data obtained for the same language, with the three different MT systems. Subsequently, we represented, for each language in part, each of the sentences in the joint training corpus as vectors, whose features represented the presence/absence of the unigrams and bigrams contained in this corpus. In order to test the performance of the sentiment classification, we employed the Gold Standard for the corresponding language, representing each sentence it contains according to the presence or absence of the unigrams and bigrams in the corresponding joint training corpus for that language. Finally, we applied SVM SMO to classify the sentences according to the polarity of the sentiment they contained. Additionally, we applied the AdaBoost and Bagging meta-classifiers to test the possibilities to minimize the impact of noise in the data. The results are presented in Tables 3 and 4, again, in terms of weighter F1 measure.

Language	SMO	AdaBoost M1	Bagging
To German	0.565*	0.563*	0.565*
To Spanish	0.419	0.494	0.511
To French	0.25	0.255	0.23

Table 3: 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.

8 Results and Discussion

Generally speaking, from our experiments using SVM, we could see that incorrect translations imply

	Bing	Google T.	Moses
To German	0.57*	0.572*	0.562*
To Spanish	0.392	0.511	0.448
To French	0.612*	0.571*	0.575*

Table 4: For each language, the SMO classifiers have been trained merging the translated data coming form different SMT systems, and tested using independently the translated test sets. *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.

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.

From Table 2, we can see that:
a) 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 8 percentages.
b) Adaboost is sensitive to noisy data, and it is evident in our experiments where in general it does not modify the SMO performances or there is a drop. Vice versa, Bagging, reducing the variance in the estimated models, produces a positive effect on the performances increasing the F-score. These improvements are larger using the German data, this is due to the poor quality of the translated data, which increases the variance in the data.

Looking at the results in Tables 3 and 4, we can see that:
a) Adding all the translated training data together drastically increases the noise level in the training data, creating harmful effects in terms of classification performance: each classifier loses its discriminative capability.
b) 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 the classifier is not able to properly learn the correct information for the positive and the negative classes, this results in the assignment of most of the test points to one class and zero to the other. In Table 3, for the French language we have significant drop in performance, but the classifier is still able to learn something from the training and assign the test points to both the classes.
c) The results for Spanish presented in Table 3 confirm the capability of Bagging to reduce the model variance and increase the performance in classification.
d) At system level in Table 4, there is no evidence that better translated test set allows better classification performance.

9 Conclusions and Future Work

In this work we propose an extensive evaluation of the use of translated data in the context of sentiment analysis. Our findings show that SMT systems are mature enough to produce reliably training data for languages other than English. The gap in classification performance between systems trained on English and translated data is minimal, with a maximum of 8

Working with translated data implies an increment number of features, sparseness and noise in the data points in the classification task. To limit these problems, we test three different classification approaches showing that bagging has a positive impact in the results.

In future work, we plan to investigate different document representations, in particular we believe that the projection of our documents in space where the features belong to a sentiment lexical and include syntax information can reduce the impact of the translation errors. As well we are interested to evaluate different term weights such as tf-idf.

Acknowledgments

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

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.

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.

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.

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.

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.

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.

C. Callison-Burch, and P. Koehn and C. Monz and J. Schroeder. 2009. *Findings of the 2009 Workshop on Statistical Machine Translation*. Proceedings of the Fourth Workshop on Statistical Machine Translation, pages 1–28. Athens, Greece.

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).

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 483490.

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 595603, Uppsala, Sweden, 11-16 July 2010.

P. Koehn. 2005. *Europarl: A Parallel Corpus for Statistical Machine Translation*. Proceedings of the Machine Translation Summit X, pages 79-86. Phuket, Thailand.

P. Koehn. 2010. *Statistical Machine Translation*. Cambridge University Press.

P. Koehn and F. J. Och and D. Marcu. 2003. *Statistical Phrase-Based Translation*, Proceedings of the North America Meeting on Association for Computational Linguistics, 48–54.

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.

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.

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.

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.

J. Savoy, and L. Dolamic. 2009. *How effective is Google's translation service in search?*. Communications of the ACM, 52(10):139–143.

R. Steinberger and B. Pouliquen and A. Widiger and C. Ignat and T. Erjavec and D. Tufiş and D. Varga. 2006. *The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages*. Proceedings of the 5th International Conference on Language Resources and Evaluation, pages 2142–2147. Genova, Italy.

J. Tiedemann. 2009. *News from OPUS-A Collection of Multilingual Parallel Corpora with Tools and Interfaces*. Recent advances in natural language processing V: selected papers from RANLP 2007, pages 309:237.

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.

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.

Language	SMT system	Nr. of unigrams	Nr. of bigrams
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.

Language	SMT	Test Set	SMO	AdaBoost M1	Bagging	Bleu Score
English		GS	0.685	0.685	0.686	
To German	Bing	GS	0.641	0.631	0.648	
		Tr	0.658	0.636	0.662	0.227
To German	Google T.	GS	0.646	0.623	0.674	
		Tr	0.687	0.645	0.661	0.209
To German	Moses	GS	0.644	0.644	0.676	
		Tr	0.667	0.667	0.674	0.17
To Spanish	Bing	GS	0.656	0.658	0.646	
		Tr	0.633	0.633	0.633	0.316
To Spanish	Google T.	GS	0.653	0.653	0.665	
		Tr	0.636	0.667	0.636	0.341
To Spanish	Moses	GS	0.664	0.664	0.671	
		Tr	0.649	0.649	0.663	0.298
To French	Bing	GS	0.644	0.645	0.664	
		Tr	0.644	0.649	0.652	0.243
To French	Google T.	GS	0.64	0.64	0.659	
		Tr	0.652	0.652	0.678	0.274
To French	Moses	GS	0.633	0.633	0.645	
		Tr	0.666	0.666	0.674	0.227

Table 2: Results obtained using the individual training sets obtained by translating with each of the three considered MT systems, to each of the three languages considered.