Resource Creation and Evaluation for Multilingual Sentiment Analysis in Social Media Texts

Alexandra Balahur*, Marco Turchi^, Ralf Steinberger*, Jose-Manuel Perea-Ortega*, Guillaume Jacquet*, Dilek Küçük*, Vanni Zavarella*, Adil El Ghali*

*European Commission Joint Research Centre
Via E. Fermi 2749, 21027 Ispra (VA), Italy
E-mail: Firstname. Lastname@jrc.ec.europa.eu
^Fundazione Bruno Kessler
Via Sommarive, 18
Povo, Trento, Italy
turchi@fbk.e

Abstract

This paper presents an evaluation of the use of machine translation to obtain and employ data for training multilingual sentiment classifiers. We show that the use of machine translated data obtained similar results as the use of native-speaker translations of the same data. Additionally, our evaluations pinpoint to the fact that the use of multilingual data, including that obtained through machine translation, leads to improved results in sentiment classification. Finally, we show that the performance of the sentiment classifiers built on machine translated data can be improved using original data from the target language and that even a small amount of such texts can lead to significant growth in the classification performance.

Keywords: sentiment analysis, multilingual sentiment analysis, machine translation, mixed language classifiers.

1. Introduction

Sentiment analysis (SA) regards the classification of texts according to the polarity of the opinions they express. SA systems are highly relevant to many real-world applications (e.g. marketing, eGovernance, business intelligence, behavioural sciences) and also to many tasks in Natural Language Processing (NLP) – information extraction, question answering, textual entailment, to name just a few. The importance of this field has been proven by the high number of approaches proposed in research, as well as by the interest that it raised from other disciplines and the applications that were created using its technology.

In our case, the primary focus is to use sentiment analysis in the context of media monitoring, to enable tracking of global reactions to events. The main challenge that we face in this endeavour is that tweets are written in different languages and an unbiased system should be able to deal with all of them, in order to process all (possible) available data.

Unfortunately, although many linguistic resources exist for processing texts written in English, for many other languages, data and tools are scarce. Following our initial efforts described in (Balahur & Turchi; 2012, 2013, 2014), in this article we extend our study on the possibility to implement a multilingual system that is able to: a) classify sentiment expressed in tweets in various languages using training data obtained through machine translation; b) to verify the extent to which the quality of the translations influences the sentiment classification performance, in this case, of highly informal texts; and c) to improve multilingual sentiment classification using small amounts of data annotated in the target language. To this aim, varying sizes of target language data are tested. The languages we explore are: Turkish, Italian, Spanish, German and French.

2. Background and Motivation

In order to produce multilingual resources for subjectivity and sentiment analysis, different approaches have been proposed, mainly based on translations of English annotated resources. They include the use of bilingual dictionaries (Banea et al., 2008) or the use of machine translation systems (Wan, 2009; Kim et al., 2010; Banea et al., 2010), in conjunction to supervised or semi-supervised learning. In our approaches presented in (Balahur & Turchi, 2012 and 2014), we employed three different machine translation systems - Bing, Google and Moses (Koehn et al., 2007) - and evaluated the impact of translation quality on the sentiment classification performance. We also tested the possibility to employ meta-classifiers, such as AdaBoost and Bagging to remove the noise introduced by machine translation. All these approaches showed that machine translation has reached a level of maturity that allows for it to be used to obtain sufficiently accurate resources for multilingual SA systems.

Finally, other research approaches the issue of sentiment dictionaries creation in other languages using a method called “triangulation”, which involves a translation step and manual corrections (Steinberger et al., 2011). They translate the data, in parallel, from English and Spanish to other languages and obtain dictionaries from the intersection of these two translations.

As far as tweet processing and sentiment analysis in tweets is concerned, Go et al. (2009) pioneered research proposing the use of emoticons (e.g. ":)", ":(", etc.) as markers of positive and negative tweets. Read (2005) employed this method to generate a corpus of positive tweets, with positive emoticons ":)", and negative tweets with negative emoticons ":(". Pak and Paroubek (2010) also generated a corpus of tweets for sentiment analysis, by selecting positive and negative tweets based on the presence of specific emoticons. These approaches employed different supervised approaches for sentiment analysis, using n-grams as features. Zhang et al. (2011) employ a hybrid approach, combining supervised learning with the knowledge on sentiment-bearing words, which they extract from the DAL sentiment dictionary (Whissell, 1989). The authors conclude that the most important features are those corresponding to sentiment-bearing words. Finally, Jiang et al. (2011) classify sentiment expressed on previously-given "targets" in tweets adding the context of the tweet to increase the text length.

In our previous work (Balahur & Turchi, 2013), we showed that employing training data obtained by machine translation, we could build statistical classifiers that could accurately categorize tweets written in different languages into positive, negative and neutral. Further on, we showed that adding training data from different languages and especially adding data from languages with similar etymology, we can significantly improve the results of the sentiment classification performance. In this paper, we present the extension of these experiments to other languages and test the influence of increasing amounts of "correct" target language data on the sentiment classification performance.

3. Datasets

In order to carry out the different experiments, we chose two datasets that are available from open competitions: SemEval 2013 Task 2 – Sentiment Analysis in Tweets (Wilson et al., 2013) English Tweets dataset and TASS (Taller de Analisis de Sentimientos y Subjetividad) 2013 Spanish Tweets datasets. As such, others can employ the data and perform similar experiments and we can also compare the results obtained with the results obtained by the participants in the official runs. Both these datasets contain tweets annotated with polarity (positive, negative and neutral). The TASS dataset contains also a finer-grained classification, into objective, positive, high positive, negative, high negative and neutral classes.

From the SemEval 2013 data, we employ the training (T*) and development (t*) sets. Their characteristics are described in Table 1. The training set is used to extract the features for the classification models and the development set is employed for testing purposes. Both these datasets have been translated to Arabic, Turkish, Russian, Italian, Spanish, German and French.

Data	#Tweet	#Pos.	#Neg.	#Neu.
T*	6688	2450	956	3282
t*	1051	386	199	466

Table 1: Characteristics of training (T*) and testing (t*) datasets from SemEval 2013 Task 2.

In order to have a Gold Standard for evaluation, in (Balahur & Turchi, 2013), we manually corrected the development set (t*) for Italian, Spanish, German and French at the level of word ordering and translation. We thus obtained the Gold Standard. Subsequently, native speakers co-authors corrected the development set in the sense of including the vocabulary a native speaker would employ to express the meaning of the sentences (i.e. translation and interpretation of the texts) leading to Gold Standard 2.

The TASS 2013 dataset was split into training (TASS*) and testing (tass*). The characteristics of these sets are presented in Table 2.

Data	#Tweet	#Pos.	#Neg.	#Neu.
TASS*	7219	2783	2124	2312
tass*	60798	22233	15844	22721

Table 2: Characteristics of training (TASS*) and testing (tass*) datasets from TASS 2013.

This dataset was employed to evaluate the impact of translated data on the sentiment analysis performance on a dataset in the original language and to test the possibility to employ "correct" data from the target language (in this case, Spanish) to improve the performance of the sentiment classification.

We can observe that the distribution of examples per polarity class is significantly different among the two datasets. From this point of view, the TASS data is more balanced across the three polarity classes, and the SemEval data has predominantly more neutral examples and a 3:1 proportion of positive to negative examples. We can also notice that the SemEval data is similar as number of examples to the training dataset supplied in TASS, a fact which will be exploited in the experiments presented in Section 5.1, where we test the possibility to improve sentiment classification performance using small amounts of original target language data.

4. Sentiment Analysis of Multilingual Data

Our sentiment analysis system is based on a hybrid approach, which employs supervised learning with a Support Vector Machines Sequential Minimal Optimization (Platt, 1998) linear kernel, on unigram and bigram features, but exploiting as features sentiment dictionaries, emoticon lists, slang lists and other social media-specific features. We do not employ any specific language analysis software. The aim is to be able to apply, in a straightforward manner, the same approach to as many languages as possible. The approach can be extended to other languages by using similar dictionaries

that have been created in our team. They were built using the same dictionaries we employ in this work and their corrected translation to Spanish. The new sentiment dictionaries were created by simultaneously translating from these two languages to a third one and considering the intersection of the translations as correct terms. Currently, new such dictionaries have been created for 15 other languages.

The sentiment analysis process contains two stages: pre-processing and sentiment classification. Pre-processing involves tokenization, normalization of language (only done for English) and the addressing of special signals of emotion in informal texts – emoticons, punctuation signs, and capitalization (which are marked correspondingly). Once the tweets are pre-processed, they are passed on to the sentiment classification module.
For the sentiment classification, we employ supervised learning using the Support Vector Machines Sequential Minimal Optimization (Platt, 2005) implementation in Weka[1], with a linear kernel, based on boolean features - the presence or absence of n-grams (unigrams, bigrams and unigrams plus bigrams) determined from the training data (tweets that were previously pre-processed as described above). Bigrams are used specifically to spot the influence of modifiers (negations, intensifiers, diminishers) on the polarity of the sentiment-bearing words. This approach was successfully employed for English and although for other languages other (additional or slightly different) features might be useful to be included, at this point we employ the same approach for all the languages considered, for comparison reasons.

5. Evaluation and Discussion

5.1. Evaluation of sentiment analysis on translated data

Our initial experiments included the evaluation of the English, Spanish, Italian, French and German data individually and considering different combinations thereof, on Gold Standard 1. In these preliminary evaluations, the datasets were corrected by non-native speakers and only the faulty order of words and word choice were edited. The results are presented in Figure 1. From this evaluation, we can see that the performance of the different pairs of languages compared to individual results, we can: a) on the one hand, see that combining languages with a comparatively high difference in performance results in an increase of the lower-performing one and b) on the other hand, in some cases, the overall performance is improved on both systems, which shows that combining this data helps to disambiguate the contextual use of specific words. Finally, the results show that the use of all the languages together improves the overall classification of sentiment in the data. This shows that a multilingual system can simply employ joint training data from different languages in a single classifier, thus making the sentiment classification straightforward, not needing any language detection software or training different classifiers.

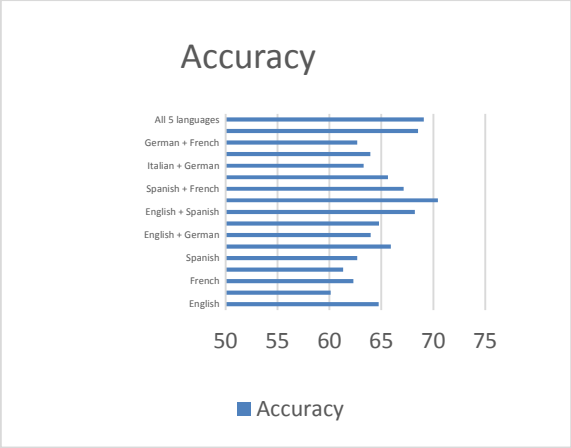

Figure 1: Results obtained classifying each language individually versus on pairs and families of languages.

Subsequently, the translated development set was also corrected by native speakers, who also modified the slang and expressions to match the "normal", informal speech in the respective language. A new language was also added, to quantify the extent to which its lack of relatedness to any of the languages considered had any influence on the final results. This language is Turkish. This new datasets were denominated Gold Standard 2. We performed the same tests as with Gold Standard 1. The results are presented in Figure 2.

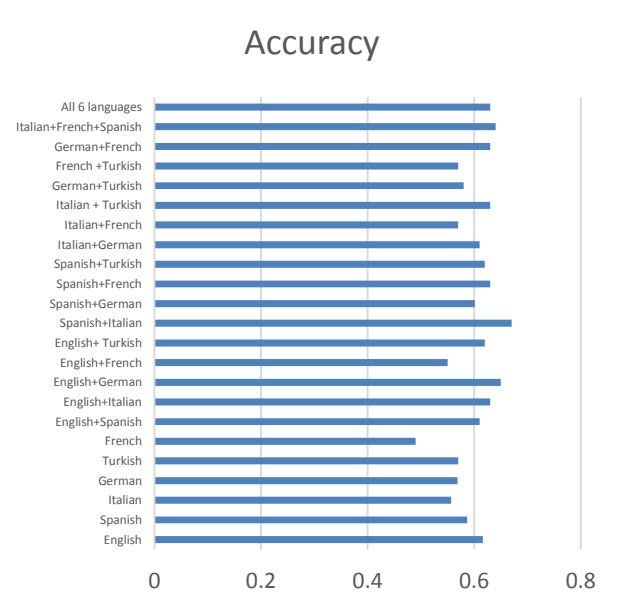

Figure 2: Results obtained classifying each language individually versus on pairs and families of languages on Gold Standard 2.

In order to see whether the results correlate with the quality of the machine translation, we computed the BLEU score (Papineni et al., 2002) between the machine translated version of the development data and the native-speaker corrected versions. The results were: 39.11 for German, 72.88 for Spanish, 54.84 for French, 59.98 for Italian and 28.77 for Turkish. In this case, the better results correlate only with Spanish. Surprising is the case of Turkish, where the low score of the machine translation

[1] http://www.cs.waikato.ac.nz/ml/weka/

did not influence the quality of the sentiment classification. At the same time, using features from this language has reinforced the relevant features from the other languages, leading to improved results in the combination thereof.

5.2. Test of the influence of target language data on the sentiment analysis performance

Another experiment we performed was meant to quantify the impact of machine-translated data in conjunction to "correct" data from the target language. In order to test this setting, we employed the SemEval data translated to Spanish and the TASS 2013 training and test sets.
We evaluated our approach using the TASS 2013 test data (tass*), under different settings: employing for training only the translated SemEval data, adding successively quarters of training data from TASS* (the training set) and using only the TASS* training set. The results are presented in Tables 3 and 4 in terms of precision, recall and F1 measure for each of the classes and globally. In Figures 3, 4 and 5, we present a visualization of these evaluations, per polarity class (positive, negative and neutral).

	Sem Eval only	SemEval + ¼ TASS*	SemEv. + ½ TASS*	SemEv. + ¾ TASS*	TASS *+ Sem Eval	TASS *only
P_{pos}	0.64	0.65	0.67	0.68	0.68	0.67
R_{pos}	0.51	0.58	0.61	0.64	0.66	0.66
P_{neg}	0.54	0.54	0.56	0.57	0.58	0.57
R_{neg}	0.23	0.51	0.59	0.62	0.64	0.64
P_{neu}	0.47	0.53	0.56	0.59	0.66	0.60
R_{neu}	0.75	0.61	0.59	0.58	0.48	0.55

Table 3: Results in terms of Precision and Recall for the classification of the TASS 2013 test set using as training the translated SemEval data and increasing quantities of TASS training data.

	Sem Eval only	Sem Eval+ ¼ TASS*	Sem Eval+ ½ TASS*	Sem Eval+ ¾ TASS*	TASS *+ SemEval	TASS* only
$F1_{pos}$	0.57	0.61	0.64	0.66	0.67	0.66
$F1_{neg}$	0.32	0.53	0.58	0.60	0.61	0.60
$F1_{neu}$	0.58	0.57	0.58	0.58	0.56	0.58
$F1_{glob}$	0.49	0.57	0.60	0.62	0.61	0.61

Table 4: Results in terms of Precision and Recall for the classification of the TASS 2013 test set using as training the translated SemEval data and increasing quantities of TASS training data.

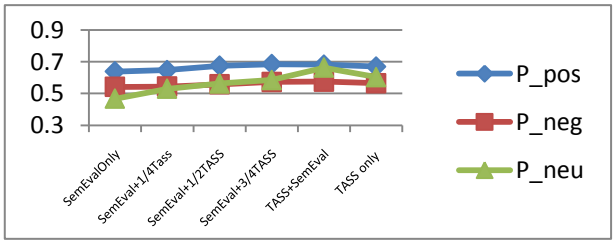

Figure 3: Results in terms of Precision (per polarity class) for the classification of the TASS 2013 test set using as training the translated SemEval data and increasing quantities of TASS training data.

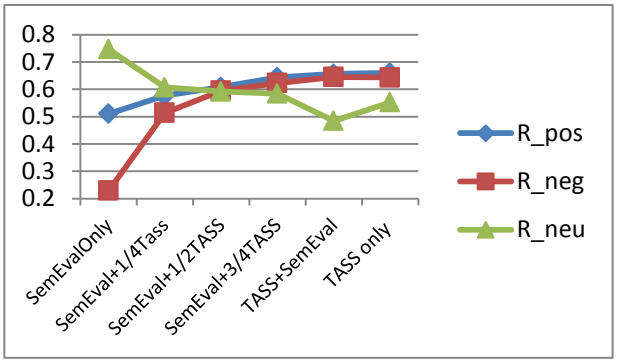

Figure 4: Results in terms of Recall (per polarity class) for the classification of the TASS 2013 test set using as training the translated SemEval data and increasing quantities of TASS training data.

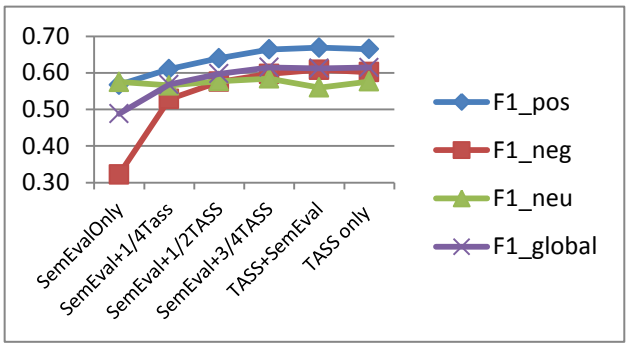

Figure 5: Results in terms of F1 measure (per polarity class) for the classification of the TASS 2013 test set using as training the translated SemEval data and increasing quantities of TASS training data.

When interpreting the results, we should first keep in mind that the SemEval and TASS training corpora are similar in size (about 7000 tweets each), but that they are distributed differently over the polarity classes. While TASS is quite balanced, SemEval has predominantly neutral examples and a proportion of 3:1 of positive to negative examples. In spite of this difference, we can see that when adding small amounts of data from the target language, the results improve, not only as far as F-measure is concerned, but also in the sense of classifying positive and negative examples more accurately (which, in fact, is our goal). Using the "TASS+SemEval" and the "TASS only" training sets leads to comparable results as far as F-measure is concerned, but at significantly different values for the three classes (result given by the $\chi 2$ test) as far as positive and negative versus neutral class are concerned. This shows that the use of training data obtained by translation is useful and the noise introduced does not hinder the results, but brings more information to better discriminate positive and negative statements from the rest.

6. Conclusions and Future Work

In this paper, we evaluated the use of machine translation to obtain and employ data for training multilingual sentiment classifiers. We showed that the use of machine

translated data obtained similar results as the use of native-speaker translations of the same data. Subsequently, we also showed that the use of multilingual data, including that obtained through machine translation, leads to improved results in sentiment classification. This is due to the fact that, using multiple languages to build the classifiers, the features that are relevant are automatically selected (as the feature space becomes sparser). Finally, we showed that the performance of the sentiment classifiers built on machine translated data can be improved using original data from the target language and that even a small amount of such texts can lead to significant growth in the classification performance.

Future work includes the use of meta-classifiers to improve sentiment classification using mixed language models and the use of larger quantities of target language data to improve the quality of machine translated data.

Finally, the Gold Standards employed in this evaluation will be made available for the research community, so that comparisons with other approaches to the multilingual sentiment analysis issue can be made.

7. Acknowledgements

The work by Dilek Kucük is supported in part by a postdoctoral research grant from TÜBİTAK.

8. References

1. Balahur, A. & Turchi, M. (2014). Comparative Experiments Using Supervised Learning and Machine Translation for Multilingual Sentiment Analysis. In Computer Speech and Language, 28(1), pp. 56-75.
2. Balahur, A. & Turchi, M. (2012). Multilingual Sentiment Analysis using Machine Translation? Proceedings of the Association for Computational Linguistics: 3rd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis, pp. 52-60, Jeju, Republic of Korea, 12 July 2012.
3. Banea, C., Mihalcea, R., Wiebe, J. (2008). A bootstrapping method for building subjectivity lexicons for languages with scarce resources. In: Proceedings of the Conference on Language Resources and Evaluations (LREC 2008), Maraakesh, Marocco.
4. Banea, C., Mihalcea, R., Wiebe, J. (2010). Multilingual subjectivity: are more languages better? In: Proceedings of the International Conference on Computational Linguistics (COLING 2010), Beijing, China. pp. 28–36.
5. Go, A., Bhayani, R., Huang, L. (2009). Twitter sentiment classification using distant supervision. Processing, pp. 1–6.
6. Jiang, L., Mo Yu, Ming Zhou, Xiaohua Liu, and Tiejun Zhao (2011). Target-dependent twitter sentiment classification. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies – Volume 1, HLT '11, pages 151–160, Stroudsburg, PA, USA. Association for Computational Linguistics.
7. Kim, J., Li, J.-J., Lee, J.-H. (2010). Evaluating multilanguage-comparability of subjectivity analysis systems. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. pp. 595–602.
8. Koehn, P., Och, F. J., Marcu., D. (2003). Statistical phrase-based translation. In: Proceedings of the North America Meeting on Association for Computational Linguistics. pp. 48–54.
9. Pak, A. & Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10), Valletta, Malta; ELRA, may. European Language Resources Association, pp. 19-21.
10. Papineni, K., Roukos, S., Ward, T., Zhu, W. J. (2002). "BLEU: a method for automatic evaluation of machine translation". ACL-2002: 40th Annual meeting of the Association for Computational Linguistics. pp. 311–318.
11. Platt, J. C. (1999). Fast training of support vector machines using sequential minimal optimization. MIT Press, Cambridge, MA, USA, pp. 185–208. URL http://dl.acm.org/citation.cfm?id=299094.299105
12. Read, J. (2005). Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In Proceedings of the ACL Student Research Workshop, ACL student '05, pages 43–48, Stroudsburg, PA, USA. Association for Computational Linguistics.
13. Steinberger, J., Lenkova, P., Ebrahim, M., Ehrman, M., Hurriyetoglu, A., Kabadjov, M., Steinberger, R., Tanev, H., Zavarella, V., Vazquez, S. (2011). Creating sentiment dictionaries via triangulation. In Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis, Portland, Oregon.
14. Wan, X., Li, H., Xiao, J. (2010). Cross-language document summarization based on machine translation quality prediction. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. pp. 917–926.
15. Whissell, C. (1989). The Dictionary of Affect in Language. In Robert Plutchik and Henry Kellerman, editors, Emotion: theory, research and experience, volume 4, The measurement of emotions. Academic Press, London.
16. Wilson, T., Kozareva, Z., Nakov, P., Rosenthal, S., Stoyanov, V., Ritter, A. (2013). SemEval-2013 task 2: Sentiment analysis in Twitter. In Proceedings of the International Workshop on Semantic Evaluation, SemEval '13, June.
17. Zhang, L., Riddhiman Ghosh, Mohamed Dekhil, Meichun Hsu, and Bing Liu. (2011). Combining lexicon-based and learning-based methods for Twitter sentiment analysis. Technical Report HPL-2011-89, HP, 21/06/2011.