Creating Sentiment Dictionaries via Triangulation

Josef Steinberger,
Polina Lenkova, **Mohamed Ebrahim**,
Maud Ehrmann, **Ali Hurriyetoglu**,
Mijail Kabadjov, **Ralf Steinberger**,
Hristo Tanev and **Vanni Zavarella**
EC Joint Research Centre
21027, Ispra (VA), Italy
Name.Surname@jrc.ec.europa.eu

Silvia Vázquez
Universitat Pompeu Fabra
Roc Boronat, 138
08018 Barcelona
silvia.vazquez@upf.edu

Abstract

The paper presents a semi-automatic approach to creating sentiment dictionaries in many languages. We first produced high-level gold-standard sentiment dictionaries for two languages and then translated them automatically into third languages. Those words that can be found in both target language word lists are likely to be useful because their word senses are likely to be similar to that of the two source languages. These dictionaries can be further corrected, extended and improved. In this paper, we present results that verify our triangulation hypothesis, by evaluating triangulated lists and comparing them to non-triangulated machine-translated word lists.

1 Introduction

When developing software applications for sentiment analysis or opinion mining, there are basically two main options: (1) writing rules that assign sentiment values to text or text parts (e.g. names, products, product features), typically making use of dictionaries consisting of sentiment words and their positive or negative values, and (2) inferring rules (and sentiment dictionaries), e.g. using machine learning techniques, from previously annotated documents such as product reviews annotated with an overall judgment of the product. While movie or product reviews for many languages can frequently be found online, sentiment-annotated data for other fields are not usually available, or they are almost exclusively available for English. Sentiment dictionaries are also mostly available for English only or, if they exist for other languages, they are not comparable, in the sense that they have been developed for different purposes, have different sizes, are based on different definitions of what sentiment or opinion means.

In this paper, we are addressing the resource bottleneck for sentiment dictionaries, by developing highly multilingual and comparable sentiment dictionaries having similar sizes and based on a common specification. The aim is to develop such dictionaries, consisting of typically one or two thousand words, for tens of languages, although in this paper we only present results for eight languages (English, Spanish, Arabic, Czech, French, German, Italian and Russian). The task raises the obvious question how the human effort of producing this resource can be minimized. Simple translation, be it using standard dictionaries or using machine translation, is not very efficient as most words have two, five or ten different possible translations, depending on context, part-of-speech, etc.

The approach we therefore chose is that of triangulation. We first produced high-level gold-standard sentiment dictionaries for two languages (English and Spanish) and then translated them automatically into third languages, e.g. French. Those words that can be found in both target language word lists (En Fr and Es Fr) are likely to be useful because their word senses are likely to be similar to that of the two source languages. These word lists can then be used as they are or better they can be corrected, extended and improved. In this paper, we present evaluation results verifying our triangulation hypothesis, by evaluating triangulated lists and comparing them

Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis, ACL-HLT 2011, pages 28–36,
24 June, 2011, Portland, Oregon, USA ©2011 Association for Computational Linguistics

to non-triangulated machine-translated word lists.

Two further issues need to be addressed. The first one concerns morphological inflection. Automatic translation will yield one word form (often, but not always the base form), which is not sufficient when working with highly inflected languages: A single English adjective typically has four Spanish or Italian word forms (two each for gender and for number) and many Russian word forms (due to gender, number and case distinctions). The target language word lists thus need to be expanded to cover all these morphological variants with minimal effort and considering the number of different languages involved without using software, such as morphological analysers or generators. The second issue has to do with the subjectivity involved in the human annotation and evaluation effort. First of all, it is important that the task is well-defined (this is a challenge by itself) and, secondly, the inter-annotator agreement for pairs of human evaluators working on different languages has to be checked in order to get an idea of the natural variation involved in such a highly subjective task.

Our main field of interest is news opinion mining. We would like to answer the question how certain entities (persons, organisations, event names, programmes) are discussed in different media over time, comparing different media sources, media in different countries, and media written in different languages. One possible end product would be a graph showing how the popularity of a certain entity has changed over time across different languages and countries. News differs significantly from those text types that are typically analysed in opinion mining work, i.e. product or movie reviews: While a product review is about a product (e.g. a printer) and its features (e.g. speed, price or printing quality), the news is about any possible subject (news content), which can by itself be perceived to be positive or negative. Entities mentioned in the news can have many different roles in the events described. If the method does not specifically separate positive or negative news content from positive or negative opinion about that entity, the sentiment analysis results will be strongly influenced by the news context. For instance, the automatically identified sentiment towards a politician would most likely to be low if the politician is mentioned in the context of negative news content such as bombings or disasters. In our approach, we therefore aim to distinguish news content from sentiment values, and this distinction has an impact on the sentiment dictionaries: unlike in other approaches, words like death, killing, award or winner are purposefully not included in the sentiment dictionaries as they typically represent news content.

The rest of the paper is structured as follows: the next section (2) describes related work, especially in the context of creating sentiment resources. Section 3 gives an overview of our approach to dictionary creation, ranging from the automatic learning of the sentiment vocabulary, the triangulation process, the expansion of the dictionaries in size and regarding morphological inflections. Section 4 presents a number of results regarding dictionary creation using simple translation versus triangulation, morphological expansion and inter-annotator agreement. Section 5 summarises, concludes and points to future work.

2 Related Work

Most of the work in obtaining subjectivity lexicons was done for English. However, there were some authors who developed methods for the mapping of subjectivity lexicons to other languages. Kim and Hovy (2006) use a machine translation system and subsequently use a subjectivity analysis system that was developed for English. Mihalcea et al. (2007) propose a method to learn multilingual subjective language via cross-language projections. They use the Opinion Finder lexicon (Wilson et al., 2005) and 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 in Banea et al. (2008b). To this aim, the authors perform three different experiments, with good 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. Finally, another approach to building lexicons for languages with scarce resources is presented in 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 obtained by translating words in the Opinion Finder lexicon. Translations are filtered using a measure of similarity to the original words, based on Latent Semantic Analysis (Landauer and Dumais, 1997) scores. Wan (2008) 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. 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 Naive Bayes classifier using the same data translated to 2 or 3 other languages. Their conclusion is that 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.

3 Approach Overview

Our approach to dictionary creation starts with semi-automatic way of colleting subjective terms in English and Spanish. These pivot language dictionaries are then projected to other languages. The 3rd language dictionaries are formed by the overlap of the translations (triangulation). The lists are then manually filtered and expanded, either by other relevant terms or by their morphological variants, to gain a wider coverage.

3.1 Gathering Subjective Terms

We started with analysing the available English dictionaries of subjective terms: General Inquirer (Stone et al., 1966), WordNet Affect (Strapparava and Valitutti, 2004), SentiWordNet (Esuli and Sebastiani, 2006), MicroWNOp (Cerini et al., 2007). Additionally, we used the resource of opinion words with associated polarity from Balahur et al. (2009), which we denote as JRC Tonality Dictionary. The positive effect of distinguishing two levels of intensity was shown in (Balahur et al., 2010). We followed the idea and each of the emloyed resources was mapped to four categories: positive, negative, highly positive and highly negative. We also got inspired by the results reported in that paper and we selected as the base dictionaries the combination of MicroWNOp and JRC Tonality Dictionary which gave the best results. Terms in those two dictionaries were manually filtered and the other dictionaries were used as lists of candidates (their highly frequent terms were judged and the relevant ones were included in the final English dictionary). Keeping in mind the application of the dictionaries we removed at this step terms that are more likely to describe bad or good news content, rather than a sentiment towards an entity. In addition, we manually collected English diminishers (e.g. *less* or *approximately*), intensifiers (e.g. *very* or *indeed*) and invertors (e.g. *not* or *barely*). The English terms were translated to Spanish and the same filtering was performed. We extended all English and Spanish lists with the missing morphological variants of the terms.

3.2 Automatic Learning of Subjective Terms

We decided to expand our subjective term lists by using automatic term extraction, inspired by (Riloff and Wiebe, 2003). We look at the problem of acquisition of subjective terms as learning of semantic classes. Since we wanted to do this for two different languages, namely English and Spanish, the multilingual term extraction algorithm Ontopopulis (Tanev et al., 2010) was a natural choice.

Ontopopulis performs weakly supervised learning of semantic dictionaries using distributional similarity. The algorithm takes on its input a small set of seed terms for each semantic class, which is to be learnt, and an unannotated text corpus. For example, if we want to learn the semantic class *land_vehicles*, we can use the seed set - *bus*, *truck*, and *car*. Then it searches for the terms in the corpus and finds linear context patterns, which tend to co-occur immediately before or after these terms. Some of the highest-scored patterns, which Ontopopulis learned about *land_vehicles* were *driver of the X*, *X was parked*, *collided with another X*, etc. Finally, the algorithm searches for these context patterns in the corpus and finds other terms which tend to fill the slot of the patterns (designated by X). Considering the *land_vehicles* example, new terms which the system learned were *van*, *lorry*, *taxi*, etc. Ontopopulis is similar to the NOMEN algorithm (Lin et al., 2003). However, Ontopopulis has the advantage to be language-independent, since it does not use any form of language-specific processing, nor does it use any language-specific resources, apart from a stop word list.

In order to learn new subjective terms for each of the languages, we passed the collected subjective terms as an input to Ontopopulis. For English, we divided the seed set in two classes: class A – verbs and class B – nouns and adjectives. It was necessary because each of these classes has a different syntactic behaviour. It made sense to do the same for Spanish, but we did not have enough Spanish speakers available to undertake this task, therefore we put together all the subjective Spanish words - verbs, adjectives and nouns in one class. We ran Ontopopulis for each of the three classes - the class of subjective Spanish words and the English classes A and B. The top scored 200 new learnt terms were taken for each class and manually reviewed.

3.3 Triangulation and Expansion

After polishing the pivot language dictionaries we projected them to other languages. The dictionaries were translated by Google translator because of its broad coverage of languages. The overlapping terms between English and Spanish translations formed the basis for further manual efforts. In some cases there were overlapping terms in English and Spanish translations but they differed in intensity. There was the same term translated from an English positive term and from a Spanish very positive term. In these cases the term was assigned to the positive category. However, more problematic cases arose when the same 3rd language term was assigned to more than one category. There were also cases with different polarity. We had to review them manually. However, there were still lots of relevant terms in the translated lists which were not translated from the other language. These *complement* terms are a good basis for extending the coverage of the dictionaries, however, they need to be reviewed manually. Even if we tried to include in the pivot lists all morphological variants, in the triangulation output there were only a few variants, mainly in the case of highly inflected languages. To deal with morphology we introduced wild cards at the end of the term stem (* stands for whatever ending and _ for whatever character). This step had to be performed carefully because some noise could be introduced. See the Results section for examples. Although this step was performed by a human, we checked the most frequent terms afterwards to avoid irrelavant frequent terms.

4 Results

4.1 Pivot dictionaries

We gathered and filtered English sentiment terms from the available corpora (see Section 3.1). The dictionaries were then translated to Spanish (by Google translator) and filtered afterwards. By applying automatic term extraction, we enriched the sets of terms by 54 for English and 85 for Spanish, after evaluating the top 200 candidates suggested by the Ontopolulis tool for each language. The results are encouraging, despite the relevance of the terms (27% for English and 42.5% for Spanish where some missing morphological variants were discovered) does not seem to be very high, considering the fact that we excluded the terms already contained in the pivot lists. If we took them into account, the precision would be much better. The initial step resulted in obtaining high quality pivot sentiment dictionaries for English and Spanish. Their statistics are in table 1. We gathered more English terms than Spanish (2.4k compared to 1.7k). The reason for that is that some translations from English to Spanish have been filtered. Another observation is that there is approximately the same number of negative terms as positive ones, however, much more highly negative than highly positive terms. Although the

Language	English	Spanish
HN	554	466
N	782	550
P	772	503
HP	171	119
INT	78	62
DIM	31	27
INV	15	10
TOTAL	2.403	1.737

Table 1: The size of the pilot dictionaries. HN=highly negative terms, N=negative, P=positive, HP=highly positive, INV=invertors, DIM=diminishers, INV=invertors.

frequency analysis we carried out later showed that even if there are fewer highly positive terms, they are more frequent than the highly negative ones, which results in almost uniform distribution.

4.2 Triangulation and Expansion

After running triangulation to other languages the resulted terms were judged for relevance. Native speakers could suggest to change term's category (e.g. negative to highly negative) or to remove it. There were several reasons why the terms could have been marked as 'non-sentiment'. For instance, the term could tend to describe rather negative news content than negative sentiment towards an entity (e.g. *dead*, *quake*). In other cases the terms were too ambiguous in a particular language. Examples from English are: *like* or *right*.

Table 2 shows the quality of the triangulated dictionaries. In all cases except for Italian we had only one annotator assessing the quality. We can see that the terms were correct in around 90% cases, however, it was a little bit worse in the case of Russian in which the annotator suggested to change category very often.

Terms translated from English but not from Spanish are less reliable but, if reviewed manually, the dictionaries can be expanded significantly. Table 3 gives the statistics concerning these judgments. We can see that their correctness is much lower than in the case of the triangulated terms - the best in Italian (54.4%) and the worst in Czech (30.7%). Of course, the translation performance affects the results here. However, this step extended the dictionaries by approximately 50%.

When considering terms out of context, the most common translation error occurs when the original word has several meanings. For instance, the English word *nobility* refers to the social class of nobles, as well as to the quality of being morally good. In the news context we find this word mostly in the second meaning. However, in the Russian triangulated list we have found *dvoryanstvo* , which refers to a social class in Russian. Likewise, we need to keep in mind that a translation of a monosemantic word might result polysemantic in the target language, thereby leading to confusion. For example, the Italian translation of the English word *champion campione* is more frequently used in Italian news context in a different meaning - *sample*, therefore we must delete it from our sentiment words list for Italian. Another difficulty we might encounter especially when dealing with inflectional languages is the fact that a translation of a certain word might be homographic with another word form in the target language. Consider the English negative word *bandit* and its Italian translation *bandito*, which is more frequently used as a form of the verb *bandire* (*to announce*) in the news context. Also each annotator had different point of view on classifying the borderline cases (e.g. *support*, *agreement* or *difficult*).

Two main reasons are offered to explain the low performance in Arabic. On the one hand, it seems that some Google translation errors will be repeated in different languages if the translated words have the same etymological root. For example both words – the English *fresh* and the Spanish *fresca* – are translated to the Arabic as جديد meaning *new*. The Other reason is a more subtle one and is related to the fact that Arabic words are not vocalized and to the way an annotator perceive the meaning of a given word in isolation. To illustrate this point, consider the Arabic word المنَاسبه, which could be used as an adjective, meaning *appropriate*, or as a noun, meaning *The occasion*. It appears that the annotator would intuitively perceive the word in isolation as a noun and not as an adjective, which leads to disregarding the evaluative aspects of a given word.

We tried to include in the pivot dictionaries all morphological variants of the terms. However, in highly inflected languages there are much more variants than those translated from English or Spanish.

We manually introduced wild cards to capture the variants. We had to be attentive when compiling wild cards for languages with a rich inflectional system, as we might easily get undesirable words in the output. To illustrate this, consider the third person plural of the Italian negative word *perdere* (*to lose*) *perdono*, which is also homographic with the word meaning *forgiveness* in English. Naturally, it could happen that the wildcard captures a non-sentiment term or even a term with a different polarity. For instance, the pattern *care%* would capture either *care*, *careful*, *carefully*, but also *career* or *careless*. That is way we perform the last manual checking after matching the lists expanded by wildcards against a large number of texts. The annotators were unable to check all the variants, but only the most frequent terms, which resulted in reviewing 70-80% of the term mentions. This step has been performed for only English, Czech and Russian so far. Table 5 gives the statistics. By introducing the wildcards, the number of distinct terms grew up significantly - 12x for Czech, 15x for Russian and 4x for English. One reason why it went up also for English is that we captured compounds like: *well-arranged*, *well-balanced*, *well-behaved*, *well-chosen* by a single pattern. Another reason is that a single pattern can capture different POSs: *beaut%* can capture *beauty*, *beautiful*, *beautifully* or *beautify*. Not all of those words were present in the pivot dictionaries. For dangerous cases like *care%* above we had to rather list all possible variants than using a wildcard. This is also the reason why the number of patterns is not much lower than the number of initial terms. Even if this task was done manually, some noise was added into the dictionaries (92-94% of checked terms were correct). For example, highly positive pattern *hero%* was introduced by an annotator for capturing *hero*, *heroes*, *heroic*, *heroical* or *heroism*. If not checked afterwards *heroin* would score highly positively in the sentiment system. Another example is taken from Russian: word meaning *to steal ukra%* - might generate *Ukraine* as one most frequent negative word in Russian.

4.3 How subjective is the annotation?

Sentiment annotation is a very subjective task. In addition, annotators had to judge single terms without any context: they had to think about all the senses of the term. Only if the main sense was subjective they agreed to leave it in the dictionary. Another subjectivity level was given by concentrating on distinguishing news content and news sentiment. Defining the line between negative and highly negative terms, and similarly with positive, is also subjective. In the case of Italian we compared judgments of two annotators. The figures of inter-annotator agreement of annotating the triangulated terms are in table 6 and the complement terms in table 7. Based on the percent agreement the annotators agree a little bit less on the triangulated terms (76.6%) compared to the complement terms (82.6%). However, if we look at Kappa figures, the difference is clear. Many terms translated only from English were clearly wrong which led to a higher agreement between the annotators (0.318 compared to 0.614). When looking at the difference between positive and negative terms, we can see that there was higher agreement on the negative triangulated terms then on the positive ones.

Metric	Percent Agreement	Kappa
HN	0.909	0.465
N	0.796	0.368
P	0.714	0.281
HP	0.846	0
N+HN	0.829	0.396
P+HP	0.728	0.280
ALL	0.766	0.318

Table 6: Inter-annotator agreement on checking the triangulated list. In the case of HP all terms were annotated as correct by one of the annotators resulting in Kappa=0.

Metric	Percent Agreement	Kappa
HN	0.804	0.523
N	0.765	0.545
P	0.686	0.405
HP	0.855	0.669
N+HN	0.784	0.553
P+HP	0.783	0.559
ALL	0.826	0.614

Table 7: Inter-annotator agreement on checking the candidates. In ALL diminishers, intensifiers and invertors are included as well.

Language	Triangulated	Correct	Removed	Changed category
Arabic	926	606 (65.5%)	316 (34.1%)	4 (0.4%)
Czech	908	809 (89.1%)	68 (7.5%)	31 (3.4%)
French	1.085	956 (88.1%)	120 (11.1%)	9 (0.8%)
German	1.053	982 (93.3%)	50 (4.7%)	21 (2.0%)
Italian	1.032	918 (89.0%)	36 (3.5%)	78 (7.5%)
Russian	966	816 (84.5%)	49 (5.1%)	101 (10.4%)

Table 2: The size and quality of the triangulated dictionaries. Triangulated=No. of terms coming directly from triangulation, Correct=terms annotated as correct, Removed=terms not relevant to sentiment analysis, Change category=terms in wrong category (e.g., positive from triangulation, but annotator changed the category to highly positive).

Language	Terms	Correct	Removed	Changed category
Czech	1.092	335 (30.7%)	675 (61.8%)	82 (7.5%)
French	1.226	617 (50.3%)	568 (46.3%)	41 (3.4%)
German	1.182	548 (46.4%)	610 (51.6%)	24 (2.0%)
Italian	1.069	582 (54.4%)	388 (36.3%)	99 (9.3%)
Russian	1.126	572 (50.8%)	457 (40.6%)	97 (8.6%)

Table 3: The size and quality of the candidate terms (translated from English but not from Spanish). Terms=No. of terms translated from English but not from Spanish, Correct=terms annotated as correct, Removed=terms not relevant to sentiment analysis, Change category=terms in wrong category (e.g., positive in the original list, but annotator changed the category to highly positive).

Language	Terms	Correct	Removed	Changed category
Czech	2.000	1.144 (57.2%)	743 (37.2%)	113 (5.6%)
French	2.311	1.573 (68.1%)	688 (29.8%)	50 (2.1%)
German	2.235	1.530 (68.5%)	660 (29.5%)	45 (2.0%)
Italian	2.101	1.500 (71.4%)	424 (20.2%)	177 (8.4%)
Russian	2.092	1.388 (66.3%)	506 (24.2%)	198 (9.5%)

Table 4: The size and quality of the translated terms from English. Terms=No. of (distinct) terms translated from English, Correct=terms annotated as correct, Removed=terms not relevant to sentiment analysis, Change category=terms in wrong category (e.g., positive in the original list, but annotator changed the category to highly positive).

Language	Initial terms	Patterns	Matched terms		
			Count	Correct	Checked
Czech	1.257	1.063	15.604	93.0%	74.4%
English	2.403	2.081	10.558	93.8%	81.1%
Russian	1.586	1.347	33.183	92.2%	71.0%

Table 5: Statistics of introducing wild cards and its evaluation. Initial terms=checked triangulated terms extended by relevant translated terms from English, Patterns=number of patterns after introducing wildcards, Matched terms=terms matched in the large corpus - their count and correctness + checked=how many mentions were checked (based on the fact that the most frequent terms were annotated).

4.4 Triangulation vs. Translation

Table 4 present the results of simple translation from English (summed up numbers from tables 2 and 3). We can directly compare it to table 2 where only results of triangulated terms are reported. The performance of triangulation is significantly better than the performance of translation in all languages. The highest difference was in Czech (89.1% and 57.2%) and the lowest was in Italian (89.0% and 71.4%).

As a task-based evaluation we used the triangulated/translated dictionaries in the system analysing news sentiment expressed towards entities. The system analyses a fixed word window around entity mentions. Subjective terms are summed up and the resulting polarity is attached to the entity. Highly negative terms score twice more than negative, diminishers lower and intensifiers lift up the score. Invertors invert the polarity but for instance inverted highly positive terms score as only negative preventing, for instance, *not great* to score as *worst*. The system searches for the invertor only two words around the subjective term.

We ran the system on 300 German sentences taken from news gathered by the Europe Media Monitor (EMM)[1]. In all these cases the system attached a polarity to an entity mention. We ran it with three different dictionaries - translated terms from English, raw triangulated terms (without the manual checking) and the checked triangulated terms. This pilot experiment revealed the difference in performance on this task. When translated terms were used there were only 41.6% contexts with correct polarity assigned by the system, with raw triangulated terms 56.5%, and with checked triangulated terms 63.4%. However, the number does not contain neutral cases that would increase the overall performance. There are lots of reasons why it goes wrong here: the entity may not be the target of the subjective term (we do not use parser because of dealing with many languages and large amounts of news texts), the system can miss or apply wrongly an invertor, the subjective term is used in different sense, and irony is hard to detect.

4.5 State of progress

We finished all the steps for English, Czech and Russian. French, German, Italian and Spanish dictionaries miss only the introduction of wild cards. In Arabic we have checked only the triangulated terms. For other 7 languages (Bulgarian, Dutch, Hungarian, Polish, Portuguese, Slovak and Turkish) we have only projected the terms by triangulation. However, we have capabilities to finish all the steps also for Bulgarian, Dutch, Slovak and Turkish. We haven't investigated using more than two pivot languages for triangulation. It would probably results in more accurate but shortened dictionaires.

5 Conclusions

We presented our semi-automatic approach and current state of work of producing multilingual sentiment dictionaries suitable of assessing the sentiment in news expressed towards an entity. The triangulation approach works significantly better than simple translation but additional manual effort can improve it a lot in both recall and precision. We believe that we can predict the sentiment expressed towards an entity in a given time period based on large amounts of data we gather in many languages even if the per-case performance of the sentiment system as on a moderate level. Now we are working on improving the dictionaries in all the discussed languages. We also run experiments to evaluate the system on various languages.

Acknowledgments

We thank Alexandra Balahur for her collaboration and useful comments. This research was partly supported by a IULA-Universitat Pompeu Fabra grant.

[1] http://emm.newsbrief.eu/overview.html

References

Alexandra Balahur, Ralf Steinberger, Erik van der Goot, and Bruno Pouliquen. 2009. Opinion mining from newspaper quotations. In *Proceedings of the Workshop on Intelligent Analysis and Processing of Web News Content at the IEEE / WIC / ACM International Conferences on Web Intelligence and Intelligent Agent Technology (WI-IAT)*.

A. Balahur, R. Steinberger, M. Kabadjov, V. Zavarella, E. van der Goot, M. Halkia, B. Pouliquen, and J. Belyaeva. 2010. Sentiment analysis in the news. In *Proceedings of LREC'10*.

C. Banea, R. Mihalcea, and J. Wiebe. 2008a. A bootstrapping method for building subjectivity lexicons for languages with scarce resources. In *Proceedings of LREC*.

C. Banea, R. Mihalcea, J. Wiebe, and S. Hassan. 2008b. Multilingual subjectivity analysis using machine translation. In *Proceedings of EMNLP*.

C. Banea, R. Mihalcea, and J. Wiebe. 2010. Multilingual subjectivity: Are more languages better? In *Proceedings of COLING*.

S. Cerini, V. Compagnoni, A. Demontis, M. Formentelli, and G. Gandini. 2007. Micro-WNOp: A gold standard for the evaluation of automatically compiled lexical resources for opinion mining. In Andrea Sansò, editor, *Language resources and linguistic theory: Typology, second language acquisition, English linguistics*. Franco Angeli, Milano, IT.

A. Esuli and F. Sebastiani. 2006. SentiWordNet: A publicly available resource for opinion mining. In *Proceeding of the 6th International Conference on Language Resources and Evaluation*, Italy, May.

S.-M. Kim and E. Hovy. 2006. Extracting opinions, opinion holders, and topics expressed in online news media text. In *Proceedings of the ACL Workshop on Sentiment and Subjectivity in Text*.

T. Landauer and S. Dumais. 1997. A solution to plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review*, 104:211–240.

W. Lin, R. Yangarber, and R. Grishman. 2003. Bootstrapped learning of semantic classes from positive and negative examples. In *Proceedings of the ICML-2003 Workshop on The Continuum from Labeled to Unlabeled Data, Washington DC*.

R. Mihalcea, C. Banea, and J. Wiebe. 2007. Learning multilingual subjective language via cross-lingual projections. In *Proceedings of ACL*.

E. Riloff and J. Wiebe. 2003. Learning extraction patterns for subjective expressions. In *Proceeding of the Conference on Empirical Methods in Natural Language Processing*.

P.J. Stone, D.C. Dumphy, M.S. Smith, and D.M. Ogilvie. 1966. The general inquirer: a computer approach to content analysis. *M.I.T. studies in comparative politics, M.I.T. Press, Cambridge, MA*.

C. Strapparava and A. Valitutti. 2004. WordNet-Affect: an affective extension of wordnet. In *Proceeding of the 4th International Conference on Language Resources and Evaluation*, pages 1083–1086, Lisbon, Portugal, May.

H. Tanev, V. Zavarella, J. Linge, M. Kabadjov, J. Piskorski, M. Atkinson, and R.Steinberger. 2010. Exploiting machine learning techniques to build an event extraction system for portuguese and spanish. *Linguamatica: Revista para o Processamento Automatico das Linguas Ibericas*.

X. Wan. 2008. Co-training for cross-lingual sentiment classification. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing*.

T. Wilson, J. Wiebe, and P. Hoffman. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of HLT-EMNLP*.