Multilingual Entity-Centered Sentiment Analysis Evaluated by Parallel Corpora

Josef Steinberger and **Polina Lenkova** and **Mijail Kabadjov**
Ralf Steinberger and **Erik van der Goot**
EC Joint Research Centre
21027, Ispra (VA), Italy
Josef.Steinberger, Polina.Lenkova, Ralf.Steinberger
Mijail.Kabadjov, Erik.van-der-Goot
@jrc.ec.europa.eu

Abstract

We propose the creation and use of a multilingual parallel news corpus annotated with opinion towards entities, produced by projecting sentiment annotation from one language to several others. The objective is to save annotation time for development and evaluation purposes, and to guarantee comparability of opinion mining evaluation results across languages. By creating this resource, we answered the question whether sentiment is consistently translated across languages so that projection can actually be an option. We describe our approach to multilingual sentiment analysis and show its performance in 7 languages of the parallel corpus.

1 Introduction

In sentiment analysis the goal is to detect and classify subjective content of a text. The text can be classified as a whole such as in product reviews, in which an overall judgment is assigned to the product. If we move to the news domain, the overall sentiment score of an article can be used for detecting bad or good news. It can be used also for detecting the changes in sentiment in a particular topic. However, if the goal is to detect sentiment expressed towards entities, the aggregated sentiment of the articles, in which the entity appears, need not to correspond to opinions expressed towards the entity. The entity can be mentioned positively in a very negative article. We have to go down and analyze each entity mention based on the surrounding context.

Solving the problem in multilingual environment and gathering large amounts of articles from many sources give advantage to detect news opinions expressed in different countries towards same persons. Also, it eliminates the biased news. However, multilinguality brings another challenge. For instance, it is not easy to develop NLP tools like parsers or taggers in many languages, also using them can cause computational problems when applied on large amounts of articles every day. Another difficulty comes with resources. Sentiment-annotated data are not usually available for other types of texts then reviews, 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.

We addressed the resource bottleneck for sentiment dictionaries, by developing highly multilingual and comparable sentiment dictionaries having similar sizes and based on a common specification (Steinberger et al., 2011).

Our sentiment system is simply based on counting subjective terms around entity mentions (mainly persons and organizations). Evaluating its performance in more languages would multiply the annotation efforts. In this paper we propose using parallel corpora to automatically project annotations from English. We study the subjectivity of the entity-centered sentiment annotation and evaluate our sentiment system in seven languages (English, Spanish, French, German, Czech, Italian and Hungarian). As a side effect this evaluation serves as a task-based evaluation of the quality of the sentiment dictionaries.

Firstly, we discuss related work in Section 2. Next, we shortly mention the development of sentiment dictionaries and briefly discuss our sentiment system (Section 3). Then we focus on the annotation of the parallel corpus in Section 4. We show the figures of inter-annotator agreement. Before we conclude all, we discuss evaluation results of our system run on the parallel corpus (Section 5).

Proceedings of Recent Advances in Natural Language Processing, pages 770–775,
Hissar, Bulgaria, 12-14 September 2011.

2 Related work

The substantial growth in subjective information on the world wide web in the past years has made sentiment analysis a task on which constantly growing efforts have been concentrated. Subjectivity in natural language refers to aspects of language used to express opinions, evaluations, and speculations (Wiebe et al., 2005). To classify statements (as traditionally to positive, neutral (objective) and negative) is not a trivial task, as many expressions carry in themselves a certain subjectivity and many expressions are used both in a subjective (even both positive and negative), as well as objective manner.

Sentiment analysis has been done at a document level, the most often for review texts, starting from the assumption that each document focuses on a single object and contains opinion from a single opinion holder. There were numerous approaches dealing with document level sentiment classification (Pang et al., 2002; Dave et al., 2003). The approaches are usually evaluated by comparing the outcome of the analysis against the number of stars given to the review.

The document level assumptions do not hold for newspaper articles or blog posts where each sentence expresses one single opinion (sentence level approaches) about a target. (Hatzivassiloglou and Wiebe, 2000; Wiebe and Mihalcea, 2006; Wilson et al., 2004) use subjectivity analysis to detect sentences from which patterns can be deduced for sentiment analysis, based on a subjectivity lexicon. Kim and Hovy (2004) try to find, given a certain topic, the positive, negative and neutral sentiments express on it and the source of opinions (the opinion holder). The authors computed the sentiment of the sentence in a window of different sizes around target.

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. Another approaches in obtaining subjectivity lexicons for other languages than English were explored in Banea et al. (2008) or Wan (2008).

In the effort to guarantee comparability of results across languages, various authors have suggested using multilingual parallel corpora. For instance, Koehn (2002) used the multilingual parallel corpus EuroParl to evaluate Machine Translation performance across language pairs. Zaanen et al. (2004) propose to use a multilingual parallel parsed corpus as the best and fairest gold standard for grammatical inference evaluation, because parallel documents can be assumed to have the same degree of language complexity. Turchi et al. (2010) use parallel corpora for the evaluation of multilingual multi-document summarisation, in which the annotation is very expensive. It also makes the evaluation results across languages directly comparable.

3 Method for multilingual sentiment analysis

The objective of this paper is to focus on the creation and use of the sentiment-annotated parallel corpus. Our sentiment analysis tools will therefore only be described briefly. Any other sentiment analysis tool could be applied to this parallel corpus instead.

Our objective is to detect positive or negative opinions expressed towards entities in the news across different languages and to follow trends over time. Entities of interest are mostly persons and organisations, but also concepts such as the '7th Framework Program' or 'European Constitution'. Entities can be mentioned positively in negative news context, and vice versa, so that document level analysis is not sufficient (Balahur et al., 2010), but opinions expressed towards the specific entity mention must be detected. As we do not have access to parsers or even part-of-speech taggers for the range of languages we intend to analyse, we chose to use an extremely simple method that does not require language-specific tools besides NER software and language-specific sentiment dictionaries: we add up positive and negative sentiment scores in six-word windows around the entities, distinguishing two positive and two negative levels of sentiment words (having values of -4, -2, 2 and 4 points, respectively). Enhancers and diminishers add or remove 1 point, negation inverts the value, except for negated high positive ('not very good' is not equivalent to 'very bad').

The sentiment dictionaries – currently available in 15 languages – were created using a triangulation method, which was described in detail in (Steinberger et al., 2011). In a nutshell: carefully elaborated English and Spanish sentiment word lists were translated into third languages. The introduction of errors through word sense ambiguity was limited by taking the intersection of both target language word lists. According to our evaluation, approximately 90% of these intersection words were correct, while only about 50% of those words were correct that were translations from either English or Spanish, but not from both. For Arabic, Czech, French, German, Italian and Russian, these word lists were manually checked and enhanced, while for Bulgarian, Dutch, Hungarian, Polish, Portuguese, Slovak and Turkish we simply used the intersecting word list. For a subset of languages (Czech, English and Russian), wild cards were manually added to the sentiment word lists in order to capture morphological variants. For the other languages, the same will be done in the future. The results in section 5 differ heavily depending on whether morphological variants are dealt with.

4 Building a sentiment-annotated parallel corpus

In this section we give details about the parallel corpus, sentiment annotation and inter-annotator agreement.

4.1 Named entity-annotated parallel corpus

We worked with data from Workshops on Statistical Machine Translation (2008, 2009, 2010)[1] which provide parallel corpora of news stories in 7 European languages: English, Spanish, French, German, Czech, Italian (only 2009) and Hungarian (only 2008 and 2009). Putting together the data from the three years resulted in 7 065 parallel sentences in five languages, and a subset in Italian and Hungarian. We ran our in-house entity recognition on the data. Only known entities (entities present in our database) were marked in the data. It gave us enough samples to run sentiment experiments although guessing other entities (and considering coreference mentions) would considerably increase the pool of samples. For English we received 1 274 entity mentions, resulting in the same number of sentence-target (S-T) pairs for testing sentiment analysis. We built golden standard annotations and projected them to other languages. Because of different performance of entity recognition we obtained fewer S-T pairs in other languages than in English.

4.2 Sentiment annotation and inter-annotator agreement

Annotating sentiment in news is clearly a subjective task. Even the same person can assign different values to the same entity mention when reviewing it in a different time. Also, we were not sure whether there is the same sentiment in all language variants of the same sentence. If so we could project it automatically after annotating the S-T pairs only in one language. We had two annotators to judge the cases. The first one, native Russian speaker with advanced knowledge of English and Italian, and the second one, a native Czech speaker with advanced knowledge of English. Each of the annotators were asked to judge randomly-ordered S-T pairs. The first annotator in both English and Italian languages and the second one in both English and Czech languages. We could thus measure the agreement on how the same annotator judged the same sentences in different languages at a different time. Also, we could see the agreement between the annotators. The results in Table 1 show that there were cases considered differently by the same annotator while reviewing them in different languages. When analyzing the disagreed cases we have not found any example in which the reason of attaching different polarity would be that the sentiment was not correctly translated with the sentence. The agreement was 87%, resp. 90%, far above random agreement which results in high Kappa. We measured 80% agreement between annotators with fair Kappa (0.65). The first annotator assigned POS to 16% of the cases, NEG to 17% and NEUT to 67%. The second one assigned non-neutral polarity more often: POS – 26%, NEG – 24% and NEUT – 50%.

Because we wanted to obtain golden standard annotations the disagreed cases were judged by the third (super-)annotator.

Many controversial cases are related to the sentences where both positive and negative sentiments are expressed. Below we present three different examples (target is in bold) of such cases and our suggestions on how to deal with them.

[1] http://www.statmt.org/wmt10/translation-task.html.

1st annot.	A1-English	A2-English	A1-English
2nd annot.	**A1-Italian**	**A2-Czech**	**A2-English**
ALL	0.87	0.90	0.80
POS	0.78	0.81	0.78
NEUT	0.91	0.94	0.86
NEG	0.78	0.87	0.79
POS/NEG	0.78	0.83	0.78
Random	0.54	0.48	0.42
Kappa	0.72	0.81	0.65

Table 1: Inter-annotator agreement. A1/A2 = Annotator 1/2.

1. Positive and negative aspects of an event/entity

   Britain's building societies could face a bill of more than 80m after the rescue of the ***Bradford & Bingley bank****.*

   The above statement seems to be quite balanced, in the sense it presents both negative and positive characteristics, which do not contradict one another. Following our guidelines, POS/NEG cases are considered to be neutral.

2. Polarized opinions about the same entity:

   According to Russian observers, the reasons for this are the welfare and stability in the country led by ***Alexander Lukashenko****, while Organization for Security and Co-operation in Europe (OSCE) explains it as vote counting frauds.*

   This sentence might seem a bit more controversial than the previous one, as the author presents two different opinions, and we could expect that he supports one of them. By examining this sentence in isolation, we cannot say which side the journalist takes. Therefore we mark it as a neutral statement.

3. One sentiment value is stronger than the other. As an illustration consider the following example:

   It's almost funny to see how ***Barack Obama****, reputedly the wisest president, is trying so hard in the matter of the Afghan war to repeat the strategy of his predecessor, having himself considered him to be the most foolish.*

   In this sentence, we have a reference to Barack Obama as the wisest president, which is obviously a positive statement about him. On the other hand, the journalist claims that the president tries to follow his predecessor, whom he strongly criticizes, which reveals a stark inconsistency in the president's policy. Therefore the overall sentiment about him is negative.

4. Sometimes, sentiment towards one entity implicates the same sentiment towards another entity

   "We are satisfied with what we have reached during the night and we highly appreciate the efforts of the two parties in order to stabilize our financial markets and protect our economy", declared ***Tony Fratto, spokesman of the White House****.*

   The sentence describes an achievement reached in the White House, which positively characterizes the entity, but also its speaker Tony Fratto, as being representative of the White House.

   In the example below, there is a positive sentiment expressed towards Krugman, and since this positive sentiment is linked to the fact that he is a leader writer of New York Times, we conclude that New York Times as well bears positive characteristics.

   55 year old Krugman is a neo-Keynesian that teaches at Princeton University and he is a well-known leader writer of the ***New York Times****.*

5. Another, probably less obvious example of the sentiment transferred from one entity to another:

   A new case of positive testing during the last ***Tour de France****: it is the Austrian Bernhard Kohl, of the team Gerolsteiner, third in classification and winner of the best grimpeur shirt.*

   It is evident that positive testing characterizes negatively Bernhard Kohl, but also brings a bad reputation to the Tour de France, which has been affected by a few cases of positive testing.

6. There are also cases where we are unable to correctly detect sentiment without using world knowledge:

However, in spite of all these arguments, the winning trumf for the Democrats is ***George Bush****.*

The sentence sounds positively, however, cosidering the fact that George Bush is Republican inverts the polarity.

5 Evaluation

We projected the sentiment polarities in golden standard data to other languages and we ran the sentiment system. Table 2 compares the system results for each language with Random baseline. Another baseline is when all cases are attached to the most frequent neutral class (All NEUT), even if this baseline is not that valuable (no sentiment analysis at all). We can see that the overall agreement with golden standard was from 66% (Italian) to 74% (English and Czech). The best two performing languages are the ones with all steps of dictionary creation finished. In all languages the system performed better than the Random baseline and on the same level as the ALL NEUT baseline. Kappa shows the difference to random agreement. It uncovers the poorest performing language - Hungarian, for which we currently have only raw triangulated dictionaries. Thus this evaluation can serve as a task-based evaluation of the quality of sentiment dictionaries: best performing English and Czech (the most advanced dictionaries) are followed by French, Italian, German and Spanish, in which the lack of all morphological variants results in lower recall, with Hungarian at the end of the list. The cases on which the system fails to capture the right polarity can be found in the previous section. Consider the subjective terms like *rescue*, *positive testing* or *winning trumf.*

Another observation is that the system performs better on negative statements than on positive ones. We think that the reason is that the gap between the negative and the neutral class is larger than the gap between the positive and the neutral class.

The per-case sentiment assignment works at the 70% level. However, it goes down if we do not consider the neutral cases - around 50%. And this is exactly what we are interested in and these are the cases that we are going to summarise and show in the news monitoring system. The question is: Is this performance good enough to assess sentiment expressed in news towards an entity? We try to answer it by the following experiment. We gathered all mentions of an entity in the corpus, emulating the time period. We computed how many times the entity was mentioned positively and negative in the golden standard. The difference would be its aggregated score (e.g. -2 means there were two more negative mentions than positive). We do the same with the system annotations. If both the golden standard and the system attached the same polarity to the entity we consider it as a correct answer. Because we process large amounts of articles every day, precision is more important than recall. Also, aggregated values close to zero are the most dangerous. One mistake in polarity assignment can invert the polarity of the whole entity within the time period. Thus, we experimented with different thresholds. For example threshold 2 means that we need the aggregated value to be at least 2 to consider the entity positive, resp. -2 to consider it negative. We report only the cases in which both the system the golden standard reported a non-neutral value to remove the borderline unreliable cases (Table 3). We can observe that with the basic threshold (1) we correctly classified 78% entities and by lifting the threshold up to 2 the system reached the performance of 93%. The only wrongly classified entity for English was *al-Qaeda*. While annotators assigned to this entity clearly negative overall sentiment (-5), many difficult cases led the system to a positive overall sentiment (+2). We did not find any wrong case in which the system did not agree on polarity with the golden standard with threshold 3. Testing higher thresholds would require analyzing a larger set.

Threshold	**1**	**2**	**3**
English	0.80 (102)	0.88 (8)	1.00 (3)
Spanish	0.58 (26)	0.75 (4)	1.00 (1)
French	0.85 (41)	1.00 (5)	1.00 (2)
German	0.75 (32)	1.00 (4)	—
Czech	0.88 (24)	1.00 (3)	1.00 (1)
Italian	0.76 (21)	1.00 (2)	1.00 (2)
Hungarian	0.75 (12)	1.00 (1)	—
Total	0.78 (258)	0.93 (27)	1.00 (9)

Table 3: Precision of aggregated sentiment for each entity across the corpus for three different thresholds which divide POS/NEG sentiment from NEUT. *precision* (*No. of entities*)

6 Conclusion

We presented the extensive evaluation of our multilingual sentiment analysis system. We con-

Language	English	Spanish	French	German	Czech	Italian	Hungarian
ALL	0.74	0.71	0.72	0.70	0.74	0.66	0.68
POS	0.44 / 0.32	0.31 / 0.08	0.43 / 0.12	0.32 / 0.10	0.48 / 0.12	0.34 / 0.18	0.38 / 0.07
NEUT	0.79 / 0.90	0.73 / 0.96	0.74 / 0.96	0.72 / 0.96	0.76 / 0.95	0.70 / 0.90	0.70 / 0.96
NEG	0.58 / 0.31	0.57 / 0.10	0.62 / 0.18	0.70 / 0.10	0.57 / 0.23	0.56 / 0.19	0.36 / 0.06
POS/NEG	0.50 / 0.31	0.43 / 0.09	0.53 / 0.15	0.45 / 0.10	0.53 / 0.18	0.44 / 0.18	0.37 / 0.06
ALL NEUT	0.72	0.71	0.71	0.70	0.73	0.66	0.69
Random	0.62	0.67	0.66	0.66	0.68	0.59	0.66
Kappa	0.31	0.11	0.17	0.12	0.20	0.17	0.05

Table 2: System's results on the parallel corpus. The cells that correspond to POS, NEUT, NEG and POS/NEG rows contain precision/recall figures.

tributed to resources of the sentiment community by building the multilingual sentiment dictionaries and annotating the parallel corpus. Working on parallel data enabled to evaluate such a system in many languages with a little annotation effort and, also, the results are comparable across the languages. The evaluation also serves as a task-based evaluation for sentiment dictionaries.

Our system is language-independent, although it needs to be fed by sentiment dictionaries for each language. So far, we created dictionaries for 15 languages with varied quality, however, we have capabilities to further improve the resources. The final goal is to feed the output of the sentiment analysis into the news monitoring system in all the 50 languages it supports.

Even if discovering the right polarity of sentiment towards an entity in a sentence is a difficult task and the system's results for non-neutral cases are modest, per-entity sentiment aggregation leads to precise conclusions when used carefully.

References

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*.

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

K. Dave, S. Lawrence, and D. Pennock. 2003. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceeding of WWW*.

V. Hatzivassiloglou and J. Wiebe. 2000. Effects of adjective orientation and gradability on sentence subjectivity. In *Proceedings of COLING*.

S.M. Kim and E. Hovy. 2004. Determining the sentiment of opinions. In *Proceedings of COLING*.

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*.

P. Koehn. 2002. Europarl: A multilingual corpus for evaluation of machine translation. In *Unpublished draft*.

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

B. Pang, L. Lee, and S. Vaithyanathan. 2002. Thumbs up? sentiment classification using machine learning techniques. In *Proceeding of EMNLP*.

J. Steinberger, P. Lenkova, M. Ebrahim, M. Ehrman, A. Hurriyetoglu, M. Kabadjov, R. Steinberger, H. Tanev, V. Zavarella, and S. Vazquez. 2011. Creating sentiment dictionaries via triangulation. In *Proceedings of the ACL's WASSA Workshop*.

M. Turchi, J. Steinberger, M. Kabadjov, and R. Steinberger. 2010. Using parallel corpora for multilingual (multi-document) summarisation evaluation. In *Proceedings of CLEF*.

X. Wan. 2008. Co-training for cross-lingual sentiment classification. In *Proceedings of ACL*.

J. Wiebe and R. Mihalcea. 2006. Word sense and subjectivity. In *Proceedings of COLING-ACL*.

J. Wiebe, T. Wilson, and C. Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2-3).

T. Wilson, J. Wiebe, and R. Hwa. 2004. Just how mad are you? finding strong and weak opinion clauses. In *Proceeding of AI*.

M. Van Zaanen, A. Roberts, and E. Atwell. 2004. A multilingual parallel parsed corpus as gold standard for grammatical inference evaluation. In *Proceedings of The Amazing Utility of Parallel and Comparable Corpora Workshop*.