The Challenges of Multi-dimensional Sentiment Analysis Across Languages

Emily Öhman and **Timo Honkela** and **Jörg Tiedemann**
University of Helsinki
firstname.lastname@helsinki.fi

Abstract

This paper outlines a pilot study on multi-dimensional and multilingual sentiment analysis of social media content. We use parallel corpora of movie subtitles as a proxy for colloquial language in social media channels and a multilingual emotion lexicon for fine-grained sentiment analyses. Parallel data sets make it possible to study the preservation of sentiments and emotions in translation and our assessment reveals that the lexical approach shows great inter-language agreement. However, our manual evaluation also suggests that the use of purely lexical methods is limited and further studies are necessary to pinpoint the cross-lingual differences and to develop better sentiment classifiers.

1 Introduction

Typically, sentiment analysis is modeled as a three-class classification task, marking utterances as either positive, negative or neutral. In some cases, this may be accompanied with a degree of polarity. However, that still treats the task as a one-dimensional one along the scale of general polarity. In this paper, we look at the challenge of a multi-dimensional approach in which we aim at a much more fine-grained classifications with eight distinct dimensions of emotion in addition to the classical sentiments of positive and negative polarity. These emotions are based on Plutchik's wheel of emotions (see Figure 1): anger, anticipation, disgust, fear, joy, sadness, surprise, and trust (Plutchik, 1980). Detecting fine-grained sentiment is important for practical applications as well as for theoretical reasons. In the context of social media, it is useful to know whether someone is, for instance, happy, angry or sad, rather than relying solely on positive or negative sentiments. This can be applied, for instance, for the detection of hate-speech or depression and can be used to monitor peoples well-being or social dynamics.

As sentiment analysis methods are often developed for English first and other languages second, it is necessary to know whether it is possible to transfer tools and resources from English to other languages to speed up the coverage of the linguistic diversity in the World. With the growing importance of social media in societal issues, as a marketing tool, opinion generator, and so forth, it is essential to be able to accurately classify sentiments and emotions also for languages other than English. For those reasons, we, therefore, focus on cross-lingual methods and multi-dimensional settings.

Previous work has focused on lexical approaches using indicator word lists that define cues for detecting certain types of sentiment. In our work, we are interested in studying the effectiveness of these purely lexical approaches and we emphasize their use across languages. We have previously conducted research on multidimensional sentiment analysis (Honkela et al., 2014) but not across language borders. Multilingual studies for conventional sentiment analysis have been done, e.g., for English and German by Denecke (2008) but not with the fine-grained multidimensional analysis. Other related studies using Plutchik's eight emotions are for example the Rule-based Emission Model by Tromp and Pechenizkiy (2014) and EmoTwitter which takes advantage of the NRC Word-Emotion Association Lexicon to produce visualizations for identifying enduring sentiments in tweets (Munezero et al., 2015).

Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media, pages 138–142, Osaka, Japan, December 12 2016.

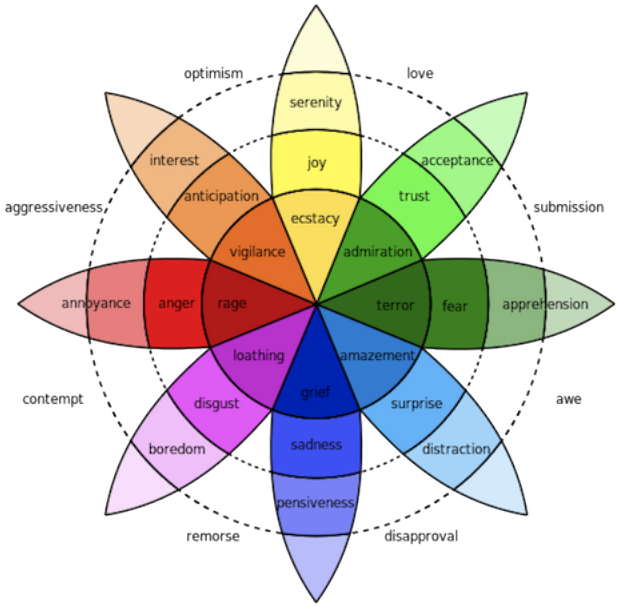

Figure 1: Plutchik's wheel of emotions [1]

One of the main research question we ask in our study is whether fine-grained sentiment and emotions are preserved across languages. Studies directly evaluating the preservation of sentiments in translation have often focused on comparing them with other methods such as whether it is better to translate the original text to English and analyze the English sentiments or to translate the lexicon from English to the "original" target language (work on Arabic (Salameh et al., 2015) and Chinese (Wan, 2008)). One study found that connotations change if texts are machine translated or manually translated and suggested that "further cross-lingual studies should not use parallel corpora to project annotations blindly" (Carpuat, 2015).

Related work does not provide a full picture of sentiment preservation in translation and we are interested in additional investigations with other data sets and setups. In particular, we would like to understand more clearly how sentiment preservation applies to the multidimensional task and whether there are differences between cases of similar versus less-related languages. For this purpose, we use lexicon-based methods and parallel data sets as a proxy for multilingual sentiment analyses on comparable texts. We also test the reliability of the purely lexical sentiment detection strategy using a small-scale manual evaluation.

The essential research questions we would like to ask are, hence, the following:

- To what extent is fine-grained sentiment preserved in translation? Are there differences between languages and their cultural embeddings?
- How reliable are purely lexical approaches in detecting multi-dimensional sentiments and emotions across languages?

To address the first question, we performed a small scale manual evaluation of movie subtitles to measure the correlation between detected sentiments in aligned subtitles. Using this set of manually classified utterances we then estimate the expected preservation of sentiment across specific language pairs. Finally, using those expectations we can measure the correlation with the automatic classification based on lexical look-up across languages to address our second question.

In the following, we first briefly describe the data sets and resources used in our study. Thereafter, we describe the manual evaluation of sentiment across languages and, finally, we discuss the results of

[1] Source: `https://en.wikipedia.org/wiki/Contrasting_and_categorization_of_emotions`

automatic multi-dimensional sentiment classification based on an existing lexical resource. We conclude with our main findings and prospects for future research.

2 Multilingual Data Resources

For our cross-lingual experiments we rely on publicly available parallel data sets. OPUS[2] provides large quantities of sentence-aligned multilingual corpora including a comprehensive collection of movie subtitles in various languages (Lison and Tiedemann, 2016). Movies certainly contain a lot of emotional contents and their predominantly colloquial style makes them a good proxy for social media data we aim at with our multidimensional sentiment analyzer.

As a comparison data set, we selected the Europarl parallel corpus (Tiedemann, 2012). Europarl represents a different genre and the translations come from professional sources, whereas the subtitle translations contain a much larger quantity of noise (due to the unreliability of user-generated / user-provided content, incomplete data sets as well as conversion and alignment errors). We are thus able to compare two different quality-levels of translation as well besides the comparison of two dissimilar genres. In both cases, we used 1.5 million lines of aligned sentences from the parallel corpora for each language, which we lemmatized using the Turku Finnish Dependency Parser for Finnish (Ginter et al., 2013) and UDPipe (Straka et al., 2016) for all other languages.

The emotion lexicon we apply is called the NRC Word-Emotion Association Lexicon (Mohammad and Turney, 2013). The lexicon is a list of originally English words and their crowd-sourced associations with Plutchik's eight basic emotions and two sentiments (Plutchik, 1980). The words have been translated by the creators of the lexicon using Google Translate. The number of annotated words per language vary between 4,043 and 14,182. We also translated the remaining words for the target languages in the same way bringing the total number of annotated words to 14,182 for all languages. The translation results were checked and, for the target languages, no clear translation errors were found making the full lexicon at least as good as the original version.

3 Multilingual Fine-Grained Sentiment Classification

In the following, we look at a purely lexicon-based approach to fine-grained sentiment analyses using the multilingual emotion lexicon presented in the previous section. For this, all lines in our data set are matched one-by-one with the items in the lexicon. The result of this process is a 10-dimensional vector for each line containing the counts of matched words that represent the sentiment or emotion of that particular dimension according to the lexicon. We can interpret the vectors in two different ways: (i) Any non-zero count indicates the presence of the sentiment in question (binarized interpretation), or, (ii) the counts represent the prevalence of the corresponding sentiments and emotions.

We can now measure the cross-lingual correlation between the sentiments detected by the lexicon-based approach by comparing the vectors created for each of the 1.5 million lines in each translation. We do this for both, the subtitle corpus and the Europarl corpus by means of individual emotions and sentiments and by means of a multidimensional comparison. For the former we apply the binarized interpretation and compute the percentage of matching sentiments detected across language borders. Table 1 lists the scores for each test case. Note that we discard all zero-score matches where no sentiment was detected in either language. This applies to the majority of lines and, therefore, would blur the picture.

The scores in the table show that for English-Finnish the subtitle data is more likely to match across languages than the Europarl data. For all the other pairs, this trend is reversed for all emotions and sentiments.

The most common emotions in the texts were the same for all languages: negative, positive, then fairly similar for trust, disgust, anger, fear, joy, sadness, and generally much lower for anticipation and surprise. This is most likely related to the higher cross-language agreement for these emotions: the more common a sentiment, the more chances of one language detecting it but it being missed by the other and therefore decreasing the cross-language agreement score.

[2] http://opus.lingfil.uu.se

	Emotion / Sentiment										
	pos.:	neg.:	anger:	anticip.:	disg.:	fear:	joy:	sad.:	surpr.:	trust:	ALL
Language	Movie Subtitles										
EN-FI	.6051	.4535	.7507	.8299	.7761	.7818	.8364	.7766	.8964	.7471	$0.752^{\pm 0.232}$
EN-SV	.5709	.4744	.7897	.8310	.7817	.7948	.7710	.7865	.8922	.7631	$0.802^{\pm 0.220}$
ES-PT	.6186	.4912	.7964	.8419	.8119	.7715	.8749	.7299	.9248	.8251	$0.746^{\pm 0.231}$
Language	EuroParl										
EN-FI	.5670	.4613	.7733	.8138	.7839	.7805	.8240	.7755	.8914	.7434	$0.788^{\pm 0.241}$
EN-SV	.3219	.4028	.7148	.6590	.7420	.6605	.7314	.6902	.7888	.4851	$0.665^{\pm 0.213}$
ES-PT	.4172	.4480	.6849	.6934	.7815	.6783	.7352	.6501	.8278	.5570	$0.692^{\pm 0.178}$
AVG:	.5168	.4552	.7516	.7782	.7795	.7446	.7955	.7348	.8702	.6868	

Table 1: Percentage of matched sentiments across languages according to lexicon-based classification. *ALL* refers to the averaged cosine similarity of the 10-dimensional sentiment vectors and the number in superscript gives the standard deviation observed in the data.

The cosine similarity scores indicate that the Finnish and English vectors are most dissimilar, with only slightly higher similarity scores for the English-Swedish pair. The Spanish-Portuguese scores, however, show higher similarity scores than either of the other two languages. One is tempted to conclude that this illustrates the cultural influences that determine the expressions of sentiments and emotions but we have to take these preliminary results with a grain of salt also based on the manual evaluation presented below, which indicates that purely lexicon-based methods are not reliable enough.

4 Manual Evaluation

In order to test the reliability of the lexicon-based method, we conducted a small scale manual evaluation on the same data set. For this, we randomly selected 100 lines of the aligned texts and annotated them by hand using Plutchik's eight emotions as well as their positive and negative sentiments.

	pos.:	neg.:	anger:	antic.:	disgust:	fear:	joy:	sad.:	surpr.:	trust:	AVG	COS
EN-FI	.923	.846	1.000	.897	.821	.923	1.000	.949	.872	.897	.913	.983
EN-SV	.909	.848	.970	.909	.788	.970	.970	.939	1.000	1.000	.930	.976

Table 2: Hand-annotation sentiment agreement across languages

Each line corresponds to one or more sentences from within a translation unit and we also considered previous and subsequent context for deciding proper classifications. We restricted ourselves to binary choices when marking one or more of the ten dimensions. Using scales for such human annotation would be an interesting extension that we would like to explore in future work. Each line was classified by two annotators and a third annotator was consulted in case of disagreement between the two.

As Table 2 shows, the manual annotation reveals that cross-language agreement is high for both language pairs and all emotions and sentiments. Using the manual annotation as gold standard we then computed precision and recall of the automatic classification. To our surprise (especially with respect to precision), both mesaures are extremely low (below 10%) for all emotions and sentiments. However, this may be caused due to the overall scarcity of emotions and our little data set in general. In order to understand better the true precision and recall of the automatic classification as compared to the hand-annotated data-set it would, of course, be highly beneficial to have a larger sample of hand-annotated data.

5 Conclusions and Discussions

The study of multilingual social media corpora is important as it provides, for instance, a possibility to compare how people in different parts of the world view various topics.

There is clearly a use for good lexicons in sentiment analysis. The extent to which these are utilized and the quality of the lexicon, especially if translated, is what influences the cross-language agreement ratings the most.

As the results show, a purely lexicon-based approach can tell us about the sentiments and emotions in a text, but that it is not as good as a gold standard. In this pilot study we can see that Spanish and Portuguese have higher cross-language agreement than English and Finnish, or English and Swedish. In the future it would be interesting to compare languages that are culturally more different such as English and Chinese, or English and Arabic or Japanese. This might reveal a clearer picture about the influence of cultural backgrounds on the expressions of emotions and sentiments in comparable texts.

With respect to our second research question, we are at this stage interested in how well lexical approaches are capable of detecting multidimensional sentiments using parallel data as a proxy for evaluation. For this, we assume that sentiments and emotions are preserved in translation and we verify this with a small-scale manual annotation. These initial results will guide us in future work to enhance the detection approach with more sophisticated methods based on supervised and semi-supervised machine learning techniques including cross-lingual representations and transfer models.

References

Marine Carpuat. 2015. Connotation in Translation. In Alexandra Balahur, Erik van der Goot, Piek Vossen, and Andrés Montoyo, editors, *WASSA@EMNLP*, pages 9–15. The Association for Computer Linguistics.

Kerstin Denecke. 2008. Using SentiWordNet for multilingual sentiment analysis. In *ICDE Workshops*, pages 507–512. IEEE Computer Society.

Filip Ginter, Jenna Nyblom, Veronika Laippala, Samuel Kohonen, Katri Haverinen, Simo Vihjanen, and Tapio Salakoski. 2013. Building a Large Automatically Parsed Corpus of Finnish. In Stephan Oepen, Kristin Hagen, and Janne Bondi Johannessen, editors, *NODALIDA*, volume 85 of *Linköping Electronic Conference Proceedings*, pages 291–300. Linköping University Electronic Press.

Timo Honkela, Jaakko Korhonen, Krista Lagus, and Esa Saarinen. 2014. Five-Dimensional Sentiment Analysis of Corpora, Documents and Words. In *Advances in Self-Organizing Maps and Learning Vector Quantization - Proceedings of the 10th International Workshop, WSOM 2014*, pages 209–218.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk, and Stelios Piperidis, editors, *LREC*. European Language Resources Association (ELRA).

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a Word-Emotion Association Lexicon. 29(3):436–465.

Myriam Munezero, Calkin Suero Montero, Maxim Mozgovoy, and Erkki Sutinen. 2015. EmoTwitter - A Fine-Grained Visualization System for Identifying Enduring Sentiments in Tweets. In Alexander F. Gelbukh, editor, *CICLing (2)*, volume 9042 of *Lecture Notes in Computer Science*, pages 78–91. Springer.

Robert Plutchik. 1980. A general psychoevolutionary theory of emotion. *Theories of emotion*, 1:3–31.

Mohammad Salameh, Saif Mohammad, and Svetlana Kiritchenko. 2015. Sentiment after Translation: A Case-Study on Arabic Social Media Posts. In Rada Mihalcea, Joyce Yue Chai, and Anoop Sarkar, editors, *HLT-NAACL*, pages 767–777. The Association for Computational Linguistics.

Milan Straka, Jan Hajič, and Straková. 2016. UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, Paris, France, May. European Language Resources Association (ELRA).

Jorg Tiedemann. 2012. Parallel Data, Tools and Interfaces in OPUS. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Erik Tromp and Mykola Pechenizkiy. 2014. Rule-based Emotion Detection on Social Media: Putting Tweets on Plutchik's Wheel. *CoRR*, abs/1412.4682.

Xiaojun Wan. 2008. Using Bilingual Knowledge and Ensemble Techniques for Unsupervised Chinese Sentiment Analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 553–561, Stroudsburg, PA, USA. Association for Computational Linguistics.