# Emotion Preservation in Translation: Evaluating Datasets for Annotation Projection

Kaisla Kajava[1], Emily Öhman[1], Piao Hui[2], and Jörg Tiedemann[1]

University of Helsinki, Finland
firstname.lastname@helsinki.fi
The University of Tokyo, Japan
firstnamelastname@p.u-tokyo.ac.jp

**Abstract.** This paper is a pilot study that aims to explore the viability of annotation projection from one language to another as well as to evaluate the multilingual data set we have created for emotion analysis. We study different language pairs based on parallel corpora for sentiment and emotion annotations and explore annotator agreement. We show that the data source is a possible one for reliable L1 data to be used in annotation projection from high-resource languages, such as English, into low-resource languages and that this is a reliable way of creating data sets for fine-grained sentiment analysis and emotion detection.

**Keywords:** Annotation Projection, Sentiment Analysis, Translation Studies.

## 1 Introduction

Most languages in the world are low-resource languages from a computational linguistics point of view [20, 21, 28], i.e. few data sets and tools exist that have been created specifically for those languages. Although this problem is being addressed more and more, the lack of resources is still leading to difficulties in creating tools such as customer-service chat bots, automatic speech recognition [1], but also services such as translation tools for hospitals.

The majority of sentiment analysis resources are for English, while many other languages suffer from the aforementioned scarce resources [18]. Cross-lingual text classification aims to apply resources in one language (L1) to another language (L2) [25]. In cross-lingual transfer learning for sentiment analysis annotations, sentiment-labeled resources in a resource-rich language L1, typically English, can be projected onto parallel unlabeled resources in a resource-poor language L2 and used to train a sentiment classifier. The result is a sentiment classifier for L2 (see e.g. [3, 17]).

However, one challenge of using translated data to train a sentiment classifier is assessing how well sentiment is preserved in translation. When a text is translated from a source language to a target language, the sentiment information of that source language may not be preserved correctly in the target language [22, 23]. This is mainly due to word choice and different languages coding emotions and sentiments differently. As a result, that text has a risk of not being representative for sentiment analysis in the target language. How well the sentiment information is preserved varies as we will show in subsequent sections.

”Translation can be defined as the result of a linguistic-textual operation in which a text in one language is re-contextualized in another language” [10]. We suggest that a ”good” translation should, at least implicitly, contain the sentiment and emotions present in the original text. If translated data is used as training data for sentiment analysis systems, it is important to assess whether the output is meaningful in that language, i.e. that it encodes the same sentiment and emotion information.

Another method to increase reliability is by weeding out low-effort annotations using a confidence score [32]. A translation that encodes a sentiment which the source text did not intend to encode is not a “good“ translation for the purpose of annotation projection and thus is not representative for sentiment analysis in a target language.

Analogously, if sentiment is preserved well in translation, translated data is a viable option for training sentiment analysis systems. In this scenario, if data compiled by annotation projection produces good classification results, cross-lingual transfer learning by annotation projection is a useful way of creating sentiment analysis data for low-resource languages. Therefore, we first investigate the preservation of sentiments and emotions in translation by manually annotating parallel data sets in English, Finnish, Italian, and French. As a side effect those data sets also constitute the gold standard test sets that we need to evaluate annotation projection approaches to cross-lingual sentiment classification. For the gold standard test set we do not do any projections from one language to another. Instead we evaluate how well this data set is suited for projection by comparing how different languages encode emotion.

In particular, we apply parallel corpora of movie subtitles as cross-lingual resources from L1 to L2 to test the feasibility of projection. A parallel corpus with aligned translations in multiple languages allows for annotation labels to be directly projected from L1 to L2. Banea et al. [2] refer to this method as constructing a direct ”bridge” between the two languages. In sentiment analysis and for the purpose of this study, parallel corpora are a useful resource for investigating how sentiment is encoded and preserved across languages, and how classifiers trained with cross-lingual transfer learning decode sentiment information.

In practice, when sentiment resources are translated from one language to another, there is not only a risk of losing sentiment information due to a poor translation but also due to the cultural encoding of sentiment. The risk of sentiment information being lost in translation applies both to machine translation and to manual translation. Thus, a translation may not be representative for sentiment analysis in the target language in the sense that it may encode a sentiment which the source text did not intend to encode. This is an issue for translated texts in general and cross-cultural texts specifically rather than related to sentiment analysis or the translation method used (i.e. machine translation vs. human translation).

If the sentiment classes of *positive*, *neutral*, and *negative* may change in translation, it seems reasonable to assume that multi-dimensional emotion classes (i.e. the 8 emotions listed by Plutchik [19]) may contain even more significant differences. A relevant question to ask is then to what degree multi-dimensional sentiment information is preserved in translation. When annotating translated text, how often do human annotators assign the same sentiment label to a translated text as to the original text? In machine classification, is the intended sentiment of the source language text preserved in the tar-

get language text when sentiment labels are projected from one text to another? These are some of the questions we hope to answer in the following pages.

## 2 Previous Work

Öhman et al. [30] used the NRC Word-Emotion Association Lexicon sentiment lexicon to investigate how multi-class sentiment information is preserved in translation. While the study found correlation between sentiment preservation across languages in a parallel data set, they concluded that lexical methods are not sufficient for preserving sentiment information cross-lingually.

How sentiment information is preserved in translated text is an important question in translation studies, in the study of cultural differences with respect to emotions and sentiment in text, and in research on annotation projection and cross-lingual transfer models for sentiment detection. This question is relevant both in terms of cultural differences when it comes to encoding sentiment in language and in terms of evaluating the quality of sentiment-labeled data sets produced by translation.

*Parallel texts should express similar sentiment*, was the assumption behind the study of Lu et al. [15] and Sarthak et al. [11]. Therefore, parallel corpora have been used for transfer learning. However, there are studies that show that emotions and sentiment are encoded differently in different languages, which is why one needs to study the preservation of sentiment in translation. As no previous studies were found that investigated the preservation of fine-grained sentiments and emotions with a multi-dimensional model, this study is trying to fill that gap.

Overall, the use of parallel corpora has been shown to be a viable option for cross-lingual transfer learning in sentiment analysis (see e.g. [11, 15, 18, 25]). This is evidenced by high inter-agreement between source and target language classification. However, no studies on cross-lingual sentiment preservation for multi-dimensional sentiment classes were found.

Machine translation methods for cross-lingual sentiment analysis have been studied and leveraged for many approaches. Machine translation has also been proposed as an alternative for adapting sentiment analysis to other languages with the assumption that automatic translation produces text with the same sentiment and emotions as the source. However, machine translation is trained on human translations and, therefore, it may suffer from the same problem, i.e. that those aspects are not preserved well. This is another reason we need this study to verify that we can apply such cross-lingual techniques with success. Ways of encoding sentiment in parallel text can give us valuable information on how different languages encode emotion and subjectivity, and to what extent these patterns are comparable across language boundaries. This is especially important to carefully consider as we know that emotions and sentiments are encoded differently in different languages, cultures, and time periods [12, 26, 27].

### 2.1 Sentiment Preservation in Translation

In previous work, the question of how sentiment is preserved in translation has to a large extent been studied with regards to machine translation and mainly as a binary or ternary classification problem.

With the assumption that sentiment may not be preserved correctly when translation quality is poor, Lohar et al. [14] trained a sentiment classifier to investigate both translation quality and sentiment preservation. They found that splitting the training data for the machine translation system into sentiment classes improved sentiment preservation in the target language.

It has been shown that errors made by machine translation systems are more likely to lead human annotators astray than to lower the performance of a classifier [18, 22]. Salameh et al. [22] found that automatic sentiment classification of machine-translated text reached a higher accuracy than manual sentiment annotation of the same text. However, they also found that some degree of sentiment information gets lost in translation, which results in a larger output of sentiment instances classified as neutral.

The cultural context of parallel texts is essential to consider with regards to sentiment preservation. Even when a translation is correct, the sentiment may be marked differently [18]. That is, the source language may encode a particular sentiment into a text but the target language may interpret that sentiment as a different one.

## 3 Data & Method

The data set used in this work was extracted from the OPUS movie subtitles corpus [13], an open source resource of parallel and aligned translations of movie subtitles. English was selected as the source language with three target languages: Finnish, French, and Italian. This yielded a data set consisting of original English sentences and their translations in the three aforementioned languages.

Although Finnish, French, and Italian are not typically considered low-resource languages, in a sentiment analysis context they are. For the purposes of this pilot study, i.e. to show that sentiment projection is a possible way to create sentiment analysis datasets for truly low-resource languages, it should not matter whether the languages used are truly low-resource or not as the projection is what is evaluated and should work reasonably similarly between any two languages. The languages used were chosen due to the availability of native or native-level speakers willing to annotate the datasets.

As very little information exists regarding the translation method of specific translations in the database, it is unclear whether the translation in question has been a professional one, an amateur one, or created by the use of machine translation and then fixed to the best of a human translator's ability, if fixed. The differences in translations can be seen in the *Opusparcus* paraphrase corpus [6] created from these same OPUS subtitles where the line has been translated from one language to another and then back again. "I thought so, too." & "That was my general impression as well." and "Have a seat." & "Sit down." are examples from *Opusparcus*. Although the phrases are very similar in meaning, it is clear that the way subtitles have been translated can significantly alter the perceived emotions conveyed by the phrase. These differences in translation methods affect our data set as well.

### 3.1 Data Validation and Preprocessing

The OPUS subtitle data set in English has been tested with preliminary results showing that the data itself works with classification tasks using simple multi-layer perceptron,

Naive Bayes, SVM, and MaxEnt classification frameworks. This was done by manually splitting the data into stratified training and testing sets and then performing classification tests on the data.

To extract meaningful data from the OPUS subtitles corpus, the corpus source files were first run through a filtering process. Similarly to Creutz [6], truly parallel one-to-one aligned sentences between English, Finnish, French, and Italian were extracted from the corpus. Documents which did not have English as the source language, sentences which did not appear in one or more of the languages, and sentences which were misaligned in one or more of the languages were filtered out. In addition, incorrectly rendered characters were manually corrected.

The extracted files were then searched for incorrect characters, typos, and previously undetected misaligned sentences. This phase in the preprocessing work additionally served as a scan for errors in the OPUS subtitles corpus as some patterns of mistakes in the corpus files where detected. The most notable of these patterns were how the character *l* was rendered as *I* (i.e. lower case 'L' and upper case 'i', respectively) in many Finnish language documents, the character *à* was rendered as *ø*, and the character *è* was rendered as *é* in many Italian language documents.

These observations can be used for improving the OPUS corpus compilation steps to reduce the amount of errors in the corpus files. The creators of the corpus have been made aware of these inconsistencies.

### 3.2 Data Set Distribution

The OPUS corpus provides metadata which allows each sentence and document to be traced back to its original source. The metadata for the extracted sentences was retained and used to examine the distribution of the data. This was useful for detecting any meaningful tendencies in the data, such as sentences assigned a specific label originating in movies of a specific genre, from a specific time, or from a specific country. The data in the final data set originates from 214 documents collected from the OPUS corpus, each document containing the subtitles of one movie.

**Table 1.** Number of sentences by sentiment class.

| ***pos*** | ***neg*** | *ang* | *ant* | *dis* | *fea* | *joy* | *sad* | *sur* | *tru* | **Total** |
|---|---|---|---|---|---|---|---|---|---|---|
| 2,857 | 3,570 | 908 | 823 | 862 | 685 | 975 | 705 | 649 | 820 | 6,427 |

Drama emerged as the most common source genre for all labels due primarily to the large amount of source documents in the corpus tagged as belonging to that genre. While some genres, such as musical, film-noir, and animation, were fairly uncommon in the data, other genres such as thriller, comedy, and romance seemed to exhibit some trends with regard to the distribution of each label. In movies tagged as thrillers, the predominant sentiment labels were *fear* and *trust*, while romance movies had a notable peak in *joy* and a fairly large amount of *anticipation* and *sadness*. Comedy proved to be the largest source of *joy*-labeled sentences after drama, and yet produced also a large

amount of *anger*, *disgust*, and *anticipation*. Considering that comedy was a common genre in the source documents, it is worth noticing that few sentences with the labels *fear*, *sadness*, *surprise*, and *trust* originated in that genre. This is fairly in line with the main theories on humor [16].

It should be noted that the sentiment data used in this study was primarily from an American context. Overall, the data consisted mostly of sentences from American movies dating from the 1970s and 1980s with the majority of those movies belonging to the genres of drama, comedy, romance, and crime. The distributional information on the OPUS data used in this work can also be used to study the correlation of genre and release time with sentiment information.

## 4 Annotation & Results

The English base data set was annotated using Sentimentator [31, 32], a web-based tool for sentiment annotation based on Plutchik's theory of emotion [19]. Sentimentator makes it possible to annotate on the sentence-level, ranging from binary and ternary annotation up to 48 more fine-grained classes which the end-user can select with an intensity slider. With the *positive*, *negative*, and *neutral* classes included, the number of possible distinct labels amounts to a total of 51.

The goal of the annotation process was to produce a data set that can be used for both binary and multi-class sentiment classification. For this purpose, the multi-class sentiment classes derived from Plutchik's core eight emotions were also treated as binary classes:

- ***Negative***: *anger*, *disgust*, *fear*, *sadness*
- ***Positive***: *anticipation*, *joy*, *trust*

*Surprise* not belonging intuitively to either binary class, sentences annotated with that label were divided into the *negative* and *positive* classes.

The annotated English data set was exported from Sentimentator and manually revised. At this stage, the revised data was annotated by a native English speaker. Sentences which the annotator considered ambiguous or neutral were removed from the data set. Sentences for which the most suitable label was disputed were revised and either assigned one of the given labels or excluded from the data set. This process yielded a final English sentiment data set based on Plutchik's core eight emotions.

Similarly, the parallel translated sentences in Finnish, French, and Italian were annotated by two native or native-level speakers of each language. Annotators were instructed to annotate the sentences so that a sentence should be assigned only one (1) label, a sentence which does not fit into any sentiment category should be skipped, and the label for each sentence should be chosen according to one's own judgment.

Annotators were informed that the sentences were presented in random order and that each sentence was to be considered an independent unit unrelated to the preceding or following sentence. The intuition behind guiding annotators to choose sentiment labels based on their own judgment was to emphasize that the goal of the task was not to find any universally correct label for a given sentence. Rather, the expression of sentiment in natural language is highly nuanced and open to interpretation. This

question of human interpretation of natural language is especially present when dealing with very short texts such as self-standing sentences [7, 24].

### 4.1 Inter-Annotator Agreement

The annotation process for the English base data set described above resulted in an overall inter-annotator agreement accuracy of 99.5%, calculated as the percentage of sentences out of the total classified into the same class by both annotators. The English base data set was further evaluated using the Cohen Kappa coefficient [5] as a measure for inter-annotator reliability.

The coefficient was used to assess the role of chance in the annotation process. It yields a score between -1 and 1, where -1 indicates total disagreement and 1 indicates total agreement. A score of 0 points to random annotation or simply that the task is hard and annotators do not always easily agree. The Cohen Kappa has been deemed a useful tool for the evaluation of multi-class classification tasks [9]. According to Galar et al. [9], this is because instead of a mere percentual accuracy score which takes into account all matches for all classes, the Cohen Kappa coefficient scores the accuracy separately and calculates an aggregate score for each class. When the data is unequally stratified in the different classes, this can help eliminate the bias of randomness in inter-annotator agreement scoring. In this type of multi-class classification, a kappa score of 0.6 is the best one could expect [9].

**Table 2.** Inter-annotator agreement in the English data set.

| *ang* | *ant* | *dis* | *fea* | *joy* | *sad* | *sur* | *tru* |
|---|---|---|---|---|---|---|---|
| 0.82 | 0.84 | 0.81 | 0.82 | 0.81 | 0.83 | 0.80 | 0.81 |

As presented in table 2, each sentiment class in the English base data set had a Cohen Kappa score of strong agreement. This indicates an annotation process reliable for evaluating sentiment preservation for the purposes of this study.

To study how sentiment information was retained in the Finnish, French, and Italian translations, the parallel data set of each language was manually annotated. Each annotator annotated in only one language to ensure that they were not influenced by parallel sentences they had previously annotated in another language. The data was annotated into the eight multi-class classes following Plutchik's core eight emotions. The annotators were instructed to follow the principles presented in section 4. Annotations in each target language were then compared against the annotations of the English base data set.

The results of the human classification were evaluated by comparing the number of equally classified parallel data instances. This produced a percentual accuracy score for the preservation of sentiment in each class as well as a total accuracy score.

The results of the annotation work indicate that sentiment is preserved quite well in translation. Table 3 shows the sentiment preservation accuracy scores for each investigated language pair for each label individually as well as in total. Overall, sentiment

information was retained best in the Finnish data, with French fairing slightly worse, and Italian diverging the most. In both English and Finnish, *joy* was the best-preserved sentiment, while in Italian it was *sadness*. *Anger* was preserved surprisingly badly, and *surprise* was the least preserved, as expected. Overall, *surprise* was the most ambiguous sentiment class, with the most sentences marked by the annotators as not belonging to any class. *Disgust* was preserved better than *anger* largely due to a perceived overlap between those two classes by the annotators.

For Italian data, the two annotators produced a near-equal total accuracy score, showing high agreement for most sentiment classes. In the case of Italian and French, both annotators agreed on the sentiment class with the highest accuracy. In Finnish, however, the highest accuracy was given to *joy* by the first annotator and *fear* by the second. The Italian annotators agreed most on the sentiment classes of individual sentences, while the Finnish annotators agreed least.

**Table 3.** Sentiment preservation accuracy.

| | ***pos*** | ***neg*** | *ang* | *ant* | *dis* | *fea* | *joy* | *sad* | *sur* | *tru* | ***Total*** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **EN → FI** | 0.87 | 0.86 | 0.81 | 0.87 | 0.88 | 0.91 | **0.92** | 0.87 | 0.74 | 0.86 | ***0.86*** |
| **EN → FR** | 0.83 | 0.83 | 0.78 | 0.83 | 0.88 | 0.83 | **0.90** | 0.87 | 0.73 | 0.82 | ***0.83*** |
| **EN → IT** | 0.81 | 0.83 | 0.81 | 0.76 | 0.83 | 0.84 | 0.86 | **0.92** | 0.70 | 0.80 | ***0.82*** |

The Cohen Kappa scores displayed in table 4 were at a moderate level of agreement (between 0.40-0.59) for all sentiment classes, which might even be considered good for such a complex and subjective task of annotation work.

**Table 4.** Inter-annotator agreement per language pair (Kappa).

| | *ang* | *ant* | *dis* | *fea* | *joy* | *sad* | *sur* | *tru* |
|---|---|---|---|---|---|---|---|---|
| **EN → FI** | 0.45 | 0.47 | 0.47 | 0.48 | 0.48 | 0.46 | 0.44 | 0.46 |
| **EN → FR** | 0.44 | 0.46 | 0.47 | 0.45 | 0.48 | 0.46 | 0.43 | 0.45 |
| **EN → IT** | 0.44 | 0.43 | 0.45 | 0.42 | 0.43 | 0.48 | 0.40 | 0.43 |

## 5 Discussion

The results achieved in the scope of this work indicate that sentiment is preserved well in translation, although something important seems to be lost in translation. While the degree of preservation may be considered sufficient for using translated data to cross-lingually train sentiment analysis systems, translated sentiment data is likely to contain samples which are not representative of their assigned sentiment class in the translated language. However, as these results have been produced in the scope of this study, it is important to evaluate their relevance for wider application.

One of the things to consider is the annotation process. In all annotation work, there is a limited number of annotators who each have their individual views and experiences on language, social expression, and so on. In addition to these characteristics, each annotator performs an annotation task in a possibly varying mindset. Not only are they influenced by their immediate environment, but they are also influenced by their attitude towards the task [29].

All of these factors make annotators biased. The fundamental difficulty of annotating fine-grained data for subjective tasks such as sentiment analysis is that humans have a relatively low likelihood of reaching high agreement on one most suitable label. This is not only due to differing opinions or experiences between individuals but also to the fact that language is expressively layered, which renders it impossible to assign a universally correct label to a single utterance or fragment of written text.

The annotators for this study were chosen based on native or fluent competence in one of the languages to be annotated. The assumption behind this was that a native speaker would be more likely to pick up on the conventions of encoding sentiment information in their own language. When comparing the two annotators of each language with each other, all annotator pairs produced a considerably similar preservation score. This suggests that there was high consensus between the annotators, pointing to reliability. Despite these numbers, the Cohen Kappa score for each class was in the moderate range of 0.4-0.59, ranging from 0.4 to 0.48. While this indicated acceptable inter-annotator agreement, there was discrepancy between the Cohen Kappa scores and the percentual preservation scores, the latter of which were $>$80%. The Cohen Kappa metric does not provide a way to determine the exact cause for a value lower than the percentual agreement, but it may be influenced by the distribution of the data set. When calculating the Cohen Kappa score of each individual sentiment class for an annotator pair, the distribution of all the classes was considered in the calculation. As no two classes had the same amount of samples, this may have influenced the Cohen Kappa metric. However, despite the difference between the two metrics, the Cohen Kappa was high enough to be considered useful.

The Finnish translations had the highest sentiment preservation and inter-annotator agreement scores, the first annotator producing a preservation percentage of 88% and the second a percentage of nearly 90%. French was preserved second-best at 86% and 87%, and Italian was preserved least well at still a high score of 84% and 85%. There are multiple ways to interpret the causes of one language having higher sentiment preservation than another. Firstly, it is possible that languages of the same language family are more likely to encode sentiment in similar ways and thus preserve sentiment information better in translation. In the case of this study, there was no Germanic language to compare with English, and as sentiment was preserved best in the Finnish data, linguistic similarity is not a viable explanation. However, it is possible that Italian and French were the two less preserved languages due to their inter-linguistic relation as Romance languages. It should be noted, however, that the percentual distances between the best (88%, 90%), the second-best (86%, 87%), and the third-best (84%, 85%) preservation scores have a step of 2-3 percentage points, which makes this seem unlikely.

Another factor to consider when looking at the differences in sentiment preservation between the language pairs is that each translation will have been produced by a

different translator. Each translator can be assumed to have made individual decisions and possibly even had varying translation skills or competency gaps [8], making it less straightforward to compare the sentiment preservation of different translations in different languages. In other words, the translation quality likely fluctuates by a wide margin between translations. Furthermore, the data likely consists of the work of a considerable number of different translators. It is therefore possible that the data in a given language has been translated using varying approaches by varying people who may or may not be native speakers. This means that a possible reason why sentiment information in Italian was preserved worst, for example, could be the quality of the translations compared to that of the Finnish or French translations. The quality may also be reduced by misalignment issues in the corpus files. While the translated sentences were checked to be correctly aligned, some sentences may have been incomplete, due either to misalignment or translation choice. As an example of non-preserved sentiment due to translator choice, the following English sentence has been annotated with the label *joy*, while the parallel Finnish sentence has been assigned the label *surprise* by both annotators:

*It is the most spectacular thing in the Senate annals.*

*Tämä on merkillisintä, mitä senaatissa on koskaan nähty.*

While the English word 'spectacular' has a positive connotation, the Finnish word 'merkillinen' suggests something surprising, possibly with a negative nuance. By contrast, the following French sentence seems to not be comparable to its original in content:

*Did you write me that letter or not?*

*Est-ce exact ou non?*

Considering cases such as the sentence pair above, some of the sentiment information which was not preserved is likely to have been lost because the translation was not exact enough. Those sentences will have relied on contextual information as well as the audiovisuals of the movie (see e.g. [4] on a discussion on the impact of audiovisual cues in translation). In addition to the annotation process and translation quality, it is relevant to consider the implications of using subtitles as data. As the English subtitles used in this work are from English language productions, they may have been produced for the hard-of-hearing or as an additional feature to be used in conjunction with the audiovisuals of a film. Therefore, they may partly rely on the expressiveness of acting, animation, or a soundtrack. It is possible that sentiment is encoded differently when intended to be an addition rather than an integral part of a work of art. The comparison between the sentiment information encoded in a work of art and a text written to express an opinion, such as a review, is also relevant.

As for individual sentiment classes, as shown in chapter 4, *joy* was preserved best in the Finnish and French translations, Italian having a significantly lower score. By contrast, in the Italian data *sadness* had a higher preservation score. Another interesting pattern in the results is the evident ambiguity of the *surprise* data agreed on by most annotators. For example, the following sentence was annotated as *surprise* in English but as *sadness* in French by one annotator, and *disgust* by the other:

*But I always thought little stenographers made little pennies.*

*Je sais bien qu'une petite secrétaire ne roule pas sur l'or.*

The ambiguity of the *surprise* class stems largely from it being perceived as having more overlapping sentiments than some of the other classes, and thus being hard to pinpoint. *Anger* and *disgust* had overlap as they may not be straightforward to distinguish from each other in the first place, let alone identify in self-standing sentences. As suggested earlier in this chapter, since it is not possible to assign a universal label to any given sentence, sentiments as related as *anger* and *disgust* are likely to often appear hand in hand. The following sentence was assigned the label *anger* in the English data set, and *sadness* in the Italian data set by both annotators:

*There are a hundred other places that really need the water!*

*Ci sono centinaia di altri posti che hanno bisogno d'acqua!*

These types of sentence pairs do not necessarily tell much about the cultural or linguistic encoding of sentiment and its interpretation but rather about the general tendency of overlap in sentiment information.

Overall, as sentiment was preserved quite well for all language pairs, and taking into account the possible biases previously mentioned in this section, the use of translated data compiled by annotation projection to train sentiment analysis systems seems a viable solution to provide certain NLP solutions to low-resource languages.

## 6 Conclusions and Future Work

In conclusion, this study found that sentiment information is preserved sufficiently in translated text for data projection to be used for sentiment analysis in particular, but also emotion analysis. The results of the annotation process indicated that a considerable amount of the sentiment information which was lost in translation was due to either 1) an incomplete translation, 2) an ambiguous choice on the part of the translator, or 3) an overlap of possible sentiment classes.

The next logical step would be to perform some actual projection of annotations and compare the results of different classification tasks to those of the hand-annotated data described in this paper.

## References

1. Adams, O., Makarucha, A., Neubig, G., Bird, S., Cohn, T.: Cross-lingual word embeddings for low-resource language modeling. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers. pp. 937–947 (2017).
2. Banea, C., Mihalcea, R., Wiebe, J.: Multilingual Sentiment and Subjectivity Analysis. Multilingual natural language processing **6**, 1–19 (2011), last accessed 2018/10/16.
3. Banea, C., Mihalcea, R., Wiebe, J., Hassan, S.: Multilingual Subjectivity Analysis Using Machine Translation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. pp. 127–135. EMNLP '08, Association for Computational Linguistics, Stroudsburg, PA, USA (2008), last accessed 2017/11/13.
4. Brew, A., Greene, D., Cunningham, P.: The interaction between supervised learning and crowdsourcing. In: NIPS workshop on computational social science and the wisdom of crowds (2010).

5. Cohen, J.: A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement **20**(1), 37 (1960).
6. Creutz, M.: Open Subtitles Paraphrase Corpus for Six Languages. arXiv preprint arXiv:1809.06142 (2018), arXiv:1809.06142 [cs.CL], last accessed 2018/11/09.
7. Digman, J.M., Takemoto-Chock, N.K.: Factors in the natural language of personality: Re-analysis, comparison, and interpretation of six major studies. Multivariate behavioral research **16**(2), 149–170 (1981).
8. Flanagan, M., Christensen, T.P.: Testing post-editing guidelines: how translation trainees interpret them and how to tailor them for translator training purposes. The Interpreter and Translator Trainer **8**(2), 257–275 (2014).
9. Galar, M., Fernández, A., Barrenechea, E., Bustince, H., Herrera, F.: An Overview of Ensemble Methods for Binary Classifiers in Multi-class Problems: Experimental Study on One-vs-one and One-vs-all Schemes. Pattern Recognition **44**(8), 1761–1776 (August 2011), last accessed 2018/08/08.
10. House, J.: Translation quality assessment: Past and present. In: Translation: A multidisciplinary approach, pp. 241–264. Springer (2014).
11. Jain, S., Batra, S.: Cross Lingual Sentiment Analysis using Modified BRAE. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. pp. 159–168. Association for Computational Linguistics (2015), last accessed 2017/12/05.
12. Konstan, D.: Translating ancient emotions : keynote address. Acta Classica : Proceedings of the Classical Association of South Africa **46**(1), 5–19 (2003).
13. Lison, P., Tiedemann, J.: Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. In: Proceedings of the European Language Resources Association. European Language Resources Association (2016).
14. Lohar, P., Afli, H., Way, A.: Maintaining Sentiment Polarity in Translation of User-Generated Content. The Prague Bulletin of Mathematical Linguistics **108**, 73–84 (2017), last accessed 2017/10/11.
15. Lu, B., Tan, C., Cardie, C., Tsou, B.K.: Joint Bilingual Sentiment Classification with Unlabeled Parallel Corpora. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. HLT '11, vol. 1, pp. 320–330. Association for Computational Linguistics, Stroudsburg, PA, USA (2011), last accessed 2017/10/11.
16. McGhee, P.E., Pistolesi, E.: Humor: Its origin and development. WH Freeman San Francisco (1979).
17. Mihalcea, R., Banea, C., Wiebe, J.: Learning Multilingual Subjective Language via Cross-Lingual Projections. In: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (2007), last accessed 2017/10/14.
18. Mohammad, S.M., Salameh, M., Kiritchenko, S.: How Translation Alters Sentiment. Journal of Artificial Intelligence Research **55**, 95–130 (2016), last accessed 2017/09/24.
19. Plutchik, R.: A general psychoevolutionary theory of emotion. Theories of emotion **1**, 3–31 (1980).
20. Ragni, A., Knill, K.M., Rath, S.P., Gales, M.J.: Data augmentation for low resource languages. In: Fifteenth Annual Conference of the International Speech Communication Association (2014).
21. Sabou, M., Bontcheva, K., Scharl, A.: Crowdsourcing research opportunities: lessons from natural language processing. In: Proceedings of the 12th International Conference on Knowledge Management and Knowledge Technologies. p. 17. ACM (2012).
22. Salameh, M., Mohammad, S., Kiritchenko, S.: Sentiment after Translation: A Case-Study on Arabic Social Media Posts. In: HLT-NAACL. pp. 767–777. The Association for Computational Linguistics (2015), last accessed 2017/09/29.

23. Tang, X., Wan, X.: Learning Bilingual Embedding Model for Cross-Language Sentiment Classification. In: Proceedings of the 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT) - Volume 02. pp. 134–141. WI-IAT '14, IEEE Computer Society, Washington, DC, USA (2014), last accessed 2017/12/10.
24. Waltz, D.L., Pollack, J.B.: Massively parallel parsing: A strongly interactive model of natural language interpretation. Cognitive science **9**(1), 51–74 (1985).
25. Wan, X.: Bilingual Co-Training for Sentiment Classification of Chinese Product Reviews. Computational Linguistics **37**(3), 587–616 (2011), last accessed 2017/12/06.
26. Wassmann, C.: Forgotten origins, occluded meanings: Translation of emotion terms. Emotion Review **9**(2), 163–171 (2017).
27. Wierzbicka, A.: Emotions across Languages and Cultures: Diversity and Universals. Cambridge University Press (1999).
28. Zoph, B., Yuret, D., May, J., Knight, K.: Transfer learning for low-resource neural machine translation. arXiv preprint arXiv:1604.02201 (2016).
29. Öhman, E.: Challenges in emotion annotation:annotator experiences from a crowd-sourced annotation task (2020), forthcoming.
30. Öhman, E., Honkela, T., Tiedemann, J.: The challenges of multi-dimensional sentiment analysis across languages. In: Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media (PEOPLES). pp. 138–142. The COLING 2016 Organizing Committee, Osaka, Japan (December 2016), last accessed 2018/10/16.
31. Öhman, E., Kajava, K.: Sentimentator: Gamifying Fine-grained Sentiment Annotation. In: Digital Humanities in the Nordic Countries 2018. CEUR Workshop Proceedings (2018), last accessed 2018/10/16.
32. Öhman, E.S., Tiedemann, J., Honkela, T.U., Kajava, K., et al.: Creating a dataset for multilingual fine-grained emotion-detection using gamification-based annotation. In: Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis. Association for Computational Linguistics (2018).