Balancing Translation Quality and Sentiment Preservation

Pintu Lohar pintu.lohar@adaptcentre.ie
Haithem Afli haithem.afli@adaptcentre.ie
Andy Way andy.way@adaptcentre.ie
ADAPT Centre, School of Computing, Dublin City University

Abstract

Social media platforms such as Twitter and Facebook are hugely popular websites through which Internet users can communicate and spread information worldwide. On Twitter, messages (tweets) are generated by users from all over the world in many different languages. Tweets about different events almost always encode some degree of sentiment. As is often the case in the field of language processing, sentiment analysis tools exist primarily in English, so if we want to understand the sentiment of the original tweets, we are forced to translate them from the source language into English and pushing the English translations through a sentiment analysis tool.

However, Lohar et al. (2017) demonstrated that using freely available translation tools often caused the sentiment encoded in the original tweet to be altered. As a consequence, they built a series of sentiment-specific translation engines and pushed tweets containing either positive, neutral or negative sentiment through the appropriate engine to improve sentiment preservation in the target language. For certain tasks, maintaining sentiment polarity in the target language during the translation process is arguably more important than the absolute translation quality obtained. In the work of Lohar et al. (2017), a small drop off in translation quality *per se* was deemed tolerable. In this work, we focus on maintaining the level of sentiment preservation while trying to improve translation quality still further. We propose a nearest sentiment class-combination method to extend the existing sentiment-specific translation systems by adding training data from the nearest-sentiment class. Our experimental results on German-to-English reveal that our approach is capable of achieving a proper balance between translation quality and sentiment preservation.

1 Introduction

The rapid development of internet technologies has given rise to a significant growth in generating and sharing of user-generated content (UGC). Internet users from all over the world stay connected via widely used social networking websites such as Twitter and Facebook by sharing information in different languages. On Twitter, tweets on different events related to sports, festivals, conferences and political events almost always encode some degree of sentiment. Accordingly, the task of sentiment analysis is important on datasets such as these. However, given the lack of such tools for most languages, this can only be achieved via an MT-based sentiment analysis approach, where the tweets are first translated from the original language to English and then sentiment analysis is performed on th English translations (Araujo et al. (2016)).

However, the 140-character limitation – recently expanded to 280 – in Twitter encourages

users to use short forms at word and phrase levels. Moreover, tweets often contain (deliberate) spelling errors, hashtags, user names and URLs which pose challenges in the translation process. In order to build corpus-based MT systems, a parallel corpus is a prerequisite, but parallel UGC data is in very short supply. In recent work based on sentiment translation system, Lohar et al. (2017) collected a parallel data comprising $4,000$ English–German tweet pairs[1] in the football domain (extracted from the FIFA World Cup 2014 Twitter feed) and built sentiment translation engines using a sentiment classification approach. In that work, the sentiment translation models were built using the (i) negative, (ii) neutral, and (iii) positive tweet pairs, respectively. Their experimental results showed that the sentiment classification approach is very useful for preserving the sentiment in the target language during translation. However, the MT quality deteriorated a little due to dividing the small corpus of $4,000$ parallel tweets into even smaller ones with different sentiment classes for training sentiment-specific MT engines.

In this work, we try to retain the degree of sentiment while at the same time minimizing any loss in translation quality. We propose to extend the sentiment-specific translation models by incorporating neighbouring sentiment data. We perform the following steps to build our extended sentiment translation system: (i) building a single baseline translation model by using the whole Twitter data regardless of the sentiment classes, (ii) building separate negative, neutral and positive sentiment translation models using the negative, neutral and positive tweet pairs, respectively, (iii) combining the negative and neutral tweet pairs to build a translation system conveying both of these sentiments, and (iv) combining the positive and neutral tweet pairs to build a translation system conveying both of these sentiments. Steps (iii) and (iv) are the main contributions in this work that extend the previously mentioned sentiment translation system of Lohar et al. (2017). The reason behind such combinations of sentiment classes is that the neutral class is relatively closer to both the negative and positive classes, compared to the distance between the negative and positive classes in terms of sentiment score. This is motivated by the fact that in their work, Lohar et al. (2017) demonstrated that while MT can alter the original sentiment, it typically transfers to the immediately neighbouring class (i.e. from negative to neutral (or vice-versa) or from positive to neutral (or vice-versa)), but rarely from positive to negative (or vice-versa).

The remainder of this paper is organised as follows. We briefly describe some relevant related work in Section 2. In Section 3, we provide an architectural overview of our sentiment classification MT system. The experimental set ups are discussed in Section 4, followed by a detailed discussion of the results in Section 5. Finally, Section 6 concludes together with some avenues for future work.

2 Related work

Translating UGC creates new challenges in the area of MT. Jiang et al. (2012) describe how to handle shortforms, acronyms, typos, punctuation errors, non-dictionary slang, wordplay, censor avoidance and emoticons, phenomena which are characteristic of UGC but not of 'normal' written forms in language. The combination of statistical machine translation (SMT) and a preprocessor was also applied to remove a significant amount of noise from tweets in order to convert them into a more readable format (Kaufmann and Kalita (2010)). Gotti et al. (2013) use an SMT system to translate Twitter feeds published by agencies and organisations. They create tuning and training sets by mining parallel web pages linked from the URLs contained in English–French pairs of tweets.

There exists quite a lot of research in the area of sentiment analysis of UGC. For example, Fang and Zhan (2015) analyse the sentiment polarity of online product reviews extracted from

[1]Recently released in Lohar et al. (2018) and available at: `https://github.com/HAfli/FooTweets_Corpus`

Amazon.com using both sentence-level and review-level categorization techniques. Gräbner et al. (2012) classify customer reviews of hotels by extracting a domain-specific lexicon of semantically relevant words based on a given corpus (Scharl et al. (2003); Pak and Paroubek (2010)). Broß (2013) focus on the following two main subtasks of aspect-oriented review mining: (i) identification of the relevant product aspects, and (ii) determining and classifying the expressions of the sentiment.

Some existing work applies MT for the task of sentiment analysis. For example, Mohammad et al. (2016) show that the sentiment analysis of English translations of Arabic text produces competitive results compared to Arabic sentiment analysis *per se*. In a similar vein, Araujo et al. (2016) reveal that simply translating the non-English input text into English and using the English sentiment analysis tool can be better than the existing language-specific efforts evaluated. In contrast, Afli et al. (2017) demonstrate that building a sentiment analysis tool for a low-resource language, namely Irish, can outperform strategies including translation as an integral sub-task. Their approach includes the following strategies: (i) using the existing English sentiment analysis resources to both manually and automatically translated tweets, and (ii) manually creating an Irish-language sentiment lexicon – Senti-Foclóir – to build the first Irish sentiment analysis system – SentiFocalTweet – which produces superior results to the first method.

Importantly, although MT can be useful for the sentiment analysis task, it can alter the sentiment of the source-language text in the target language during the translation process (Mohammad et al. (2016)). For example, a text in the source language (say Arabic) with positive sentiment may not retain its positivity when translated into the target language (say English). To address such problems, Lohar et al. (2017) propose a sentiment classification approach to build sentiment-specific translation models that aim at maintaining the sentiment polarity of the source-language text during the translation process. The results revealed that it is possible to increase the sentiment preservation score by using the sentiment translation systems. In the present paper, we extend that work by incorporating the nearest neighbour sentiment classes in order to build extended sentiment translation engines that incorporate the sentiment of the neighbouring sentiment class. To the best of our knowledge, no existing work has attempted such an approach to addressing the problems of maintaining translation quality and sentiment preservation in parallel.

3 Architecture of the sentiment translation system

Figure 1 shows the architecture of our proposed sentiment translation system using the nearest neighbour sentiment classes. The complete workflow of the whole system can be described in following steps:

(i) sentiment classification is performed on the whole corpus,

(ii) the negative and the neutral tweet pairs are grouped together as the nearest neighbour sentiment classes,

(iii) apart from being used for combination, the neutral tweet pairs are also kept for separate usage,

(iv) the positive and the neutral tweet pairs are grouped together as the nearest neighbour sentiment classes,

(v) from the above three corpus sets, three different translation systems are built: negative_neutral, neutral and positive_neutral models, respectively,

(vi) the test data is divided into negative, neutral and positive sentiment classes,

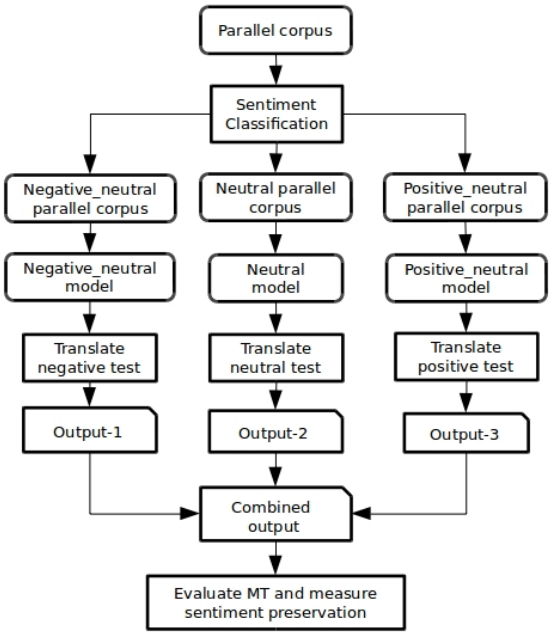

Figure 1: Sentiment translation using nearest neighbour sentiment classes

(vii) the negative, neutral and the positive test data are translated by the negative_neutral, neutral and the positive_neutral translation models, respectively,

(viii) the outputs produced by these three translation models are combined, and

(ix) the combined output is used to measure the MT and sentiment-preservation quality.

4 Experiments

4.1 Data statistics

Lohar et al. (2017) held out a small subset of only 50 tweet pairs per sentiment class (negative, neutral and positive, so 150 sentence pairs in total) for testing purposes in order to maintain as large an amount as possible for training the sentiment translation systems. However, it has to be acknowledged that it is difficult to judge the system's performance with only 150 test pairs. We therefore maintain two different distributions of the whole data set and use each of them separately in the two different experimental set-ups (*Exp1* and *Exp2*). We hope that by slightly increasing the size of the development and test sets, our analysis of the performance of the proposed system with these two different set-ups will be somewhat more informative.

Exp. setup	Train	Development			Test		
		#neg.	#neu.	#pos.	#neg.	#neu.	#pos.
Exp1	3,700	50	50	50	50	50	50
Exp2	3,400	100	100	100	100	100	100

Table 1: Data statistics

4.2 Resources and tools

All the translation models are built by using the widely used open source Moses SMT toolkit (Koehn et al. (2007)). The word and phrase alignments are obtained by using the Giza++ tool (Och and Ney (2003)). Once the translation models are built, we tune all the sentiment translation systems for both experimental set-ups (*Exp1* and *Exp2*) via minimum error rate training (Och (2003)).

4.3 Evaluation process

We use the automatic MT evaluation metrics BLEU (Papineni et al. (2002)), METEOR (Denkowski and Lavie (2014)) and TER (Snover et al. (2006)) to evaluate the absolute translation quality obtained. We measure the sentiment preservation score by calculating what percentage of the tweets belongs to the same sentiment class before and after translation.

5 Results

Table 2 shows the results obtained from the first experimental setup (*Exp1*) which is comparable with the previous results obtained in the work of Lohar et al. (2017) because the data distribution is the same (i.e. 150 tweet pairs for each of the development and test data sets).

System (Exp1)	BLEU	METEOR	TER	Sent_Pres.
Twitter_Baseline	50.3	60.9	31.9	66.66%
Twitter_SentClass	48.2	59.4	34.2	72.66%
Twitter_NearSent	49.0	60.1	34.0	66.66%

Table 2: Experiment 1: performance comparison on small test sets

The 2^{nd} and the 3^{rd} rows in Table 2 show the comparison between the two different systems (i.e., Twitter_Baseline and Twitter_SentClass) that were reported in Lohar et al. (2017). Twitter_Baseline is the translation system where sentiment classification is not used, whereas Twitter_SentClass is the system where it is switched on.

In contrast, "Twitter_NearSent" is our proposed system (including nearest sentiment class-combination) with the same amount of training, development and test data. As can be seen in this table, Twitter_SentClass obtains the highest sentiment preservation score of 72.66% but the BLEU, METEOR and TER scores are worse than the Baseline. Lohar et al. (2017) state that as expected, as the sentiment-classification approach divides the whole corpus into smaller parts with specific sentiment classes, the translation models built from each data set are much smaller than the baseline model and so the performance decreases accordingly. However, the sentiment preservation score is significantly *increased* (from 66.66% to 72.66%) which was the main objective of that work. In this work, our objective is to reduce this gap. More precisely, we are interested in obtaining better MT scores than the Twitter_SentClass system but at the same achieving the better sentiment preservation than the Baseline, which we hope to achieve using the Twitter_NearSent system. Although the sentiment preservation score is not increased at all, we do indeed manage to increase the MT scores: BLEU and METEOR increase by 0.8 (1.7% relative) and 0.7 (1.2% relative) points, respectively, while TER decreases by 0.2 points (1.5% relative improvement).

These results demonstrate that using such a small amount of test data set (i.e. only 150 tweet pairs) is insufficient to confirm the utility of our approach. We therefore decided to rerun the whole set of experiments on the new distribution (*Exp2*, see the 3^{rd} row of Table 1) of the training, development and test data (i.e. 100 sentence-pairs per sentiment for both the development and test data sets, respectively). Table 3 shows the results produced during our second

stage of experiments using the "Exp2" setup. As this data distribution is different from "Exp1", the results obtained are different and so are not directly comparable with the results shown in Table 2.

System (Exp2)	BLEU	METEOR	TER	Sent_Pres.
Twitter_Baseline	51.3	62.5	31.0	52.33%
Twitter_SentClass	47.3	59.1	35.2	60.33%
Twitter_NearSent	48.3	59.6	34.4	60.0%

Table 3: Experiment 2: performance comparison on larger test sets

It can be observed from this table that the "Twitter_SentClass" system produces the highest sentiment-preservation score of 60.33% – 8% (or 15.3% relative) better than the Baseline – but the MT scores are much lower than the Baseline. For example, the BLEU score is reduced by exactly 4 points (from 51.3 down to 47.3, an almost 8% relative reduction in translation quality), with the other automatic metrics corroborating this deterioration in performance. In contrast, our proposed system (Twitter_NearSent) raises the BLEU score by exactly 1 point (from 47.3 to 48.3, a 2.1% relative improvement) while at the same time the sentiment-preservation score is reduced by only 0.33%. In parallel, METEOR score is also increased from 59.1 to 59.6 (a 0.85% relative improvement) and the TER score is also improved (drops from 35.2 to 34.4, a 2.3% relative improvement). These results suggest that using a comparatively larger test set gives us a clearer picture of the performance of the various engines than the smaller one with "Exp1" set-up (see Table 2).

The above observations are important in terms of the balance between translation quality and sentiment preservation. Our new system is still able to significantly improve the sentiment-preservation score (from 52.33% to 60.0%) over the Baseline which is very close to the maximum value obtained (60.33%), yet at the same time manages to increase the MT score compared to the original engines built in Lohar et al. (2017).

6 Conclusion and Future Work

The current paper presented a novel extension to building sentiment-translation engines based on the combination of the nearest sentiment-class parallel data. Our new translation models are built by: (i) combining the negative and the neutral tweet pairs in order to build our first nearest neighbour sentiment translation model (*Negative_neutral model*), and (ii) combining the positive and the neutral tweet pairs to build the second nearest neighbour sentiment translation model (*Positive_neutral model*). We performed experiments in two stages with two different data distributions as the initial experiment was conducted on the smallest data set which we assumed to be insufficient to confirm the results to any great extent. The results obtained in the second stage of our experiments revealed that we can significantly increase the sentiment-preservation score (very close to the maximum value obtained) while at the same time obtaining improved MT scores than the sentiment-translation engines in Lohar et al. (2017).

Accordingly, our approach maintains a better balance between translation quality *per se* and sentiment preservation. In future, we will apply our approach to other forms of UGC such as user feedback, blogs, and reviews and with larger data sets. We will also conduct experiments on language pairs other than German-to-English. We also noticed that some of the tweets in our data sets are less related to the main topic (football) than originally anticipated, so it will be interesting to see how we can combine such out-of-domain data with true in-domain tweet pairs when building the next batch of sentiment-translation engines.

Acknowledgments

The ADAPT Centre for Digital Content Technology at Dublin City University is funded under the Science Foundation Ireland Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund.

References

Afli, H., McGuire, S., and Way, A. (2017). Sentiment translation for low resourced languages: Experiments on irish general election tweets. In *18th International Conference on Computational Linguistics and Intelligent Text Processing*, Budapest, Hungary. 10 pages.

Araujo, M., Reis, J., Pereira, A., and Benevenuto, F. (2016). An evaluation of machine translation for multilingual sentence-level sentiment analysis. In *Proceedings of the 31st Annual ACM Symposium on Applied Computing*, pages 1140–1145, New York, USA.

Broß, J. (2013). *Aspect-Oriented Sentiment Analysis of Customer Reviews Using Distant Supervision Techniques*. PhD thesis, Freie Universität Berlin, Berlin, Germany.

Denkowski, M. and Lavie, A. (2014). Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380, Baltimore, USA.

Fang, X. and Zhan, J. (2015). Sentiment analysis using product review data. *Journal of Big Data*, 2(5). 14 pages.

Gotti, F., Langlais, P., and Farzindar, A. (2013). Translating government agencies' tweet feeds: Specificities, problems and (a few) solutions. In *Proceedings of the Workshop on Language Analysis in Social Media*, pages 80–89.

Gräbner, D., Zanker, M., Fliedl, G., and Fuchs, M. (2012). Classification of customer reviews based on sentiment analysis. In *Proceedings of the International Conference on Information and Communication Technologies*, pages 460–470, Helsingborg, Sweden. Springer Vienna.

Jiang, J., Way, A., and Haque, R. (2012). Translating user-generated content in the social networking space. In *Proceedings of the Tenth Biennial Conference of the Association for Machine Translation in the Americas*, San Diego, USA. 9 pages.

Kaufmann, M. and Kalita, J. (2010). Syntactic normalization of Twitter messages. In *Proceedings of the 8th International Conference on Natural Language Processing*, pages 149–158, Kharagpur, India.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 177–180, Prague, Czech Republic.

Lohar, P., Afli, H., and Way, A. (2017). Maintaining sentiment polarity in translation of user-generated content. *The Prague Bulletin of Mathematical Linguistics*, 108:73–84.

Lohar, P., Afli, H., and Way, A. (2018). Footweets: A bilingual parallel corpus of world cup tweets. In *LREC-2018, Proceedings of the 11th International Conference on Language Resources and Evaluation*, Miyazaki, Japan. (to appear).

Mohammad, S. M., Salameh, M., and Kiritchenko, S. (2016). How translation alters sentiment. *Journal of Artificial Intelligence Research*, 55(1):95–130.

Och, F. J. (2003). Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pages 160–167, Sapporo, Japan.

Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29:19–51.

Pak, A. and Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of the Seventh conference on International Language Resources and Evaluation*, pages 1320–1326, Valletta, Malta.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania.

Scharl, A., Pollach, I., and Bauer, C. (2003). Determining the semantic orientation of web-based corpora. In *4th International Conference on Intelligent Data Engineering and Automated Learning*, pages 840–849, Hong Kong, China.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In *AMTA 2006: Proceedings of the 7th Conference of the Association for Machine Translation in the Americas, "Visions for the Future of Machine Translation*, pages 223–231, Cambridge, Massachusetts, USA.