PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 108 JUNE 2017 73–84

Maintaining Sentiment Polarity in Translation of User-Generated Content

Pintu Lohar, Haithem Afli, Andy Way

ADAPT Centre, School of Computing, Dublin City University, Dublin, Ireland

Abstract

The advent of social media has shaken the very foundations of how we share information, with Twitter, Facebook, and Linkedin among many well-known social networking platforms that facilitate information generation and distribution. However, the maximum 140-character restriction in Twitter encourages users to (sometimes deliberately) write somewhat informally in most cases. As a result, machine translation (MT) of user-generated content (UGC) becomes much more difficult for such noisy texts. In addition to translation quality being affected, this phenomenon may also negatively impact sentiment preservation in the translation process. That is, a sentence with positive sentiment in the source language may be translated into a sentence with negative or neutral sentiment in the target language. In this paper, we analyse both sentiment preservation and MT quality *per se* in the context of UGC, focusing especially on whether sentiment classification helps improve sentiment preservation in MT of UGC. We build four different experimental setups for tweet translation (i) using a single MT model trained on the whole Twitter parallel corpus, (ii) using multiple MT models based on sentiment classification, (iii) using MT models including additional out-of-domain data, and (iv) adding MT models based on the phrase-table fill-up method to accompany the sentiment translation models with an aim of improving MT quality and at the same time maintaining sentiment polarity preservation. Our empirical evaluation shows that despite a slight deterioration in MT quality, our system significantly outperforms the Baseline MT system (without using sentiment classification) in terms of sentiment preservation. We also demonstrate that using an MT engine that conveys a sentiment different from that of the UGC can even worsen both the translation quality and sentiment preservation.

1. Introduction

The world of social media has experienced significant growth in the last decade. With the advent of Web 2.0, we are all publishers these days, which means that the

 Corresponding author: pintu.lohar@adaptcentre.ie
Cite as: Pintu Lohar, Haithem Afli, Andy Way. Maintaining Sentiment Polarity in Translation of User-Generated Content. The Prague Bulletin of Mathematical Linguistics No. 108, 2017, pp. 73–84.
doi: 10.1515/pralin-2017-0010.

amount of UGC created is enormous, multilingual, diverse and of varying quality. Accordingly, building robust, high-quality MT engines can be problematic, especially when users deliberately decide to violate linguistic norms in the languages they speak (cf. Jiang et al. (2012)). Twitter, one of the largest social media websites, enables people throughout the world to share information and express their opinion (in the form of tweets) in the language of their choice. Many Twitter users follow others who do not tweet in their preferred language. In such a case, tweets in a specific language need to be translated into the language of choice of such users. As well as the 140-character restriction mentioned above, tweets are often generated using mobile devices, which contributes further to the poor quality of language, including spelling and other errors, omission of diacritics etc. Tweets also contain hashtags, user handles, retweets etc., all of which makes tweet translation a difficult task. This task can be done directly (tweet-to-tweet), or indirectly via tweet normalization (Kaufmann and Kalita, 2010; Jiang et al., 2012).

Leaving quality *per se* to one side for one moment, errors in translation can negatively impact the sentiment of the source-language tweet, e.g. a tweet in English conveying positive sentiment may not retain its positivity after being translated into Japanese. Especially in business contexts, where large multinational companies want to find out what their users think of their products and services, sentiment preservation of the original tweets is arguably as important as the overall translation quality. Accordingly, in this work, we mainly focus on incorporating sentiment classification within our MT systems to investigate the extent to which the sentiment of tweets in the source language is preserved in the target language. Our aim is to improve sentiment preservation from source-to-target language tweets while at the same time minimizing any performance degradation in translation. We use parallel Twitter data set consisting of $4,000$ English tweets from the FIFA World Cup 2014 and there translations into German. We conduct four experiments on tweet translation: (i) a Baseline translation model built from the whole parallel corpus of tweets is used to translate the tweets, (ii) the data is divided according to specific sentiment classes to build different translation models for positive, negative and neutrally-sentimented tweets, (iii) the Twitter data is amalgamated with comparably much larger out-of-domain data sets,[1] and the sentiment translation model is combined with (a) small and (b) large out-of-domain models in order to apply phrase-table fill-up (Bisazza et al. (2011)).

The remainder of this paper is organized as follows. Section 2 highlights related work in this area. We describe our sentiment classification system in Section 3. Section 4 presents the different experiments, while Section 5 provides the empirical evaluation results, together with an analysis of our findings. Finally, we conclude and outline possible future work in Section 6.

[1]They *are* UGC, but the domains are not football-related.

2. Related Work

A significant amount of work has been done in the area of translation of UGC, and especially sentiment translation. The earliest work we are aware of is that of Kanayama et al. (2004), who use a transfer-based MT engine to translate text documents to a set of sentiment units. A graph-based approach using SimRank to transfer sentiment information from a source language to a target language is presented in Scheible et al. (2010). Saif et al. (2016) examine sentiment analysis in Arabic, a (relatively) resource-poor language. They use two approaches to examining the sentiment of Arabic social media posts: (i) translate the focus language text into a resource-rich language such as English, and apply a powerful English sentiment analysis system on the text, and (ii) translate resources such as sentiment-labeled corpora and sentiment lexicons from English into the focus language, and use them as additional resources in the focus-language sentiment-analysis system. They show that the sentiment analysis of English translations of Arabic texts produces competitive results, with respect to the Arabic sentiment analysis, and the Arabic sentiment analysis systems benefit from the use of automatically translated English sentiment lexicons. Balahur and Turchi (2012) deal with the problem of sentiment detection in three different languages (French, German and Spanish) using three distinct MT systems: Bing,[2] Google,[3] and Moses (Koehn et al., 2007). These systems are used to translate the *training* data so that English sentiment analysis can be applied to the output. In a similar vein, Araujo et al. (2016) show that simply translating the input text (the *test* data) from a specific language to English and then using one of the existing methods for English can be better than the existing language-specific efforts evaluated.

In parallel with the area of sentiment translation, crosslingual sentiment analysis (CLSA) has also undergone significant evolution. Lin et al. (2014) develops a model to carry out aspect-specific sentiment analysis in a target language using the knowledge learned from a source language. The task of crosslingual sentiment lexicon learning by automatically generating target-language sentiment lexicons from available English sentiment lexicons is addressed in Gao et al. (2015). Jain and Batra (2015) use the recursive auto-encoder architecture to develop a CLSA tool using sentence-aligned corpora between a resource-rich (English) and a resource-poor (Hindi) language. He et al. (2015) propose a semi-supervised learning approach with "space transfer" to tackle the task of cross-language sentiment classification. The work in Balahur and Turchi (2013) shows that the joint use of training data from multiple languages (especially those pertaining to the same family of languages) significantly improves the results of the sentiment classification. Baker et al. (2012) incorporate related aspects of meaning such as modality into the translation process in order to both maintain semantics across translation and improve translation quality. However, to the best

[2] https://www.bing.com/translator

[3] https://translate.google.com/

of our knowledge, none of the work to date has attempted a sentiment classification approach aimed at preserving the sentiment in translation. Our proposed method integrates the sentiment classification approach in building different translation models based on specific sentiment classes. Then the particular sentiment-translation model is used to translate the tweets with that sentiment polarity. This output is compared against a Baseline system built with all Twitter data, as well as systems based on phrase-table fill-up method.

3. Sentiment classification

3.1. Manual sentiment classification

We use a Twitter data set comprising 4,000 English tweets from the FIFA World Cup 2014 , their manual translations into German and the annotated sentiment scores (prepared by anon). As might be expected, these tweets are rather informal in nature e.g. the English tweet "GOAAAAL ♡ ♥ ♡ ♥" is translated as "TOOOOR ♡ ♥ ♡ ♥" in German in order to emphasize the positive emotion in the target language. We consider the tweets with manually annotated sentiment scores as our 'gold standard' data. The tweets are categorised into the following three classes: (i) negative tweets with sentiment score ≤ 0.4, (ii) neutral tweets with sentiment score $\simeq 0.5$ and (iii) positive tweets with sentiment score ≥ 0.6. Once the tweet categorization was complete, we held out a very small subset – 50 tweets per sentiment (negative, neutral and positive) – for tuning and testing purposes because we wanted to maintain as large an amount as possible for training the MT systems. For phrase-table fill-up, we include parallel sentence pairs from (i) an English–German parallel Flickr data set[4]o train a small out-of-domain model, and (ii) a much larger data set, namely the English-German parallel "News-Commentary" corpus[5] to build a large out-of-domain model. These data are also merged with the Twitter data to create additional training resources. The objectives here were to see the effects on both MT quality and sentiment preservation when the out-of-domain data is included.

The statistics of the number of parallel data used for training, tuning and testing is shown in Table 1. Of course, 3,700 training examples (tweets in this case) is not a large amount of data in the first place, and in our non-Baseline models we reduce this data size still further. Nonetheless, as will be seen in Section 4, good results *can* be achieved with such very small amounts of training data – albeit on admittedly small test sets – contrary to the perceived wisdom in the field.

The manually annotated sentiment scores are available only for the Twitter data because the Flickr and the News data are much larger, and so their manual annotation

[4] http://www.statmt.org/wmt16/multimodal-task.html#task1

[5] http://data.statmt.org/wmt16/translation-task/training-parallel-nc-v11.tgz

Data	Train	Development			Test		
		#negative	#neutral	#positive	#negative	#neutral	#positive
Twitter	3,700	50	50	50	50	50	50
Flickr	29,000	50	50	50	50	50	50
News_comm	235,843	50	50	50	50	50	50

Table 1: Data statistics

is practically infeasible. Therefore we apply an automatic sentiment analysis tool (see Section 3.2) to extract the sentiment scores for these data sets.

3.2. Automatic Sentiment classification

This approach involves automatic extraction of the sentiment scores of the tweets (or sentences) and their classification into negative, neutral and positive tweets (sentences) with the same criteria for scoring discussed in Section 3.1. We use a lexicon-based sentiment analysis (SA) system especially designed for tweets in low-resourced languages (Afli et al., 2017). This system makes use of SentiWordNet (Esuli and Sebastiani, 2006), an opinion lexicon derived from WordNet (Miller, 1995) where each word is associated with numerical scores (from zero to one) indicating the strength of being positive, negative or neutral. SentiWordNet word values have been semi-automatically computed based on training a set of ternary classifiers, each capable of deciding the polarity of the synset. The process begins with pre-processing of the raw tweets in following three modules: (i) **tokenization**: splitting the tweet into very simple tokens such as numbers, punctuation and words of different types; (ii) **sentence splitting**: segmenting the text into sentences, if there is more than one in the tweet. This module is required for the part of speech (PoS) tagger. (iii) **PoS tagging**: producing a PoS tag as an annotation on each word or symbol. Afterwards, SentiWordNet is used to score each PoS-tagged word in the tweet. Subsequently, exponential weighting and the words magnitude scoring techniques are applied on the tokenised and split text. Finally, in order to obtain the overall sentiment score of each tweet, the scores are added and normalized by the number of tweet words.

We evaluate the performance of the sentiment analysis tool of Afli et al. (2017) in classifying sentiments correctly. Out of the 4,000 tweets, 2,994 tweets are correctly classified when compared to the gold standard manual sentiment classification, giving a performance accuracy of 74.85%.

4. Experiments

4.1. Sentiment translation

We consider German as the source language and English as the target in order to be able to use the English SA tool for the English translation of the German tweets. For the Twitter data, we divide the train data of 3,700 tweet pairs into negative, neutral

Data	Sentiment Classification	#Negative	#Neutral	#Positive	#Total
Twitter	manual	919	1,308	1,473	3,700
Twitter	automatic	630	1,343	1,727	3,700
Flickr	automatic	9,677	11,065	8,258	29,000
News_comm	automatic	111,337	14,306	113,200	238,843

Table 2: Data distribution after sentiment classification

and positive tweet pairs using both the manual and automatic sentiment classification approaches. In contrast, since the manually annotated versions of Flickr and News data are unavailable, we apply the automatic SA tool on these data in order to extract the sentiment scores. Table 2 shows the distribution of negative, neutral and positive tweet/sentence pairs after manual and automatic sentiment classification.

In order to build the translation models, we use the Moses statistical MT (SMT) toolkit, which uses Giza++ (Och and Ney, 2003) for word and phrase alignment. The models are tuned using minimum error rate training (Och, 2003). Each of the translation models is built from the parallel data with a specific sentiment category and sentiment classification approach, respectively, e.g. a 'positive sentiment' translation model (see Table 2) is built from the 1,727 positive tweet pairs. The translation models conveying the particular sentiment types are referred to as "negative", "neutral" and "positive", respectively, whereas the single model trained on the whole 3,700-tweet pairs is termed the "Baseline". Note that our smallest system is built with just 630 tweet-pairs. Despite the fact that this may be the smallest SMT system ever published, as will be seen in Table 4, good results can nonetheless be achieved. The architectural overview of the sentiment translation system is shown in Figure 1. Note that the two boxes 'Output combination1/2' only merge the different polarity translations at tweet-level prior to the whole 150-tweet test set being sent for evaluation; there is no intention to suggest that parts of individual tweets are reassembled here into 'whole tweet' translations. More precisely, the whole 150-tweet test set is comprised of 50-tweets per sentiment class and each of them is translated using the corresponding translation model and the outputs are combined.

The main experiment consists of three different approaches; (i) translation without sentiment classification, (ii) translation with manual sentiment classification and (iii) translation with automatic sentiment classification which are discussed in the following sections. We also conduct experiments on data concatenation, and use phrase table fill-up method (Bisazza et al., 2011) to see whether it is possible to increase the translation quality and at the same time maintain the sentiment preservation.

4.1.1. **Translation without sentiment classification**

In this set-up, we build three translation models: (i) one with Twitter data, (ii) one with Twitter data combined with Flickr data, and (iii) one with Twitter, Flickr and News data combined.

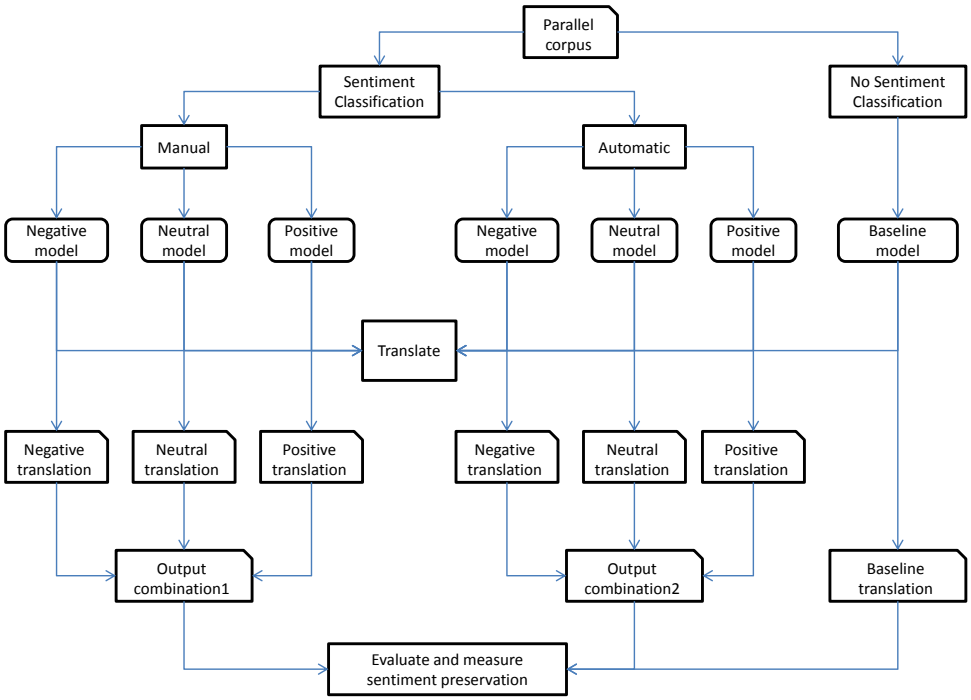

Figure 1: Architecture of the Sentiment Translation System

4.1.2. **Translation with manual sentiment classification**

In this approach, the three sentiment-translation models (with negative, neutral and positive sentiments) trained on the Twitter data with the gold standard sentiment annotations (the 'oracle', henceforth) translate the appropriate test set with the same sentiment polarity.

4.1.3. **Translation with automatic sentiment classification**

Here, we apply the SA tool to all the data sets and then train the sentiment-translation models under each sentiment class. This experiment is designed to test the expected fall off in accuracy with automatic sentiment classification.

In addition to this, we also make use of the phrase-table fill-up method using (i) one with Flickr data, and (ii) one with News data.

Translation model	Oracle	Sent_Clas.	BLEU	METEOR	TER	Sent_Pres.
Twitter	✓	✓	48.2	59.4	34.2	72.66%
Twitter	×	✓	48.1	58.9	34.6	68.0%
Twitter (Baseline)	✓	×	50.3	60.9	31.9	66.66%
Twitter + Flickr	×	✓	48.5	59.8	33.9	71.33%
Twitter + Flickr	×	×	50.7	62.0	31.3	62.66%
Twitter + Flickr + News_Comm	×	✓	50.3	62.3	31.0	**75.33%**
Twitter + Flickr + News_Comm	×	×	**52.0**∗	**63.4**∗	**30.1**∗	73.33%
Twitter (wrong MT engine)	✓	✓	46.9	57.9	35.4	47.33%

Table 3: Experimental evaluation: With data concatenation

5. Results

We conduct our experiments taking into account both the translation quality *per se* as well as the sentiment polarity preservation. The results are summarized in Table 3 which shows that, where only the Twitter data is used, the best BLEU, METEOR and TER scores are obtained when no sentiment classification (referred to as "Sent_Clas.") is applied ('Twitter (Baseline)'), i.e. when all Twitter data is merged, regardless of sentiment. The scores improve further when the Flickr data is used as additional training data, despite the fact that it is out-of-domain; when no sentiment classification is applied, the improvements here are 0.4, 1.1 and 0.6 BLEU, METEOR and TER points, respectively (see output rows 3 and 5 of Table 3). Moreover, further addition of out-of-domain News data produces the best BLEU, METEOR and TER scores of 52.0,63.4 and 30.1, respectively (row 7). We also perform statistical significance test with MultEval (Clark et al., 2011). The systems that perform significantly better than the Baseline with $p < 0.05$ are marked with "$*$".

However, we note that for Twitter data, the sentiment preservation score (termed as "Sent_Pres.") is higher when using the SMT systems in combination with the sentiment classification approach (72.66% for the Twitter oracle data). Without the oracle sentiment analysis, sentiment preservation dips to 68% (with sentiment classification), but when sentiment classification is switched off altogether in the Baseline model, the score is reduced further to only 66.66%. When the Flickr data is made available as additional training data, similar behaviour is seen; if we look at row 5, we can see that the sentiment preservation score is a full 10% less (a 16% relative reduction) than in row 1. When all the data merged together, using sentiment classification produces the highest sentiment preservation score of 75.33% (see row 6).

As might be expected, dividing an already tiny Twitter parallel corpus into different parts for translation model training causes a degradation in MT quality, but not by much: just 2.1 BLEU points compared to the Baseline (see row 1 and 3). When Flickr data is added, the BLEU, METEOR and TER scores decrease by 2.2, 2.2 and 2.6 points, respectively, but the sentiment preservation score increases by 8.67% (from 62.66% to 71.33%). When all data are concatenated, the BLEU, METEOR and TER scores decrease here too but the sentiment preservation score increases from 73.33% to 75.33%.

The last row in Table 3 shows that the wrong MT engines[6] produces the lowest MT evaluation and sentiment preservation scores. As is well-known, using the phrase-table fill-up method can improve MT quality, as this is used to plug the gaps of the smaller in-domain MT system (Bisazza et al., 2011). Accordingly, we conduct experiments with an aim to increasing the translation quality and observing any accompanying degradation in sentiment polarity preservation. The results are shown in Table 4. It can be observed that the scores remain similar (almost no improvement) in all cases. The probable reason is that the addition of Flickr and News data adds certainty in terms of the probabilities in the phrase-table in the data concatenation approach, which do not effectively carry over in the phrase-table fill-up method. However, the sentiment preservation scores decrease in both cases. Additionally, Table 5 shows some of the interesting results obtained. We can compare the translations generated by combining outputs by sentiment classification with the translations produced using the Baseline model. Example 1 (the reference) is a tweet with negative sentiment but both of the two systems fail to produce proper translation because the word "terrible" which is the main word representing negative emotion still remains untranslated in both cases. In general,

Data	Fill-up	BLEU	METEOR	TER	Sent_Pres.
Twitter	×	48.2	59.4	34.2	**72.66%**
Flickr	✓	48.0	59.0	34.4	69.33%
News_Comm	✓	48.4	59.4	34.3	71.33%

Table 4: Experiment evaluation using fill-up method

Ex.	Reference	sentiment translation models	Baseline model
1	*Howard Webb is a terrible ref #WorldCup*	*Howard Webb is a schrecklicher ref #WorldCup*	*Howard Webb is a schrecklicher ref #WorldCup*
2	*injured Neymar out of World Cup 2014*	*verletzter Neymar out the WC2014*	*verletzter Neymar out of World Cup 2014*
3	*penalty shootouts are too intense !*	*penalty shoot is to intensiv !*	*penalties is to intensiv !*
4	*damn chile is nice !!!! #WorldCup*	*freeking Chile is good !!! #WorldCup*	*damn Chile is good !!! #WorldCup*
5	*a bit boring ...*	*a little boring ...*	*some boring ...*
6	*im with Germany*	*I stand to Deutschlands side*	*I stand to Germany's side*
7	*as getting I, GO CHILE !*	*completely mache I it GO CHILE !*	*as getting I, GO CHILE !*

Table 5: Comparison of translations by sentiment translation models and Baseline model

the Baseline model produces better translations as compared to sentiment-specific models (see examples 2, 4, 6 and 7 in Table 5). However, there are few cases where

[6]We perform a test by translating (i) negative tweets by positive model, (ii) neutral tweets by negative model, and (iii) positive tweets by neutral model. However, any of the different combination can be applied; our objective is to arbitrarily choose one of them and investigate the effect on translation and change in sentiment polarity

the sentiment-classified models outperforms the Baseline model (examples 3 and 5). This is a very interesting observation that can motivate the application of sentiment classification approach towards improving not only the sentiment preservation but also the MT quality for particular texts. Finally, Table 6 shows some results on how

Ex.	Reference	Right MT engine	Wrong MT engine
1	*little break on the #WorldCup for an amazing #Wimbledon final!*	*small Pause from the #WorldCup for a amazing #Wimbledon final!*	*kleine Pause of the #WorldCup for a erstaunliches #Wimbledon final!*
2	*yes !!!!!*	*yes !!!!!*	*so !!!!!*
3	*a bit boring ...*	*a little boring ...*	*some was ...*

Table 6: Comparison between sentiment polarities using the right and wrong MT engine

the sentiment polarity can change by using wrong MT engines. The tweet in example 1 with positive sentiment, when translated by a wrong MT engine produces an incomprehensible translation that makes it very difficult to identify its sentiment polarity. Furthermore, for the tweets in examples 2 and 3, using wrong MT engines produces semantically very much different translation from the reference and can not be assigned either of the positive or negative sentiment. These results imply that it is essential to translate the tweets by the MT engines conveying the same sentiment.

6. Conclusion and Future Work

In this paper we investigated the performance of the sentiment classification approach in order to measure the MT quality and sentiment preservation for CLSA. We propose a strategy of dividing the data used to train the Baseline SMT system into different subsets based on specific sentiment categories – positive, negative and neutral – to build a suite of sentiment translation engines. We showed that, despite a small deterioration in translation quality, the sentiment classification approach significantly improves sentiment preservation. We would argue that this trade-off is well worth making, especially in industrial sectors where it is critical that user sentiment in one (less spoken) language is accurately rendered when translated into the language of choice (typically, English). Further experiments also suggest that it is essential to carefully select the proper MT engine conveying the same sentiment polarity as that of the UGC in order to improve the accuracy of sentiment polarity preservation in the target language. In future, we would like to make use of the SA tools for both the source and the target languages and then apply our proposed approach. Another possibility is to further refine the sentiment classes with additional sentiment categories (strong positive, strong negative etc.,) in order to build more specific translation models and combine their output for evaluation.

Acknowledgements

This research is supported by Science Foundation Ireland in the ADAPT Centre (Grant 13/RC/2106) (www.adaptcentre.ie) at Dublin City University.

Bibliography

Afli, Haithem, Sorcha McGuire, and Andy Way. 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, 2017.

Araujo, Matheus, Julio Reis, Adriano Pereira, and Fabricio Benevenuto. 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, 2016.

Baker, Kathryn, Michael Bloodgood, Bonnie J. Dorr, Chris Callison-Burch, Nathaniel W. Filardo, Christine Piatko, Lori Levin, and Scott Miller. Modality and Negation in Simt Use of Modality and Negation in Semantically-informed Syntactic Mt. *Computational Linguistics*, 38(2):411–438, June 2012. ISSN 0891-2017.

Balahur, Alexandra and Marco Turchi. Multilingual Sentiment Analysis Using Machine Translation? In *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis*, pages 52–60, Jeju, Republic of Korea, 2012.

Balahur, Alexandra and Marco Turchi. Improving Sentiment Analysis in Twitter Using Multilingual Machine Translated Data. In *International Conference on Recent Advances in Natural Language Processing*, pages 49–55, Hissar, Bulgaria, 2013.

Bisazza, Arianna, Nick Ruiz, and Marcello Federico. Fill-up versus interpolation methods for phrase-based SMT adaptation. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 136–143, San Francisco, USA, 2011.

Clark, Jonathan H., Chris Dyer, Alon Lavie, and Noah A. Smith. Better Hypothesis Testing for Statistical Machine Translation: Controlling for Optimizer Instability. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies:*, pages 176–181, Portland, USA, 2011. Association for Computational Linguistics.

Esuli, Andrea and Fabrizio Sebastiani. SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining. In *In Proceedings of the 5th Conference on Language Resources and Evaluation*, pages 417–422, Genoa, Italy, 2006.

Gao, Dehong, Furu Wei, Wenjie Li, Xiaohua Liu, and Ming Zhou. Cross-lingual Sentiment Lexicon Learning with Bilingual Word Graph Label Propagation. *Computational Linguistics*, 41(1):21–40, Mar. 2015. ISSN 0891-2017.

He, Xiaonan, Hui Zhang, Wenhan Chao, and Deqing Wang. Semi-supervised Learning on Cross-Lingual Sentiment Analysis with Space Transfer. In *Proceedings of the IEEE First International Conference on Big Data Computing Service and Applications*, pages 371–377, Washington, DC, USA, 2015.

Jain, Sarthak and Shashank Batra. Cross Lingual Sentiment Analysis using Modified BRAE. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 159–168, Lisbon, Portugal, 2015.

Jiang, Jie, Andy Way, and Rejwanul Haque. 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*, pages 1–9, San Diego, USA, 2012.

Kanayama, Hiroshi, Nasukawa Tetsuya, and Watanabe Hideo. Deeper Sentiment Analysis Using Machine Translation Technology. In *Proceedings of the 20th International Conference on Computational Linguistics*, pages 494–500, Geneva, Switzerland, 2004.

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

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 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, 2007.

Lin, Zheng, Xiaolong Jin, Xueke Xu, Weiping Wang, Xueqi Cheng, and Yuanzhuo Wang. A Cross-Lingual Joint Aspect/Sentiment Model for Sentiment Analysis. In *Proceedings of the 23rd ACM International Conference on Information and Knowledge Management*, pages 1089–1098, Shanghai, China, 2014.

Miller, George A. WordNet: A Lexical Database for English. *Journal of Communications of the ACM*, 38(11):39–41, November 1995. ISSN 0001-0782.

Och, Franz Josef. 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, 2003.

Och, Franz Josef and Hermann Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29:19–51, March 2003. ISSN 0891-2017.

Saif, Mohammad M., Mohammad Salameh, and Svetlana Kiritchenko. How Translation Alters Sentiment. *Journal of Artificial Intelligence Research*, 55(1):95–130, January 2016. ISSN 1076-9757.

Scheible, Christian, Florian Laws, Lukas Michelbacher, and Hinrich Schütze. Sentiment Translation Through Multi-edge Graphs. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 1104–1112, Beijing, China, 2010.

Address for correspondence:
Pintu Lohar
pintu.lohar@adaptcentre.ie
ADAPT Centre, School of Computing, Dublin City University,
Glasnevin, Dublin 9,
Dublin, Ireland