FooTweets: A Bilingual Parallel Corpus of World Cup Tweets

Henny Sluyter-Gäthje[1], Pintu Lohar[1], Haithem Afli[1-2] and Andy Way[1]
[1]ADAPT Centre
School of Computing
Dublin City University
Dublin, Ireland
{FirstName.LastName}@adaptcentre.ie
[2] Cork Institute of Technology
haithem.afli@cit.ie

Abstract

The way information spreads through society has changed significantly over the past decade with the advent of online social networking. Twitter, one of the most widely used social networking websites, is known as the real-time, public microblogging network where news breaks first. Most users love it for its iconic 140-character limitation and unfiltered feed that show them news and opinions in the form of tweets. Tweets are usually multilingual in nature and of varying quality. However, machine translation (MT) of twitter data is a challenging task especially due to the following two reasons: (i) tweets are informal in nature (i.e., violates linguistic norms), and (ii) parallel resource for twitter data is scarcely available on the Internet. In this paper, we develop *FooTweets*, a first parallel corpus of tweets for English–German language pair. We extract 4,000 English tweets from the FIFA 2014 world cup and manually translate them into German with a special focus on the informal nature of the tweets. In addition to this, we also annotate sentiment scores between 0 and 1 to all the tweets depending upon the degree of sentiment associated with them. This data has recently been used to build sentiment translation engines and an extensive evaluation revealed that such a resource is very useful in machine translation of user generated content.

Keywords: tweets, parallel data, sentiment translation

1. Introduction

Due to the continuously developing Internet technology, there are countless digital media and social networking sites, all of which have a unique characteristic and purpose. Social media has evolved from being cyber world geek buzz to a massive platform for businesses, entrepreneurs, professionals and organizations that seek greater recognition and identification at a very economical price. However, business information sharing is not the only aspect of web services, e.g. people from all over the world with different cultural backgrounds stay connected and communicate via widely used social networking websites such as Twitter, Instagram, Facebook etc. Twitter is an extremely useful social networking tool for different event, business and news organisation that want to reach out to people (and are ready for a reply). Recently, Twitter has gained massive popularity and the number of Twitter users has increased significantly during the last few years.

However, Twitter users are often encouraged to write informal texts due to the 140-character limitation.[1] They follow many other users who tweet in different languages so, the tweets are multilingual in nature and often need to be translated from a specific (source-) language to the (target-) language of choice. In addition, tweets include spelling errors, hashtags, user handles, retweets, short forms etc. As a result, translation of such noisy texts becomes a difficult task. To the best of our knowledge, bilingual parallel corpora of tweets are hardly available on the Internet. The development of such corpus is therefore extremely important for MT of such noisy user-generated content (UGC). In this work, we extract 4,000 English tweets from the FIFA World Cup 2014 and apply the following steps in order to build the first bilingual parallel tweet corpus for the English–German language pair: (i) we translate all the English tweets into German with some translation guidelines (discussed in detail in Section 3), and (ii) each tweet is assigned a sentiment score between 0 and 1 depending upon the degree of emotion associated with it.

The remainder of this paper is organised as follows. Section 2 provides a brief literature survey of this field. In Section 3, we discuss some translation guidelines followed during the corpus development. The sentiment-annotation procedure is explained in Section 4. In Section 5, we briefly discuss the usefulness of this corpus with an example of our recent work on sentiment translation system using a the sentiment classification approach (Lohar et al. (2017)). Finally, we conclude and point out some possible future work in Section 6.

2. Related work

Parallel data for Twitter is scarcely available on the Internet. One of the available corpora is "microtopia", a parallel corpus of microblogs (Ling et al. (2014)). Recently, TweetMT (Vicente et al. (2016)) has been introduced as a parallel corpus of tweets in four language pairs that combine five languages (Spanish from/to Basque, Catalan, Galician and Portuguese). Ling et al. (2013) present a framework to crawl parallel data from microblogs in order to find parallel resources from single posts, with translations of the same sentence in two languages. Hajjem et al. (2013) create an Arabic–French comparable corpus, the first comparable corpus collected from Twitter. Despite this apparent lack of data, some work has been carried out in the area of tweet translation. Kaufmann and Kalita (2010) combine a statistical machine translation (SMT) system with a preprocessor and successfully remove the majority of noise from a tweet, which results in increasing its readability in the target lan-

[1] recently expanded to 280

guage. The work in Gotti et al. (2013) reports experimental results obtained from translating Twitter feeds published by agencies and organizations, using an SMT system. They mine parallel web pages linked from the URLs contained in English–French pairs of tweets in order to create the tuning and training material. Jiang et al. (2012) propose strategies to handle shortforms, acronyms, typos, punctuation errors, non-dictionary slang, wordplay, censor avoidance and emoticons.

3. Translation guidelines

This section describes the main guidelines we followed during the manual translation process. As Tweets may contain shortforms, typos, wordplays etc., all of which are often deliberately introduced especially due to the character limitation. Such characteristics of tweets pose challenges in translations into another language. We therefore place an emphasis on the following three strategies while translating the tweets: (i) informal to informal translation, (ii) informal to formal translation, and (iii) sentiment preservation. Following sections describe each of the main guidelines in detail.

3.1. Informal to informal translation

The tweets often contain informal texts such as short forms, stylistic effects etc. For example, the English tweet "GOAAAAL ♡ ♥ ♡ ♥" implies that the Twitter user expresses a positive emotion and deliberately introduces a stylistic effect (that is, the repetition of "A" in the word "GOAL") while writing the tweet. We place a strong emphasis on such behaviour by the users and translate the tweets accordingly into the target language, the result of which is essentially informal in nature as well. The above example tweet is therefore translated into German as "TOOOOR ♡ ♥ ♡ ♥" in order to retain the same degree of sentiment in the target language.

3.2. Informal to formal translation

As mentioned earlier in Section 3., the Twitter users are encouraged to use short forms at word or phrase level in order to fit all the contents within the specified characters limitation. Accordingly, most of the time they intentionally make acronyms for a group of words or a phrase. For example, nowadays it has become more popular to write *lol* instead of writing *laughed out loud*. In a similar vein, the phrase "going to" is often shortened to *gonna*. In addition to this, Twitter users often create short forms of individual words by omitting one or more characters from them. For example, the word "*you*" is contracted to *u* by removing the letters "*y*" and "*o*" so that it sounds almost the same as the original word. Such behaviour challenges in the translation process. It is therefore necessary to scan through the tweets carefully and identify such *noisy* content. Once these elements are found, we translate them with special attention so that their translations are formal. Therefore, the informal "u" is translated as formal "dir" in German. Such *informal-to-formal* translation is definitely useful in building MT engines in order to make the translation of tweets easier, which is otherwise a difficult task.

3.3. Sentiment preservation

As many of the tweets convey a certain degree of sentiment, they draw our special attention during the translation process maintaining the original sentiment. In addition, the deliberate stylistic effects applied on the tweets encourage us to perform the translation accordingly. For example, the tweet "YEEEEEESSSS!!!" contains a higher level of positive sentiment than the tweet "YES!!!". Considering this phenomenon, we translate the tweets based not only on their literal meaning but also the way they are expressed. The above example, therefore, can be translated as "JAAAAAAA!!!" into German.

4. Sentiment score annotation

Once the translation is performed on all the 4,000 English tweets, we manually assign sentiment scores (from 0 to 1)to each of them. However, as our intention was to categorise the tweets into three different classes namely *negative*, *neutral* and *positive*, we categorise them using the following criteria: (i) *negative*, if sentiment score < 0.5, (ii) *neutral*, if sentiment score $= 0.5$ and (iii) *positive*, if sentiment score > 0.5 .

With the above criteria for categorisation, it is perfectly valid to assign any sentiment score from 0 to less than 0.5 for the negative-sentimented tweets as it does not affect the sentiment class as long as the score remains in this range. In a similar manner, the tweets that convey positive sentiment can be assigned any score that is greater than 0.5 but is less than or equal to 1. With this consideration, we decided to use the following three different values for sentiment scores to make this task easier: (i) 0.3 for negative tweets, (ii) 0.5 for neutral tweets, and (iii) 0.7 for positive tweets.

Note that this categorisation technique is valid only if there are only these three sentiment classes. It may be required to decide on other values if we include other sentiment classes such as *strong negative*, *strong positive* etc. However it is not required in this case as our main focus is not on the exact sentiment score but only on these three sentiment classes.

Table 1 shows some example translations along with the sentiment scores assigned to them. There are three different values for sentiment scores (0.3, 0.5 and 0.7) that depend upon the sentiment class the tweet belongs to. Example 1 in this table is a negative-sentimented tweet and hence is assigned a score of 0.3 according to our criteria of sentiment annotation. We can see that "I am" is shortened as "Im" but it is translated as "bin ich" in German, which is essentially informal-to-formal translation. Similarly in example 2 (positive-sentimented tweet), the phrase "going to" is informally written as "gonna". Note that the first segment is "not gonna lie..." where the pronoun "I" is missing but it is obvious from the context. This informal segment is translated into "ich werde nicht lügen" which is formal in German. Subsequently the word "next" is contracted to "nxt" (with a character omission) but translated as "nächstes", a formal word in German. However, sometimes users write almost-formal texts in their tweets. For example, the item number 4 in Table 1 is an example where all the words are written correctly. This tweet is grammatically correct except that the verb "is" is missing (i.e., it should be "Luis

Ex.	English tweet	German translation	Sentiment score
1	now that Neymar cant play, Im so nervous for Brazil	jetzt, da Neymar nicht spielen kann, bin ich so aufgeregt wegen Brasilien	0.3
2	not gonna lie... Germanys national anthem is awesome.	ich werde nicht lügen... Deutschlands Nationalhymne ist genial.	0.7
3	sorry Brazil.... dont host a worldcup nxt time	tut mir leid Brasilien.... veranstalte nächstes Mal keine Weltmeisterschaft	0.3
4	Luis Suárez suspended for nine matches and banned for four months from any football-related activity	Luis Suárez für neun Spiele gesperrt und vier Monate von jeder Fussballtätigkeit ausgeschlossen	0.3
5	shame on Messi.. wack Argentine team.. no clear cut chance created at all...	Schande über Messi.. Schwaches argentinisches Team.. überhaupt keine klaren Chancen erspielt...	0.3
6	just making sure its still there!!!	Stelle nur sicher, dass es immer noch da ist!!!	0.5
7	Yesssss!!!!!!!!!!!! Golazo!!!!!!!	Jaaaaa!!!!!!!!!!!! Golazo!!!!!	0.7

Table 1: Tweets and their translations along with the sentiment scores

Suárez *is* suspended..."). It can be observed that the example 6 is a tweet that can be considered as having neutral sentiment as it is very difficult to associate either negative or positive sentiment with it. Finally, example 7 contains stylistic effect as the word "Yes" is deliberately written as "Yesssss" in order to express the positive emotion of the user. As expected, this word is translated as "Jaaaaa" into German to retain its positivity during the translation process. This is an example of *informal-to-informal* translation which we specially consider for sentiment preservation. Upon completing the translation of all the tweets, we obtain a distribution of the three sentiment classes. Table 2 shows the distribution of negative, neutral and positive tweet pairs.

Negative	Neutral	Positive	Total
1,019	1,408	1,573	4,000

Table 2: Data distribution

5. Usefulness of the corpus

According to the best of our knowledge, the data we developed in this work is the first ever published parallel Twitter corpus for English–German language pair. Although it is a very small-sized corpus having only $4,000$ tweet pairs, it can play a significant role in building MT engines as it contains different levels of informal parallel texts. It is, therefore, expected that the MT models built from it are capable of translating many informal texts although not everything, as it is extremely difficult to cover all types of variations of a word. In addition, annotating sentiment scores to all the tweets opens up a number of opportunities in future for sentiment analysis of the tweets as well. This corpus is useful in translating tweets and at the same time preserving the sentiment during translation by building a suite of sentiment translation engines (Lohar et al. (2017)). In the following section, we briefly discuss this work in order to highlight the importance of our corpus.

5.1. Tweet translation and sentiment preservation

The work in Lohar et al. (2017) presents a sentiment translation system based on a sentiment classification approach. The idea is to divide the English–German parallel twitter corpus into three different parts based on the following characteristics of the tweets: (i) negative corpus with sentiment score ≤ 0.4, (ii) neutral corpus with sentiment score $\simeq 0.5$, and (iii) positive corpus with sentiment score ≥ 0.6 . Afterwards, three different translation models are built from each of the above parallel corpus set and referred to as negative, neutral and positive translation model, respectively. The objective to translate the tweets using sentiment-specific translation models and at the same time preserve the sentiment during translation. They translate German tweets into English in order to be able to use the sentiment analysis tool in English especially designed for tweets (Afli et al. (2017)).

5.1.1. Experiments

As the corpus is small in size, a very small subset of 50 tweets per sentiment (negative, neutral and positive) is held out for tuning and testing purposes in order to maintain as large an amount as possible for training purpose. The remaining $3,700$ tweet pairs are considered as the training data and are similarly divided into negative, neutral and positive tweet pairs. The translation models are built using the Moses SMT tool (Koehn et al. (2007)) using Giza++ (Och and Ney (2003)) for word and phrase alignment. Afterwards, the models are tuned using minimum error rate training (Och (2003)). The additional resources used are English–German parallel Flickr data[2] and "News-Commentary (News)" data[3] in order to build larger MT engines. The evaluation process consists of two different types of measurements: (i) MT quality and (ii) sentiment preservation. For MT evaluation, the automatic evaluation metrics BLEU (Papineni et al. (2002)), METEOR (Denkowski and Lavie (2014)) and TER (Snover et al.

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

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

Translation model	Sent_Clas	BLEU	METEOR	TER	Sent_Pres
Twitter	✓	48.2	59.4	34.2	72.66%
Twitter (Baseline)	×	50.3	60.9	31.9	66.66%
Twitter + Flickr + News	✓	50.3	62.3	31.0	**75.33%**
Twitter + Flickr + News	×	**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

Reference	**Sentiment translation system**	**Baseline**
Bosnia and Herzegovina really got f*** over man	Bosnia and Herzegowina eliminated echt demolished	Bosnia and Herzegovina were a abgezogen
when USA lost, but were still moving onto the next round	even if USA today we in the next round	could usa loses the next round
Brazil 5 WorldCup championship Argentine 2 WorldCup championship so Ill go with Brazil	Brazil 5 time world champion Argentina 2 time world champion so Im for Brazil	Brazil 5 time world champions Argentina 2 time world champions so for Brazil

Table 4: Examples where sentiment is altered by the Baseline system

Reference	**Right MT engine**	**Wrong MT engine**
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!
yes !!!!!	yes !!!!!	so !!!!!
a bit boring...	a little boring ...	some was ...

Table 5: Comparison between sentiment polarities using the right and wrong MT engine

(2006)) are used. In order to measure the sentiment preservation, the fraction of all of the source-language tweets in the test set that remain under the same sentiment class after translation is calculated.

In addition, the authors also performed a random test by translating the (i) negative tweets by the positive model, (ii) neutral tweets by the negative model, and (iii) positive tweets by the neutral model. The aim was to arbitrarily choose one of any of the model-selection combinations so that the tweets with a specific sentiment class is translated by the translation model with a different sentiment, in order to see the effects on translation quality and sentiment preservation. The resultant translation system was termed as the "Wrong MT engine" whereas the "Right MT engine" was their sentiment translation system.

5.1.2. Results

The results are summarised in Table 3. It can be observed that when only the Twitter data is used, better BLEU, METEOR and TER scores are obtained without using the sentiment classification ("Sent_Clas") approach ('Twitter (Baseline)'). In contrast, the sentiment preservation score ("Sent_Pres") is higher when using the sentiment classification (72.66%) method whereas switching it off causes the score to be reduced to 66.66%. The best BLEU, METEOR and TER scores of 52.0, 63.4 and 30.1, respectively, are obtained with the concatenation of additional Flickr and News data. The sentiment classification approach still manages to increase the sentiment better in this case too (from 73.33% to 75.33%). The last row in this table shows that the wrong MT engines produce the lowest scores both in terms of MT quality and sentiment preservation.

Table 4 shows how the sentiment translation system is capable of preserving the sentiment in the target language whereas the Baseline alters the sentiment during translation. Finally, Table 5 highlights the fact that the sentiment polarity is changed by using the wrong MT engines. This is a very interesting result which suggests that it is essential to translate a specific text by using the translation system that is built from the data whose sentiment matches the input text.

6. Conclusion and Future Work

In this work, we developed the first corpus (*FooTweets*) of English–German parallel tweets. We followed some translation guidelines that are very important for translating such noisy texts. In addition to this, we manually annotated the sentiment scores for all the 4,000 tweets in order to facilitate the task of sentiment analysis. Initially we restricted the sentiment classes to only negative, neutral and positive. However, in future, it can easily be extended with some other sentiment classes such as strong negative, strong positive etc. Although these processes require a significant amount of time, in future, we would like to increase the size of our corpus as it will help improve the quality of the Twitter translation engines. We have made this corpus publicly available for access(*FooTweets*[4]). We hope that this parallel resource will be helpful for the researchers who are interested in the area of MT and sentiment translation systems. It may also open up a number of opportunities for future

[4] Available at: `https://github.com/HAfli/FooTweets_Corpus`

work for other natural language processing tasks related to UGC.

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

Haithem Afli, 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.

Michael Denkowski and Alon Lavie. 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, 2014.

Fabrizio Gotti, Philippe Langlais, and Atefeh Farzindar. 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, 2013.

Malek Hajjem, Maroua Trabelsi, and Chiraz Latiri. Building comparable corpora from social networks. In *Proceedings of the 7th Workshop on Building and Using Comparable Corpora*, pages 11–16, Reykjavik, Iceland, 2013.

Jie Jiang, 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.

Max Kaufmann 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.

Philipp Koehn, 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.

Wang Ling, Guang Xiang, Chris Dyer, Alan Black, and Isabel Trancoso. Microblogs as parallel corpora. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 176–186, Sofia, Bulgaria, 2013.

Wang Ling, Luis Marujo, Chris Dyer, Alan Black, and Isabel Trancoso. Crowdsourcing high-quality parallel data extraction from twitter. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 426–436, Baltimore, USA, 2014.

Pintu Lohar, Haithem Afli, and Andy Way. Maintaining sentiment polarity in translation of user-generated content. *The Prague Bulletin of Mathematical Linguistics*, 108(1):73–84, 2017.

Franz Josef Och. 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.

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

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 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, 2002.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. A study of translation edit rate with targeted human annotation. In *Proceedings of Association for Machine Translation in the Americas*, pages 223–231, Massachusetts, USA, 2006.

Iñaki San Vicente, Iñaki Alegria, Cristina España-Bonet, Pablo Gamallo, Hugo Gonçalo Oliveira, Eva Martínez Garcia, Antonio Toral, Arkaitz Zubiaga, and Nora Aranberri. Tweetmt: A parallel microblog corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, pages 2936–2941, Portorož, Slovenia, 2016.