A Systematic Comparison Between SMT and NMT on Translating User-Generated Content

Pintu Lohar[1], Maja Popovic[1], Haithem Afli[1,2] and Andy Way[1]

ADAPT Centre
School of Computing
Dublin City University
Dublin, Ireland
[1]{FirstName.LastName}@adaptcentre.ie
[2]haithem.afli@cit.ie

Abstract. Twitter has become an immensely popular platform where the users can share information within a certain character limit (280 characters) which encourages them to deliver short and informal messages (tweets). In general, machine translation (MT) of tweets is a challenging task. However, for translating German tweets about football into English, it has been shown that a moderate translation performance in terms of the BLEU score can be achieved using the phrase-based translation engines built on a tiny parallel Twitter data set [1]. In this work, we propose to further increase the translation quality using the neural machine translation models and applying the following strategies: (i) we back translate a set of out-of-domain English tweets released by "Harvard data set" in 2017 into German and add the synthetic parallel data to the tiny parallel data used in [1]; (ii) as tweets are short in general, we extract short text pairs from the large news-commentary parallel data and add it to the tiny Twitter parallel data set in order to restrict the length of the out-of-genre text segments. We build both phrase-based and neural MT systems (PBMT and NMT) using the above data combinations in order to perform a systematic comparison between the two approaches on translating tweets. Our experimental results reveal that the NMT system performs significantly worse than the PBMT system when using only the tiny Twitter data set for MT training. In contrast, when additional data is used for training, the results show huge improvements of the NMT system and produce very similar BLEU scores as the PBMT system even with only few hundred thousands of additional synthetic parallel data.

1 Introduction

The world of social media has undergone drastic changes during the last few years. The generation and sharing of information have become much easier than before with the advent of popular social media platforms such as Twitter, Facebook, etc. Recently, Twitter has become very popular because of its unique features. On Twitter, the users spread information in form of short and informal messages (tweets) with maximally 280 characters. Such character limitation encourages users to write short messages, although sometimes they do it deliberately. In general, tweets are noisy in terms of linguistic norms. Usually, this noise does not pose problems for human understanding of tweets, but it creates challenges for machine translation (MT). Another challenge for machine translation is sparseness of bilingual (translated) tweets, because the performance of MT systems depends on amount of bilingual training data. In this work, we perform a systematic comparison between statistical phrase-based MT (PBMT) and neural MT (NMT) of tweets using different amounts and types of training corpora. We translate the German football tweets into English by PBMT and NMT systems trained on in-domain *FIFA 2014* English–German tweet pairs [2], out-of-domain *Harvard* data set of English tweets [3][1], as well as out-of-domain[2] and out-of-genre[3] short text segments from *news-commentary* English–German parallel corpus[4].

The in-domain *FIFA 2014* data set contains only $4,000$ tweet pairs and this is certainly not enough for MT training. Therefore we accompany it with the *Harvard* data set of English tweets. However, these tweets are available only in English, they are not translated into German. Therefore, we translate these English tweets into German by an English-to-German MT system thus creating the synthetic parallel data. The MT system mentioned above is trained on the combination of the small twitter data set and the short text pairs form the *News* data. The reason behind including only the short *News* texts is that the translation model built on the combination of the Twitter data and the whole *News* data produced worse translations in our earlier attempts. Morevear, subsequently adding *Europarl* corpus[5] further worsen the translation quality. However, this back-translation process generates only around $50k$ of additional parallel segments which is still not sufficient enough for building MT engines. We therefore further add short parallel segments from *news-commentary* English–German parallel corpus[6] as mentioned earlier in this section. These short text pairs are extracted according to the length of English–German tweet pairs so that the out-of-genre data resembles our in-domain data at least in terms of segment length. All of the above data are used in different combinations to build a suite of PBMT and NMT systems. Our experimental evaluation reveals that the NMT system built on the smallest parallel tweet corpus performs much worse than the PBMT system trained on this data, as anticipated. However, successive addition of synthetic data and the News data set reduces the gap significantly. Most importantly, our best NMT system performs on par with the best performing PBMT system even by adding only a few hundred thousands of additional synthetic data and short text pairs from the *News* data set. The

[1] `https://dataverse.harvard.edu/dataset.xhtml?id=3047332`

[2] not football related

[3] They are not user generated content

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

[5] `www.statmt.org/europarl/v7/de-en.tgz`

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

remainder of this paper is organised as follows. Section 2 highlights on the history of related works in this field of study. We describe in details about our research goals in Section 3 followed by the experimental setups discussed in Section 4. The evaluation results are illustrated in Section 5. Finally, we conclude the main contribution of this work and point out the further avenues of research in Section 6.

2 Related Work

A considerable amount of work has been done on social media analysis recently, especially the sentiment analysis and the translation of user-generated content. In [4] they evaluate 28 top academic and commercial systems in tweet sentiment classification across five distinctive data sets in order to assess Twitter sentiment analysis. A deep learning approach has been proposed to perform sentiment analysis of tweets for predicting polarities at both message and phrase levels [5]. They use an unsupervised neural language model to train initial word embeddings that are further tuned by a deep learning model on a distant supervised corpus. A number of research works also attempt to translate user-generated contents. For example, Twitter translation of microblog messages from the Twitter domain by using a translation-based crosslingual information retrieval system is done in [6]]. They find relevant Arabic Twitter messages given English Twitter queries, and apply a standard pipeline for unsupervised training of PBMT to retrieval results.

[7] proposed a framework by using PBMT system for translating Arabic User-Generated Content by integrating an error correction system prior to the translation phase. Some papers investigate translating tweets in order to map sentiment labels to the target language and be able to perform the sentiment analysis in this language [[8], [9], [10]]. Moreover, several researchers attempted to build parallel corpus for user-generated content, since the lack of large parallel corpora represents one of the major challenges for translation. For example, [11] crawl a considerable amount of parallel sentence pairs from micro-blogs and release the data publicly. They extract over 1M Chinese-English parallel segments from Sina Weibo (the Chinese counterpart of Twitter) using only their public APIs. Their extracted parallel data yields improvements in translating microblog text and edited news commentary. Researchers also employ the automatic collection and crowd-sourcing approaches to build a parallel corpus of Tweets such as *TweetMT* [12]. The information retrieval method is also employed in translating hashtags in Twitter [13]. [1] investigated a suite of MT systems for sentiment translation trained on a small bilingual Twitter data and attempted to preserve the sentiment of tweets with a loss in translation quality. However, they do not report results for the state-of-the-art NMT approach, only for PBMT systems. Scarce training data are even more challenging for the NMT approach, as well as discrepances between training and test domains [14]. A number of recent publications compare PBMT and NMT approaches systematically by performing error analysis and identifying main advantages and disadvantages of each ap-

proach (e.g. [15], [16]). However, no results on UGC content, neither on scarce resource scenario were reported. This work investigates tweet translations in the scarce resource scenario by the NMT approach, compares the NMT and PBMT approaches, and investigates the usage of different training corpora for both MT systems.

3 Research goals

The main goals of our work are

(i) to compare machine translations of tweets using two MT approaches, namely phrase-based and neural MT

(ii) to compare the usage of different amounts and types of training corpora for each of the two approaches

Phrase-based approach for machine translation had been state-of-the-art for many years. The neural approach has recently emerged as the first technology able to challenge the long-standing dominance of phrase-based approaches. In PBMT, different models (translation, reordering, target language, etc.) are trained independently and combined in a log-linear scheme in which a tuning algorithm assigns a different weight to each model. On the contrary, in NMT all the components are jointly trained to maximise translation performance. NMT systems have a strong generalisation power, and are better capable of modelling long-distance phenomena. In only three years, the NMT approach has surpassed the performance of PBMT in majority of aspects, especially regarding fluency. However, whereas for PBMT it is possible to achieve decent MT translations even with small amounts of parallel texts, NMT is much more sensitive to the amount of training data. Therefore they usually perform worse than PBMT in low-resource settings, and also show lower performance on out-of-domain data.[14]. Back-translation [17] has become a widely used approach to augment the training corpora for NMT using monolingual data: a set of data in the target language is translated by a target-to-source MT system, and these translations are then used as the missing source language text for training.

To the best of our knowledge, no systematic comparison in this direction has been performed for translating Tweets. Therefore, we investigate the following scenario for translating German football tweets into English:

scarce in-domain parallel corpus Train PBMT and NMT systems on a very small in-domain parallel text, which contains the same type of text as our development and test data. The advantage of this corpus is that corresponds to the texts which should be translated. The disadvantage is that the corpus is very small.

adding out-of-domain back-translated corpus Add more tweets to the corpus, albeit about a different topic. For this scenario, an additional challenge is the fact that the tweets are available only in English, there is no bilingual parallel corpus. Therefore, we use the back-translation strategy – we use an existing English-to-German PBMT system trained on the small Twitter parallel corpus and a part of *News* data (explained in Section 1) to translate the English tweets into German. This synthetic parallel corpus is then added to the original in-domain corpus. The advantage of the corpus is that it is larger than the original corpus, and it belongs to the same genre as the test, i.e. also contains tweets. The disadvantage is that it contains different topic (no football). Also, despite of being larger than the original corpus, it is still rather small.

adding out-of-domain and out-of-genre parallel corpus In order to further increase the size of the training corpus, we use the News parallel texts, which differ from the development and test sets both in terms of domain (topics) as well as in terms of genre (style). Because of these discrepancies, we select only the part of the corpus with segment lengths similar to the lengths of tweets (i.e. short texts). To find the length, we examine all the tweets and found that the lengthiest tweet (in terms of number of words) consists of 32 words. We therefore consider the *News* texts that contains equal to or less than 32 words. The advantage of this corpus is its size. The disadvantage is that it does not correspond to the test data either regarding topic (domain) or regarding style (genre).

4 Experimental Set-up

4.1 MT systems

PBMT system In order to build the PBMT models, we use the open-source phrase-based statistical translation tool called Moses [18], which uses Giza++ [19] for word and phrase alignment. We build 3-gram language models using the SRILM toolkit [20]. The maximum phrase length for phrase-based training is set to 7. in our experiments. The models are tuned using minimum error rate training [21].

NMT system Our NMT engines are built by using the freely available open source NMT toolkit called *OpenNMT*[7] [22]. We use the default parameter settings of *OpenNMT*, such as, *RNN* as the default type of encoder and decoder, $word\_vec\_size = 500$, $training\_steps = 100,000$ and so on. In fact, the parameter values can be used in numerous combinations and each combination may lead to different result. In this work, we have not explored other combinations, which can be done in future in order to investigate the difference in results (if any). However, the default parameter settings we use is one of the most applied optimal settings.

[7] `https://github.com/OpenNMT/OpenNMT-py`

4.2 Data sets

Table 1 shows the number of segments in the three data sets used for the experiments. As the $4,000$ tweet pairs is clearly an extremely small corpus for MT training, we held out only small amounts of data for development and test purposes because we wanted to keep as much data as possible for training. Therefore, we used $3,000$ segments for training, only 500 for development and 500 for test purposes. The additional data consists of followings.

Harvard data set This data set contains English tweets collected by crawling Twitter's REST API using the Python library *tweepy 3*[8] . The tweets belong to the 20 most popular twitter users (with the most followers) such as *Katy Perry*, *Barack Obama*, etc.

Short news texts The out-of-domain *News* data consists of about $216,000$ short segments. These short text segments consists of up to 32 words (as explained in Section 3).

Table 1. Data sets

Data set	**#Segments**
Twitter Football (in domain)	$3,000$
Development	500
Test	500
Twitter Harvard (out of domain)	$52,542$
News (out of genre and domain)	$216,742$

The data combinations used for training the PBMT and NMT systems is shown in Table 2. The smallest PBMT and NMT systems are built on the initial $3,000$ in domain football-related tweet pairs. Afterwards, these in domain tweets are successively accompanied with the *Harvard* data in order to build the larger translation models. Finally, the *News* data is further added for training the largest translation models of our experiments. In this work, our purpose to include the short *News* texts is to investigate the effects on translation quality when similar type of (in terms of length) out-of-domain data is used to accompany the much smaller in-domain parallel data.

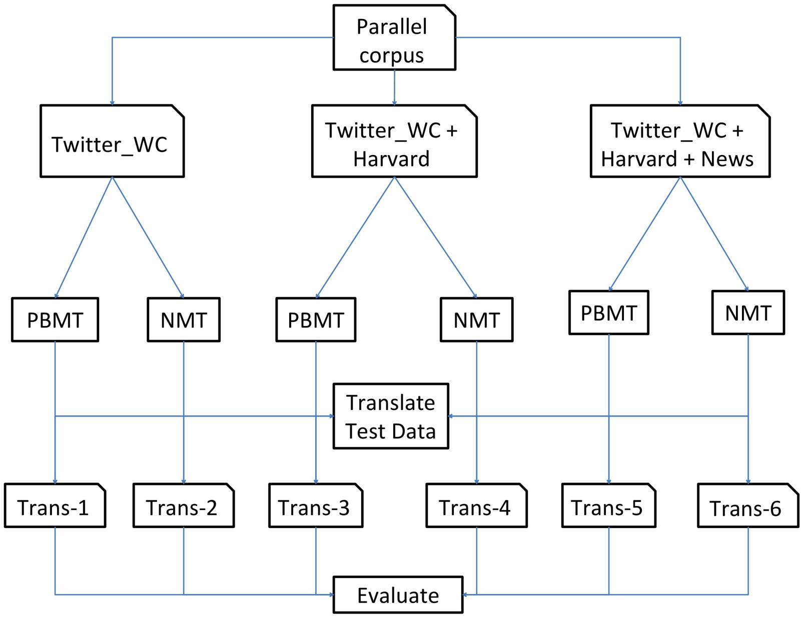
Figure 1 shows the whole system architecture of our experiments. It can be seen from the figure that we use three different data combinations as follows: (i) Twitter data (*Twitter_WC*), (ii) the combination of Twitter and the *Harvard* data set (*Twitter_WC* +*Harvard*), and (iii) Twitter, *Harvard* and the short text pairs of *News* data altogether (*Twitter_WC* +*Harvard* + *News*). Both the PBMT and NMT models are built on each of these data combinations. As a

[8] `https://github.com/felHR85/Tweepy-3`

Table 2. Data combinations used for PBMT and NMT training

Training Data	**#Segments**
Twitter_WC	3,000
Twitter_WC+Harvard	55,542
Twitter_WC+Harvard + News	272,284

Fig. 1. Illustration of the training of three PBMT and three NMT systems using different bilingual parallel corpora.



result, a total of 6 different models are built. Each of them is used to translate the test data and hence 6 different translations are produced. Finally, we evaluate all the translation outputs by using the *BLEU*, *METEOR* and *TER* metrics as explained in the following section.

4.3 Evaluation

We assess the MT performance using three widely spread automatic measures: BLEU [23], Meteor [24] and TER [25]. All metrics produce a numeric score based on similarity between the given MT output (hypothesis) and the corresponding human reference translations.

BLEU is calculated as precision of translated n- grams (sequences of n words) by comparing them with n-grams in reference translation. BLEU scores range between 0 (complete mismatch) and 100 (perfect match with the reference translation). However, it should be taken into account that there is no unique correct translation so a perfect exact match to one particular human translation is very hard to generate. Therefore, the BLEU scores of high quality (rated by humans) translations usually do not go over 60-70.

METEOR is based on the harmonic mean of precision and recall, whereby recall is weighted higher than precision. Along with standard exact word (or phrase) matching, it has additional features, i.e. stemming, paraphrasing and synonymy matching. The range of METEOR scores is the same as for BLEU.

TER is based on edit distance between translation hypothesis and the reference. In addition to standard substitutions, deletions and insertions, it takes into account shift cost for reordering operation. The number of operations is normalised with the length of the reference translation. Contrary to BLEU and METEOR, the TER score reflects mismatch between two texts, 0 being the perfect match. For the same reasons mentioned in the BLEU description, TER scores rarely reach zero.

5 Results

Table 3. BLEU, METEOR and TER scores for each of the six MT outputs generated by two MT approaches and three bilingual training sets.

MT approach	Training	BLEU	METEOR	TER
PBMT	Twitter_WC	46.6	39.1	33.5
	Twitter_WC+Harvard	48.6	41.4	30.9
	Twitter_WC+Harvard + News	**50.0**	**42.2**	29.9
NMT	Twitter_WC	0.8	6.8	88.4
	Twitter_WC+Harvard	45.0	38.8	34.7
	Twitter_WC+Harvard + News	**50.0**	41.9	**29.6**

The results in form of the BLEU, METEOR and TER scores are shown in Table 5, and the following trends can be observed:

- PBMT with different training data
  - training on the extremely scarce in-domain training corpus already enables decent scores
  - all scores are moderately improving with increase of the training corpus.

- NMT with different training data
  - training of an NMT system with a scarce in-domain corpus is practically useless
  - adding back-translated tweets improves the system largely so that it reaches the performance of the PBMT system training on the scarce corpus only
  - adding News data further improves the system and the scores become same as for the PBMT system trained on the same full corpus

- PBMT vs. NMT in different settings
  - PBMT performance is better for very scarce training data – in accordance with findings in previous work koehn
  - the more training data is used (even though being back-translated, out-of-domain), the closer is the performance of the two approaches

6 Conclusions and Future Work

This paper demonstrates the comparative study of PBMT and NMT systems for translating specific type of user-generated content, in this case, twitter data about football World Cup. Our major contributions regarding our research goals are:

(i) when trained on about 270k segments, NMT and PBMT performance are on the same level in terms of automatic MT metrics

(ii) using smaller amounts of training data significantly deteriorates the performance of the NMT system, and moderately deteriorates the performance of the PBMT system

The NMT system trained on the tiny Twitter corpus is practically useless even though all the data are in-domain compared to the test set. Back-translated data improved the system largely, and adding more out-of-domain data improved further more. Such improvement confirms that NMT engines are very data hungry and can perform much better when more data is supplied for training, even though these data are coming from different domain and/or genre, or the source part is artificial (back-translated). Considering the fact that parallel Twitter data for MT training are scarcely available, our experiments show potentials for creating additional parallel corpora for Twitter by employing back-translation and inclusion of out-of-domain parallel resources. Our best performing NMT system is built on the combination of only $3k$ in-domain tweets, $50k$ of back-translated *Harvard* tweets and $200k$ short text pairs from *News* data, which is still scarce. This opens up a number of possibilities for further improving the NMT systems for translating tweets, such as investigating back-translation of tweets and different scenarios for including out-of-domain data. As we are aware that parallel Twitter data for MT training are scarcely available, we also look forward to incorporating other forms of user-generated contents such as customer feedback, reviews etc with the tiny parallel Twitter data used in this work. As of now, we have explored only the *News* texts as an out-of-domain data. Morever, as the *Europarl* corpus is a fix-domain and did not work well for our experiments, we plan to utilise other types of mix-domain parallel resource such as *common crawl* corpus[9] in order to extend our work.

7 Acknowledgments

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

References

1. Lohar, P., Afli, H., Way, A.: Maintaining sentiment polarity in translation of user-generated content. The Prague Bulletin of Mathematical Linguistics **108** (2017) 73–84
2. Sluyter-Gäthje, H., Lohar, P., Afli, H., Way, A.: Footweets: A bilingual parallel corpus of world cup tweets. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). (2018)
3. Bin Tareaf, R.: Tweets dataset - top 20 most followed users in twitter social platform (2017)
4. Zimbra, D., Abbasi, A., Zeng, D., Chen, H.: The state-of-the-art in twitter sentiment analysis: A review and benchmark evaluation. ACM Trans. Manage. Inf. Syst. **9** (2018) 5:1–5:29
5. Severyn, A., Moschitti, A.: Twitter sentiment analysis with deep convolutional neural networks. In: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '15 (2015) 959–962
6. Jehl, L., Hieber, F., Riezler, S.: Twitter translation using translation-based cross-lingual retrieval. In: Proceedings of the Seventh Workshop on Statistical Machine Translation. WMT '12 (2012) 410–421
7. Afli, H., Aransa, W., Lohar, P., Way, A.: From arabic user-generated content to machine translation : Integrating automatic error correction. (2016)
8. Peisenieks, J., Skadiņš, R.: Uses of machine translation in the sentiment analysis of tweets. (2014)
9. Balahur, A., Turchi, M.: Multilingual sentiment analysis using machine translation? In: Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis. WASSA '12 (2012) 52–60

[9] `http://www.statmt.org/wmt13/training-parallel-commoncrawl.tgz`

10. Balahur, A., Turchi, M.: Improving sentiment analysis in twitter using multilingual machine translated data. In: RANLP. (2013)
11. Ling, W., Xiang, G., Dyer, C., Black, A., Trancoso, I.: Microblogs as parallel corpora. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics (2013) 176–186
12. naki San Vicente, I., naki Alegria, I., na Bonet, C.E., Gamallo, P., Oliveira, H.G., Garcia, E.M., Toral, A., Zubiaga, A., Aranberri, N.: Tweetmt: A parallel microblog corpus. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). (2016)
13. ISLA, S.C.: hashtags : Joint translation and clustering. (2011)
14. Koehn, P., Knowles, R.: Six challenges for neural machine translation. In: Proceedings of the First Workshop on Neural Machine Translation, Vancouver, Canada, Association for Computational Linguistics (2017) 28–39
15. Toral Ruiz, A., Sánchez-Cartagena, M.: A multifaceted evaluation of neural versus phrase-based machine translation for 9 language directions. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2017), Valencia, Spain, Association for Computational Linguistics (ACL) (2017) 1063–1073
16. Bentivogli, L., Bisazza, A., Cettolo, M., Federico, M.: Neural versus phrase-based machine translation quality: a case study. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP 2016), Austin, Texas, Association for Computational Linguistics (2016) 257–267
17. Sennrich, R., Haddow, B., Birch, A.: Improving neural machine translation models with monolingual data. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, (ACL 2016), Berlin, Germany (2016)
18. 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., Herbst, E.: Moses: Open source toolkit for statistical machine translation. In: ACL. (2007)
19. Och, F.J., Ney, H.: A systematic comparison of various statistical alignment models. Computational Linguistics **29** (2003) 19–51
20. Stolcke, A.: Srilm - an extensible language modeling toolkit. In: INTERSPEECH. (2002)
21. Och, F.J.: Minimum error rate training in statistical machine translation. In: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics, Sapporo, Japan (2003) 160–167
22. Klein, G., Kim, Y., Deng, Y., Senellart, J., Rush, A.M.: OpenNMT: Open-source toolkit for neural machine translation. In: Proc. ACL. (2017)
23. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: A method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL 2002). ACL '02, Philadelphia, Pennsylvania, Association for Computational Linguistics (2002) 311–318
24. Banerjee, S., Lavie, A.: Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In: Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, Association for Computational Linguistics (2005) 65–72
25. Snover, M., Dorr, B., Schwartz, R., Micciulla, L., Makhoul, J.: A study of translation edit rate with targeted human annotation. In: In Proceedings of Association for Machine Translation in the Americas (AMTA 2006), Cambridge, MA (2006) 223–231