Making the Most of Synthetic Parallel Texts: Portuguese-Chinese Neural Machine Translation Enhanced with Back-Translation

Rodrigo Santos(✉), João Silva, and António Branco

NLX—Natural Language and Speech Group, Department of Informatics,
Faculdade de Ciências, University of Lisbon, Campo Grande,
1749-016 Lisbon, Portugal
{rsdsantos,jsilva,antonio.branco}@di.fc.ul.pt

Abstract. The generation of synthetic parallel corpora through the automatic translation of a monolingual text, a process known as back-translation, is a technique used to augment the amount of parallel data available for training Machine Translation systems and is known to improve translation quality and thus mitigate the lack of data for under-resourced language pairs. It is assumed that, when training on synthetic parallel data, the original monolingual data should be used at the target side and its translation at the source side, an assumption to be assessed. The contributions of this paper are twofold. We investigate the viability of using synthetic data to improve Neural Machine Translation for Portuguese-Chinese, an under-resourced pair of languages for which back-translation has yet to demonstrate its suitability. Besides, we seek to fill another gap in the literature by experimenting with synthetic data not only at the source side but also, alternatively, at the target side. While demonstrating that, when appropriately applied, back-translation can enhance Portuguese-Chinese Neural Machine Translation, the results reported in this paper also confirm the current assumption that using the original monolingual data at the source side outperforms using them at the target side.

Keywords: Neural Machine Translation · Synthetic parallel texts · Back-translation · Portuguese · Chinese · Under-resourced translation pair

1 Introduction

Neural Machine Translation (NMT) is known for its good performance and fluent output, but also for requiring large quantities of parallel data to unfold its potential in terms of delivering quality translations. Most of the current research and existing language resources concern the English language, leaving the vast majority of the other language pairs understudied and with relatively little to no data available for NMT engines to be developed. Portuguese-Chinese (PT-ZH) is

P. Quaresma et al. (Eds.): PROPOR 2020, LNAI 12037, pp. 121–130, 2020.
https://doi.org/10.1007/978-3-030-41505-1_12

the language pair being addressed in this study, and it has very little resources available [4], which is somehow disconcerting, especially when one considers the large number of speakers of either language.[1]

Creating synthetic parallel corpora from monolingual data has been used with favorable results as a valid option to overcome the lack of resources in MT for under-resourced language pairs, as monolingual data is much more readily available and in much greater quantity than parallel data [22]. A common technique to achieve this is known as *back-translation*, through which a monolingual text is automatically translated by an existing seed MT system, giving rise to synthetic parallel data, where each sentence s_o from the original monolingual corpus is paired up with its (synthetic) translation s_{mt}. This parallel corpus can then be used to train further MT systems in both directions, either with s_o as source and s_{mt} as target, or with s_o as target and s_{mt} as source.

This paper addresses the viability of improving NMT for PT-ZH, a language pair under-resourced for MT, with synthetic parallel texts. This work studies which translation direction benefits the most from using the synthetic data. That is, it compares (i.i) *synthetic target*: using the generated synthetic data s_{mt} on the target side and the original monolingual data s_o on the source side, with (i.ii) *synthetic source*: doing it the other way around, where the generated synthetic data s_{mt} is used on the source side and the original monolingual data s_o is used on the target side.

This paper also assesses the impact of progressively increasing the amount of synthetic training data through back-translating monolingual texts. Thus, on the one hand, experiments (i.i) and (i.ii) are undertaken for a range of synthetic data sets of increasing sizes, all generated with the same seed MT engine. On the other hand, the results of those experiments are compared with (ii) *bootstrapping*, where a succession of NMT models are trained with a succession of back-translated data of increasing size such that a given model is trained with the synthetic data created with the NMT model trained in the previous stage.

To create the synthetic data, we opted for using Chinese monolingual texts and (machine) translating them into Portuguese given that the quality of the Chinese text is secured by its publishing source and the level of quality of the Portuguese text outcome can be assessed by the authors of this paper, native speakers of Portuguese.

The experiments and their results presented in this paper demonstrate that back-translation can enhance Portuguese-Chinese NMT. They also deliver important lessons, namely lending credence to the assumption that the synthetic source approach, which has been used in the literature, outperforms the synthetic target approach; and that by resorting to a single, initial seed MT engine to generate the synthetic data, both synthetic source and target approaches outperform the bootstrapping approach, which is based on a succession of MT models successively retrained on the synthetic data generated by the models of the previous stages.

[1] Data fromethnologue.com ranks Chinese as the language with the most speakers among the approximately 7,000 languages in the world, and Portuguese as the sixth.

The remainder of the paper is organized as follows. Section 2 presents previous work in the literature that is more closely related to the present paper. In Sect. 3, a short overview is provided, of Transformer, the NMT architecture used throughout the experiment reported here. Section 4 describes the experiments that were performed as well as the data sets that were used and Sect. 5 presents the results of the experiments. Finally, Sect. 6 closes the paper, with the discussion of the results and conclusions.

2 Related Work

Back-translation was initially used to create synthetic parallel data in the context of Statistical MT, with encouraging results [22]. Since then it has also been successfully applied to NMT [17] and is now a common practice among the most recent work in the field.[2]

Research on back-translation has addressed several issues, such as (i) exploring methods for picking the sentences that form the synthetic corpus [8]; (ii) testing ways to improve the quality of the obtained corpus by applying to it some filtering [10]; and (iii) assessing the impact on performance of varying the ratio of real to synthetic data in the training set [15].

Other studies have attempted an iterative approach where a system trained on synthetic data is used create an additional batch of synthetic data, which is then added to the training set and used to train a presumably better system that will be used to back-translate even more data, iteratively building towards MT systems and parallel data of higher quality [9,23].

Regarding the language pair in the current study, it is worth noting that there is little research on parallel texts [6,13] and on NMT [16] for PT-ZH reported in the literature and, to the best of our knowledge, there is no published research results on applying back-translation to this language pair.

Also of note is that papers that resort to back-translation use the translated text as source and the original text as target, under the assumption that the model will be able to produce better translations if the original non-synthetic data, which are of presumably good quality, are the target that the system will aim to produce, instead of trying to learn to translate when the target is formed by the output of an MT system since these are presumably noisy synthetic sentences. However, there is not much research either for or against this assumption.

3 NMT Architecture

When using machine translation, it is necessary to process sequences of arbitrary length. To allow NMT to cope with input of variable length, the by now familiar encoder-decoder sequence-to-sequence architecture was proposed [19], where an

[2] In the most recent Conference on Machine Translation (WMT'19), nearly two-thirds of the participating systems used back-translation in some way [2].

encoder module, formed by recurrent units, takes the input sentence, one token at each time step, and encodes it into a vector of fixed size. Then, a decoder module, also formed by recurrent units, takes this vector and decodes it, one token at each time step, to produce a target translation.

This architecture achieved good results but still suffered from a major drawback, namely that it forces the encoder to encapsulate the representation of the whole source sentence information into a single vector of fixed size. This bottleneck was overcome by the so-called attention mechanism [1,12], which allows the decoder to access all encoder states, from all time steps, combining them through a weighted sum, thus releasing the encoder from the burden of having to encode the whole sentence in a single vector.

The mechanism of attention has been further exploited by the Transformer model [21], which discards all recurrent units and replaces them with the attention mechanism, resulting in an architecture that has better performance and faster training times. Given this, the Transformer has become the state of the art for NMT and is the architecture used throughout the experiments reported in this paper. In this Section we provide a short overview of Transformer, and refer the interested reader to [21] for more details.

When training, the input to Transformer are the sequences of embeddings of the words in the source and target sentences. However, since the model lacks recurrent units to implicitly track word position in the sentence, this information is explicitly integrated by adding positional embeddings. The input sequences are then fed to the encoder and decoder stacks, which use multi-head self-attention on their inputs, concatenate the output of each head and run the result through a dense layer. On the decoder side, the inputs are masked to block the leakage of future information. For each decoder layer, an additional multi-head attention layer assigns different weights to every encoder state in an averaged sum, similar to the original attention mechanism. Finally, the output of the decoder stack is passed to a softmax that produces a probability distribution over the target vocabulary, and the word with the highest probability is predicted by the model. This is repeated until the target sentence is fully predicted.

4 Experimental Setup

4.1 Seed Corpus and MT System

To obtain a synthetic parallel corpus, one needs a monolingual corpus and a seed MT system with which to back-translate it.

Seed Corpus and MT System. We use the Transformer [21] NMT architecture throughout this work as it is the current state of the art, resorting to the implementation in the Marian framework [11]. The various models trained here follow the setup from the base model described in [21], with 6 encoder and decoder layers, 8 attention heads, and an embedding size of 512. To obtain the seed MT system, Transformer was trained on the UM-PCorpus [6], a PT-ZH parallel corpus with around 1 million sentences from five domains, namely news, technology, law, subtitles, and general.[3] The UM-PCorpus further serves as the *seed corpus* inasmuch as the various back-translated corpora will be added to it.

Test Corpus. Throughout this work, we use the first 1,000 sentences of the corpus News Commentary 11 [20] as the test set (NC11). This corpus is similar to the "newstest" test set used for evaluation in most published research on NMT, and is composed of well curated, high quality translations from the news domain.

Seed System Performance. The seed system scores 13.38 BLEU for the ZH → PT direction and 10.72 BLEU for the PT → ZH direction when evaluated on NC11. These scores are in line with the best results obtained in the literature [16], and will serve as the baseline for the experiments in the present paper.

Monolingual Corpus for Back-Translation. As the monolingual input used to generate the synthetic parallel corpora, we resorted to 6 million Chinese sentences from MultiUN [20], a corpus composed of documents of the United Nations.

Text Pre-processing. Every corpus is pre-processed either with the Moses tokenizer,[4] for Portuguese text, or with the Jieba segmentation tool,[5] for Chinese text. Vocabularies with 32,000 sub-word units [18] are learned separately[6] for both languages of the seed corpus.

4.2 Experiments

Having established a seed corpus and MT system, a test set and a monolingual corpus to back-translate, the following three experiments were undertaken.

Synthetic Source. The approach to training a system on back-translated data commonly found in the literature consists of using the original monolingual corpus on the target side and the synthetic data (obtained by translating the original data) on the source side. This is an option adopted in our experiments as well.

Additionally, given that it is important to monitor the impact of progressively increasing the amount of back-translated data in relation to the seed parallel data, we created three sub-corpora of the synthetic parallel corpus: one with the first 1 million sentences, another with the first 3 million sentences, and yet another with the full 6 million sentences. Each one of these sub-corpora was added in turn to the seed parallel corpus of 1 million sentences and used to train three NMT systems for PT → ZH, which are different from the seed system. Each one of these three systems is trained on a different amount of data and a different

[3] The developers of UM-PCorpus also released an additional set of 5,000 sentence pairs (1,000 pairs from each domain) that we used for development purposes.

[4] We use the implementation from https://github.com/alvations/sacremoses.

[5] You may find Jieba at https://github.com/fxsjy/jieba.

[6] We use the implementation from https://github.com/rsennrich/subword-nmt.

ratio of parallel to synthetic data, namely $S^s_{1:1}$, trained on 2 million sentences (1:1 ratio); $S^s_{1:3}$, trained on 4 million sentences (1:3 ratio); and $S^s_{1:6}$, trained on 7 million sentences (1:6 ratio).

Synthetic Target. The case against using synthetic data on the target side comes from the expectation that the new model would aim at producing translations that are noisier, and thus of less quality, than the seed system. However, it is also possible that the eventual negative effect of an increased number of noisy sentences on the target side during training is canceled or even reverted by an increase in the grammatical diversity of those same sentences. And that performance may nevertheless happen to get improved more with synthetic sentences added to the target side, than with them added to the source side.

The back-translated corpus is used to obtain the same three sub-corpora used in the approach described above, with 1 million, 3 million, and 6 million sentences, and each is added in turn to the seed 1 million parallel corpus. The resulting extended corpora are used to train three new MT systems, namely $S^t_{1:1}$ on 2 million sentences (1:1 ratio), $S^t_{1:3}$ on 4 million sentences (1:3 ratio) and $S^t_{1:6}$ on 7 million sentences (1:6 ratio), but now in the ZH → PT direction.

Bootstrapping. The quality of a synthetic corpus is better when the quality of the MT system used to do the back-translation is also better. Since adding synthetic parallel data to the training set should allow creating a better MT system, this suggests that a bootstrapping approach may yield good results. In this approach, an initial portion of synthetic parallel data is used to augment the seed corpus and the resulting, larger data is used to train a second MT system. This system is then used to generate more synthetic parallel data, of presumably better quality than that produced from the seed system in the previous stage. This synthetic data, generated with this second MT system, is used to augment the seed parallel corpus and the result used to train a third MT system. And so on, with similar bootstrapping steps being iterated for larger portions of data.

In this experiment, we take the seed model for the ZH → PT direction and use it to create a synthetic parallel corpus with 1 million sentence pairs, which is added to the seed corpus and used to train a new MT system for ZH → PT.[7] Note that, up to this point, the result is the same as the $S^t_{1:1}$ system trained on 2 million sentence pairs with synthetic data on the target side, described above.

However, in this experiment, this $S^t_{1:1}$ system is used to back-translate 2 million new sentences which are added to the training data, for a total of 4 million sentence pairs, of which 1 million are from the original seed parallel corpus, 1 million from back-translation with the seed system, and 2 million from back-translation with $S^t_{1:1}$. This corpus of 4 million pairs is used to train a new system, $S^t_{1:1:2}$, which is then used to back-translate 3 million new sentences. These new

[7] Given the direction ZH → PT provided superior results than the direction PT → ZH in the two non-bootstrapping experiments (as reported in detail in Sect. 5), the direction ZH → PT was the one focused on in the bootstrapping approach.

synthetic parallel sentences are added to the training data, which is used to train yet another model, $S^t_{1:1:2:3}$. This last model was trained on a total of 7 million sentences, of which 1 million are from the original seed parallel corpus, 1 million from back-translation with the seed system, 2 million from back-translation with $S^t_{1:1}$, and 3 million from back-translation with $S^t_{1:1:2}$.

Table 1. BLEU scores (higher values are better)

Approach	Parallel corpus			
	Seed	1:1	1:3	1:6
Synthetic on target side	13.38	14.04	**14.12**	13.39
Bootstrap (target side)	13.38	**14.04**	13.59	11.46
Synthetic on source side	10.72	11.45	**11.84**	11.59

5 Results

This Section describes the evaluation results obtained for the experiments undertaken.

Following common practice in the literature, the evaluation of MT performance resorts to the BLEU metric [14], here implemented by the `multi-bleu.perl` script, part of the Moses[8] toolkit. The performance scores for the different experiments are in Table 1.

Results for Synthetic Source. The bottom line of Table 1 displays the performance scores obtained by incorporating the original monolingual corpus on the target side and the back-translated synthetic data on the source side and training PT → ZH models on this data.

The PT → ZH seed model, which uses no synthetic data, gets 10.72 BLEU points. Every other data point on that table line achieves a score higher than this seed model, with the largest score belonging to the $S^s_{1:3}$ model, with 3 million additional synthetic sentence pairs (a ratio of 1:3 of parallel to synthetic data). However, while the trend of increasing BLEU scores, and thus of better translation performance, is visible up to the 3 million sentence data set, translation quality degrades when one adds 6 million synthetic sentences.

The decrease in quality is probably linked to the increase in the ratio of the synthetic parallel data to the original parallel data. Whereas until this data point the addition of noisier data was overcome by the increase of diversity in the training set, with a 1:6 ratio of real to synthetic data the translation quality starts to decrease, with the noise in the synthetic sentences hurting performance.[9]

[8] https://www.statmt.org/moses/.

[9] In future work, further experimentation with introduction of noisy sentences could be explored by resorting to text generated by grammars [3,5,7].

Results for Synthetic Target. The first line of Table 1, in turn, shows the scores when the synthetic data is used in the target side for training ZH → PT models. We see improvements in BLEU up to a 1:3 ratio, from the 13.38 points of the seed ZH → PT model to 14.12 points of the $S^t_{1:3}$ model.

Once again, the $S^t_{1:6}$ model, trained on 6 million synthetic sentences, shows a decrease in quality when compared with the previous data point ($S^t_{1:3}$) and is only 0.01 BLEU points above the corresponding seed model, confirming what had been observed in the previous experiment, that a 1:6 ratio of original parallel data to synthetic parallel data starts to hurt translation performance.

Results for Bootstrapping. For the bootstrapping approach, whose scores are displayed in the second line of Table 1, one only sees an improvement in the first iteration, with the $S^t_{1:1}$ model (which is identical to the model in the synthetic target approach). In the subsequent stages, the performance decreases and the last iteration originates the only model ($S^t_{1:1:2:3}$) that is worse than its corresponding seed model, at 1.42 BLEU points below that starting point.

This experimental result seems to indicate that a rapid decrease in quality occurs, which may be a sign that back-translation is performed by increasingly worse models. A model trained on low-quality data will generate a low-quality synthetic parallel corpus, which when used to trained yet another model will only exacerbate the problem, with the initial positive increase (with the $S^t_{1:1}$ model) apparently not having the strength to generate synthetic data that leverages the performance of the systems in the subsequent stages.

6 Discussion and Conclusions

In this paper, we report on research concerning the viability of improving NMT for the language pair PT-ZH by (progressively) increasing the amount of training data through the back-translation of monolingual Chinese texts.

We experimented with different approaches concerning how to resort to synthetic data, with the results obtained having experimentally demonstrated that: (i) back-translation improves NMT performance for PT-ZH, with every approach experimented with surpassing the seed system to some degree; (ii) creating extended parallel texts by having the original monolingual data on the target side and the generated synthetic translations on the source side provide for the best performance improvements, strengthening this previously untested assumption found in the literature; (iii) bootstrapped NMT engines with recurrent back-translation deliver worse performance than progressive engines that rely on back-translation of increasingly larger data sets by a single seed engine; and (iv) progressive back-translation enters a decaying slope after reaching a peak of performance, rather than maintaining a steady increase: this adds to the literature on back-translation, where it is usually assumed that more synthetic data leads to better or similar MT performance.

This last aspect raises an important question.

While most literature points towards a steady increase in performance until reaching a plateau, the current work contradicts them by finding a drop in quality beyond a certain point of added synthetic data. Further studies on back-translation should focus on whether this behavior is found only for under-resourced languages, or even just for this language pair; and if other languages pairs used in other studies have a different tipping point, reached only at even larger quantities of synthetic data.

Acknowledgements. The research reported here was partially supported by PORTULAN CLARIN—Research Infrastructure for the Science and Technology of Language, funded by Lisboa 2020, Alentejo 2020 and FCT—Fundação para a Ciência e Tecnologia under the grant PINFRA/22117/2016.

References

1. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. In: Proceedings of the International Conference on Learning Representations (ICLR) (2015). Available as arXiv preprint arXiv:1409.0473
2. Barrault, L., et al.: Findings of the 2019 conference on machine translation (WMT19). In: Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1), pp. 1–61 (2019)
3. Branco, A., Costa, F.: Noun ellipsis without empty categories. In: The Proceedings of the 13th International Conference on Head-Driven Phrase Structure Grammar, pp. 81–101 (2006)
4. Branco, A., et al.: The Portuguese Language in the Digital Age. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-29593-5
5. Branco, A.H., Costa, F.: A computational grammar for deep linguistic processing of portuguese: Lxgram, version a.4.1. Technical report, University of Lisbon (2008)
6. Chao, L.S., Wong, D.F., Ao, C.H., Leal, A.L.: UM-PCorpus: a large Portuguese-Chinese parallel corpus. In: Proceedings of the LREC 2018 Workshop "Belt & Road: Language Resources and Evaluation", pp. 38–43 (2018)
7. Costa, F., Branco, A.: Aspectual type and temporal relation classification. In: Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, pp. 266–275 (2012)
8. Edunov, S., Ott, M., Auli, M., Grangier, D.: Understanding back-translation at scale. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 489–500 (2018)
9. Hoang, V.C.D., Koehn, P., Haffari, G., Cohn, T.: Iterative back-translation for neural machine translation. In: Proceedings of the 2nd Workshop on Neural Machine Translation and Generation, pp. 18–24 (2018)
10. Imamura, K., Fujita, A., Sumita, E.: Enhancement of encoder and attention using target monolingual corpora in neural machine translation. In: Proceedings of the 2nd Workshop on Neural Machine Translation and Generation, pp. 55–63 (2018)
11. Junczys-Dowmunt, M., et al.: Marian: fast neural machine translation in C++. In: Proceedings of ACL 2018, System Demonstrations, pp. 116–121 (2018)

12. Kuang, S., Li, J., Branco, A., Luo, W., Xiong, D.: Attention focusing for neural machine translation by bridging source and target embeddings. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1767–1776 (2018)
13. Liu, S., Wang, L., Liu, C.H.: Chinese-Portuguese machine translation: a study on building parallel corpora from comparable texts. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), pp. 1485–1492 (2018)
14. 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 of the Association for Computational Linguistics, pp. 311–318 (2002)
15. Poncelas, A., Shterionov, D., Way, A., de Buy Wenniger, G.M., Passban, P.: Investigating backtranslation in neural machine translation. In: 21st Annual Conference of the European Association for Machine Translation, pp. 249–258 (2018)
16. Santos, R., Silva, J., Branco, A., Xiong, D.: The direct path may not be the best: portuguese-chinese neural machine translation. In: Moura Oliveira, P., Novais, P., Reis, L.P. (eds.) EPIA 2019. LNCS (LNAI), vol. 11805, pp. 757–768. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-30244-3_62
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 (Volume 1: Long Papers), pp. 86–96 (2016)
18. Sennrich, R., Haddow, B., Birch, A.: Neural machine translation of rare words with subword units. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1715–1725 (2016)
19. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: Neural Information Processing Systems, pp. 3104–3112 (2014)
20. Tiedemann, J.: Parallel data, tools and interfaces in OPUS. In: Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012), pp. 2214–2218 (2012)
21. Vaswani, A., et al.: Attention is all you need. In: Neural Information Processing Systems, pp. 5998–6008 (2017)
22. Wu, H., Wang, H., Zong, C.: Domain adaptation for statistical machine translation with domain dictionary and monolingual corpora. In: Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1, pp. 993–1000 (2008)
23. Zhang, Z., Liu, S., Li, M., Zhou, M., Chen, E.: Joint training for neural machine translation models with monolingual data. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)