More Data Is Better Only to Some Level, After Which It Is Harmful: Profiling Neural Machine Translation Self-learning 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, 1749-016 Campo Grande, Lisbon, Portugal
{rsdsantos,jsilva,antonio.branco}@di.fc.ul.pt

Abstract. Neural machine translation needs a very large volume of data to unfold its potential. Self-learning with back-translation became widely adopted to address this data scarceness bottleneck: a seed system is used to translate source monolingual sentences which are aligned with the output sentences to form a synthetic data set that, when used to retrain the system, improves its translation performance. In this paper we report on the profiling of the self-learning with back-translation aiming at clarifying whether adding more synthetic data always leads to an increase of performance. With the experiments undertaken, we gathered evidence indicating that more synthetic data is better only to some level, after which it is harmful as the translation quality decays.

Keywords: Machine translation · Back-translation · Synthetic corpus

1 Introduction

When compared with alternative approaches to translation, neural machine translation (NMT) is known to need a large volume of training data to unleash all its potential, which could create a bottleneck to its application to the vast majority of language pairs, for which little parallel corpora exist. Fortunately, since the seminal study of [12], the so-called back-translation technique offered a solution to alleviate this drawback and became widely used to improve NMT. It can be seen as a form of self-learning approach where the performance of a machine translation system is increased by increasing the amount of its training data with synthetic parallel texts that are produced by the previous version of that system, with inferior translation performance.

To create the synthetic data one needs (i) a monolingual corpus in the desired target language, (ii) an MT system previously trained for the target→source language direction and (iii) to translate the data in (i) with the system in (ii). As expected, the synthetic corpora produced via back-translation is of lower quality than the original parallel corpora used to train the seed system. Nonetheless,

G. Marreiros et al. (Eds.): EPIA 2021, LNAI 12981, pp. 727–738, 2021.
https://doi.org/10.1007/978-3-030-86230-5_57

back-translation has been used to improve NMT as the increase of translation quality with the increase of the volume of training data has been found to offset the decay of quality with the additional synthetic data.

While validating the old maxim that "there is no data as more data", this gives hope to make NMT progress dependent mostly on the availability of increasingly larger computational power, to crunch increasingly larger amounts of synthetic training data, rather than on the availability (or the scarceness) of naturally occurring, non-synthetic parallel data.

Some studies have assessed the variation in the volume of synthetic corpora, [6,10] a.o., confirming that an increase in the synthetic data volume leads to an improved NMT performance. However, to the best of our knowledge, only [4] have studied the self-learning curve provided by back-translation when the training (synthetic) data receives successive increments.

Edunovo et al. [4] resort to back-translation for the German-English and French-English language pairs. They explore various methods to obtain the back-translated sentences by experimenting with: (i) beam and greedy search; (ii) unrestricted and restricted sampling over the target vocabulary; and (iii) by inserting noise into the beam search output. The plain beam search was the worst performing method and beam with added noise (beam+noise) the best. Moreover, in the plain beam search method, a fall in quality is noticeable after some amount of synthetic data is reached, while this is not apparent for the beam+noise method in the experiments reported.

Our goal in this paper is to empirically address the question of whether there is a limit for pushing back-translation and its benefits. In particular, we investigate whether this tipping point behaviour is associated only to the beam search method, or if it is associated to back-translation in general, even if eventually with different tipping points and curve shapes for different methods and initial conditions.

We are seeking to obtain empirical evidence on whether larger synthetic data always bring better translation (lesson: "more data is better"), or rather whether larger synthetic data eventually faces a ceiling for the improvement of translation quality, and if yes, whether that ceiling is approached asymptotically ("more data is either better or not harmful") or reached as a global maximum after which performance decays ("more data is better only to some level, after which it is harmful").

Knowing the answers to these questions is important in order to understand what is the strength that can be expected from self-learning NMT with back-translation and how to make the most efficient use of this technique.

To pursue this goal, we focused on NMT self-learning curves with back-translation. We perform experiments on a highly resourced language pair, viz. German-English, and on a under resourced pair, viz. Portuguese-Chinese. This permits to study the effects of back-translation on different language pairs as well as on different scenarios concerning the availability of resources for NMT.

We found that the performance gains obtained with back-translation do not extend indefinitely, and that in fact there is a point where they peak, after which quality keeps falling. This was observed for both back-translation methods, beam and beam+noise, and also for both pairs of languages, the high-resourced and the low-resourced language pair. We found also that the beam+noise method provides better results in the highly resourced than in the under resourced scenario.

The remaining of this paper is organized as follows. Related work is presented in Sect. 2 and Sect. 3 addresses the methods of obtaining synthetic parallel corpora. Section 4 describes the setup for the various experiments carried out. Sections 5 and 6 present and discuss the results obtained. Finally, Sect. 7 closes this document with concluding remarks.

2 Related Work

Back-translation was initially proposed for Statistical MT, whose data-driven methods also benefit from additional synthetic parallel data [17]. Following this approach, [12] implemented the first NMT system that resorted to back-translated data. They compared back-translation to the mere addition of plain monolingual target data, on both source and target sides, and found that while both methods improved performance, back-translation leads to a significantly larger gain.

Motivated by this result, the usage of back-translation became common practice with NMT. In the most recent Conference on Machine Translation WMT2019 [2], nearly two-thirds of the participating systems used back-translation in some way.

Papers that followed [12] studied the impact of the variation in the quantity of synthetic data provided for training [10], as well as the better methods for obtaining synthetic data [4].

Other studies focused on filtering the synthetic corpus [6] in order to increase its average quality. [5,18], in turn, studied an iterative approach where a system trained on synthetic data is used to create an additional batch of synthetic data, which is then added to the training set and used to train the next version of the system, and so on.

While related to the results of the above mentioned papers, our research question is different from theirs. Rather than seeking to further perfect back-translation, here we are interested in gaining insight about the learning curves of NMT systems supported by back translation.

3 Methods for Back-Translation

When using synthetic parallel corpus to train NMT systems, the natural (non-synthetic) sentences are used in the target language side and their synthetic counterparts, obtained via machine translation from the natural ones, in the source side, rather than vice-versa. In this way, rather in the opposite fashion, the trained system is able to produce better sentences since the natural sentences

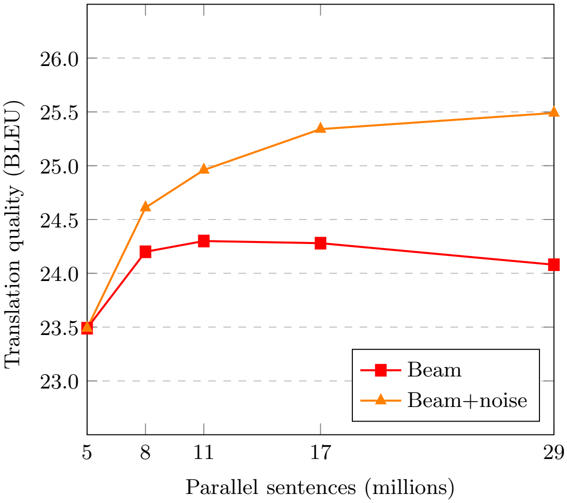

Fig. 1. Learning curves with back-translation with beam and beam+noise methods for English→German. Reproduced from [4].

are usually of good quality—even though the source sentences can be noisy at times.

The synthetic corpus produced with the seed NMT system is then concatenated with the seed parallel corpus, leading to a larger corpus, with more sentence diversity, which supports the training of a translation system with better performance than the seed one. The quality of this system is tied not only to the quantity but also to the quality of the synthetic parallel data, which is impacted by the method used to obtain the synthetic translation.

3.1 Beam Search

The most common approach used to obtain the synthetic translations involves the use of beam search [12], which eventually consists of picking, for a given source sentence, its best (i.e. most likely) translation and adding that source-translation pair to the synthetic corpus.

While, at first blush, picking the most likely translation could seem the best approach, it turns out not to be so, with other methods outperforming it substantially [4]. This happens because always picking the most likely translation leads to less varied translations, as alternative translations that are not the most likely one—though often close to it—will never be picked to integrate the synthetic parallel corpus.

Another problem found with this approach is that performance drops after a certain amount of back-translated synthetic data is added if that data has been obtained with beam search. This behaviour is visible in Fig. 1, adapted from [4]. The first data point in the figure represents the score of the seed model—for the

Monolingual sentence (German)	Der schnelle braune Fuchs springt über den faulen Hund .
Beam search output (English)	The quick brown fox jumps over the lazy dog .
Randomly delete words	The quick brown jumps over the lazy dog .
Randomly replace words with filler tokens	The quick <BLANK> jumps over the lazy <BLANK> .
Randomly shuffle words	quick <BLANK> The jumps over lazy . the <BLANK>

Fig. 2. Adding noise to the output of beam search

beam and the beam+noise methods—trained on 5 million English-German parallel natural sentences. Each subsequent point represents the addition of different amounts of synthetic parallel data.[1] A drop in quality is visible after adding 12 million synthetic parallel sentences when the synthetic corpus is obtained via beam search, but not when it is obtained with the beam+noise method.

3.2 Beam Search+Noise

Different methods have been proposed to mitigate the problem of lack of diversity in back-translated data. One of them is to inject noise [8] into the translations obtained, by randomly changing the order of words, randomly erasing words, or randomly replacing words with a filler token.

We use here the same technique as [4], and add noise to the synthetic sentences by applying the following operations in sequence: (i) deleting words with a probability of 0.1, (ii) replacing words by a filler token—we use the token <BLANK>—with a probability of 0.1, and (iii) randomly shuffling words no further than three positions apart.

Edunovo et al. [4] are not explicit concerning the exact method they used to implement the third step. We based our implementation on the description provided in [8]. Each word is initially assigned an index, $1 \dots n$, corresponding to its position in the sentence. Next, for each word, an uniformly random real number from the range $[0; 4[$ is picked and added to the index of that word. Finally, the words are placed in the sentence following the sorting by their indexes. Figure 2 illustrates the several steps of adding noise to the beam search outputs.

From Fig. 1, [4] observe that adding noise to the output of beam search outperforms the method of using only beam search, and that no decrease in quality was found as the size of the synthetic corpus is exetended (for the amount of synthetic parallel data they resorted to).

[1] This model is trained on a total of 17 million sentences: 5 million of the seed corpus and 12 million of the back-translated corpus. We will use this notation throughout this work, with the first point in the plot representing the seed system, and the subsequent corpora resulting from the addition of the seed corpora with the synthetic corpora.

4 Experimental Setup

This section presents the NMT architecture we used in this paper, as well as the corpora and pre-processing steps resorted to for each experiment. This is followed by the description of the experiments carried out.

4.1 NMT Architecture

We adopt the Transformer model [16] with the "Base" settings, which consist of 6 encoder and decoder layers, 8 attention heads and embedding size of 512.

The Transformer model follows the standard sequence to sequence architecture [14] where an Encoder stack encodes the input sequence, regardless of its length, into a vector of fixed dimensionality, that is then decoded into the target sequence by a Decoder stack.

The main innovations of the Transformer model are in (i) how it relies solely on the attention mechanism [1], dispensing with any of the recurrent modules of previous architectures; and (ii) how it resorts to multiple heads of attention and self-attention, all in all ending with a model that has better performance and needs less training time.

We use the Transformer implementation in the Marian framework [7] and we train with a patience of 10 over the validation on the development corpus every 5,000 updates.

4.2 Corpora

Portuguese-Chinese. In the low-resource scenario, to create the seed system, we used the data sets used in [11]. The Portuguese-Chinese UM-PCorpus [3] has 1 million parallel sentence pairs and was used to train the seed system. The 5,000 sentence pairs of the UM-PCorpus together with the training corpus will be used as development data. The first 1,000 sentences from the News Commentary v11 corpus [15] were used as test set.

The monolingual corpus from which the synthetic parallel data will be obtained is MultiUN [15], composed of documents of the United Nations with close to 11 million sentences.

Every corpus is pre-processed with the Moses tokenizer,[2] for Portuguese, and with the Jieba segmentation tool,[3] for Chinese. Vocabularies with 32,000 sub-word units [13] are learned separately[4] for both languages of the seed corpus. We do so because for this language pair translation quality is lower when vocabularies are learned together and embedding layers are shared [11].

English-German. For the German-English experiments, we trained the seed system on the same data set as [4], that is all the WMT 2018 parallel data, with the exception of the ParaCrawl corpus. We also remove all sentence pairs where

[2] https://github.com/alvations/sacremoses.

[3] https://github.com/fxsjy/jieba.

[4] https://github.com/rsennrich/subword-nmt.

one of the sentences is longer than 250 words, and every pair with a length ratio between source and target larger than 1.5 or smaller than 0.5.

Differently from [4], who use newstest2012 for development and testing, we used newstest2012 corpus for development, and the newstest2019 corpus for testing.

Like in [4], the monolingual corpus from which the synthetic parallel data was obtained is the German monolingual newscrawl data distributed with WMT 2018. We filter this monolingual corpus by removing duplicates and sentences longer than 250 words, with the resulting data set having slightly more than 226 million sentences.

The corpora were pre-processed with the Moses tokenizer, and a joined vocabulary with 32,000 sub-word units was learned.

4.3 Experiments

In order to assess if back-translation always improves translation quality with the addition of new synthetic data, we carried out three experiments.

We experiment with the two methods for obtaining synthetic parallel data presented in Sect. 3 for the Portuguese→Chinese pair. For English→German, we experimented with the beam+noise method, as the learning curve with the beam method had already been shown in [4] to have a n-shape with a tipping point for this language pair, depicted in Fig. 1.

Portuguese→Chinese. The Portuguese-Chinese language pair has very few resources available for it [3] and is under-represented in the NMT literature. Back-translation appears as a promising option to improve translation quality for this pair.

The experiment with this language pair permits to study the impact of back-translation in a scenario with more demanding conditions, namely with less language resources and for languages from two very disparate language families.

For Portuguese-Chinese, and for each method (beam and beam+noise), we trained four systems with different amounts of added synthetic data, namely with 1, 3, 6 and 10 million synthetic sentence pairs obtained with the initial seed system.

English→German. The use of the English-German language pair is the natural follow-up of the experiments undertaken in [4], focusing on the beam+noise method. This permits also an experimental scenario that contrasts with the previous one in that we are dealing with one of the most researched language pair, with many more language resources to support NMT.

To study the behaviour of the beam+noise method on English-German, we trained various systems with different amounts of added synthetic parallel data. The seed model was trained with the same 5 million parallel sentences as in [4]. and was then used to produce synthetic parallel data, with which we trained six new models with 2, 5, 10, 20, 30, and finally 100 million sentences.

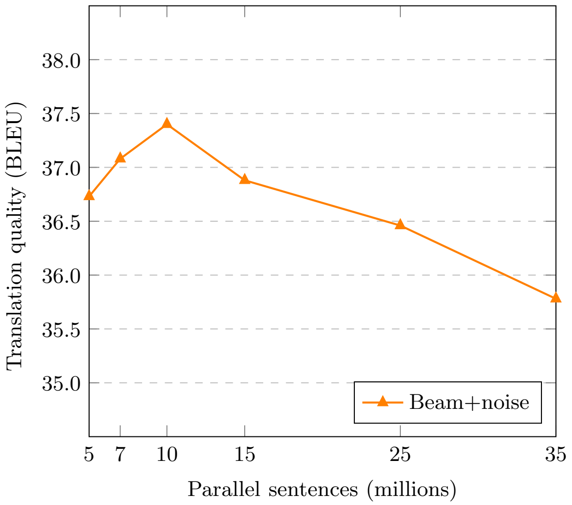

Fig. 3. Learning curve with back-translation with the beam+noise method for English→German

Note that since there are differences in the test corpus and model we are using and the ones used in [4], we cannot continue their work from their last data point, we need to redo performance scores for this experiment.

5 Results

The performance of every NMT model was evaluated with the BLEU metric [9], implemented by the `multi-bleu.perl` script, part of the Moses toolkit.[5]

5.1 English→German

The results for the English→German experiments with the beam+noise method are depicted in Fig. 3. Like with the beam method only (cf. Fig. 1), one observes also here a dip in translation quality after a certain amount of synthetic data is added, with the learning curve peaking at 37.40 BLEU points when 5 million synthetic sentences are added, an improvement over the seed model with 36.73 BLEU points.

We note that, while the Transformer "base" model may not support performance gains with back-translation as large as the Transformer "big" model used in [4], performance improvements can nevertheless still be obtained with the 'base" model.

For the largest amount of added data, 100 million synthetic sentence pairs, translation performance suffers a hard blow, with the model achieving only

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

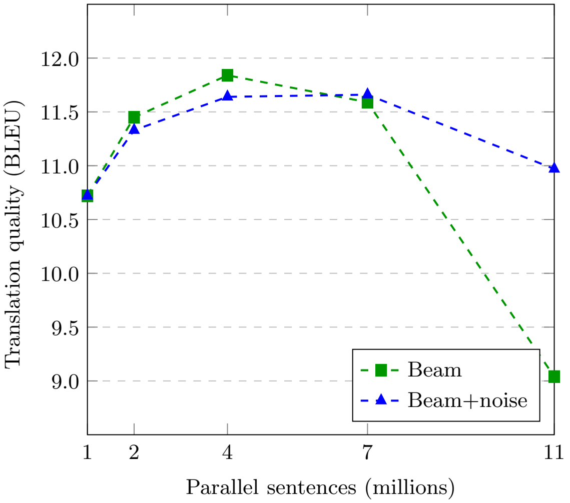

Fig. 4. Learning curves with back-translation with the beam and beam+noise methods for Portuguese→Chinese

28.65 BLEU points—the plot in Fig. 3 is not representing this data point as that would adversely impact the readability of the other points.

5.2 Portuguese→Chinese

Beam Search. As depicted in Fig. 4, back-translation for the Chinese - Portuguese language pair with the beam search method suffers also a drop in quality after a certain amount of synthetic data is added. The performance of the seed model for Portuguese→Chinese has 10.72 BLEU points. The models enriched with back-translation have improvements until 3 million synthetic sentences are added, reaching 11.84 BLEU points. After this, performance scores start falling, confirming that back-translation with beam search has a peak, after which adding more synthetic sentences only harms translation performance.

Beam+Noise. As can be observed in Fig. 4, the same behaviour occurs with the beam+noise method except that to reach the tipping point, a larger amount of synthetic data than with the beam method alone is needed.

The seed model is the same as for the beam search method, with 10.72 BLEU points. The following three data points see cumulative improvements, peaking at 11.66 BLEU points for the model trained with 6 million synthetic sentences added (7 million in total). This is also where the beam+noise method outperforms the beam method for the first time. But while the beam+noise method outperforms the beam method from that point onward, its maximum (11.66) is lower than the maximum (11.84) achieved by the beam search method. For the next model trained with the largest amount of data, 11 million, there is a drop in quality. This dip, however, is considerably smoother than what is observed for

Table 1. Seed model vs Tipping model

		Seed performance (BLEU)	Seed size (million sent.)	Tipping performance (BLEU)	Tipping size (million sent.)
En→De	Beam+noise	36.73	5	37.40	10
	Beam	23.49	5	24.30	11
Pt→Zh	Beam+noise	10.72	1	11.66	7
	Beam	10.72	1	11.84	04

Table 2. Performance gains

		Delta performance (BLEU)	Delta Size (million sent.)	Performance gain [tip. BLEU/seed BLEU] (%)	Performance gain rate [delta performance/ delta size] (BLEU points/million sent.)
En→De	Beam+noise	0.67	5	1.82%	0.13
	Beam	0.81	6	3.45%	0.14
Pt→Zh	Beam+noise	0.94	6	8.77%	0.16
	Beam	1.12	3	10.45%	0.37

the beam method, which confirms that the beam+noise method performs better than the beam method for larger amounts of data.

When comparing Figs. 1 and 4, one observes that the superiority of the beam+noise over the beam method is inverted. This is in line with similar inversion already noticed by [4] on two other methods, where the relative ranking of the beam and sampling methods depends on the amount of data used to create the seed model, with the sampling method outperforming the beam method in a highly resourced scenario, and vice-versa in an under resourced scenario.

6 Discussion

To help profiling the self-learning with back-translation, key scores and figures are combined and gathered in Tables 1 and 2.

The performance gain is larger in the under-resourced scenario (10.45% and 8.77% with the beam and the beam+noise method, respectively), illustrated by the Portuguese-Chinese language pair, than in the highly-resourced scenario (3.45% and 1.82% with the beam-noise and the beam method, respectively), illustrated by the English-German scenario. Back-translation is more effective in an under-resource scenario as it provides a better boost of translation quality when the seed system is trained with smaller amounts of data.

Back-translation is also more efficient in an under-resourced scenario as, proportionally, it requires a smaller extension of the seed data to obtain the same level of performance enhancement. The performance gain rate is larger for Portuguese-Chinese(0.37 and 0.16 BLEU points/M sentences with the beam and the beam+noise method, respectively) than for English-German(0.14 and 0.13). With the best performance gain observed (10.45%)—in the case of Pt→Zh + Beam —, the tipping performance attained is a humble 11.84 BLEU score. To attain the best tipping performance observed (37.40 BLEU score)—in the case of En→De + Beam+noise —, back-translation helps with a humble 1.82% of performance gain with respect to the seed system. Interestingly, in terms of absolute size of the volume of synthetic data necessary to be added to reach the tipping performance, that increment lies in the short range of 3M to 6M sentences, irrespective of the type of scenario at stake.

All in all, taking into account the research question in this paper, the key lesson learned when profiling NMT self-learning with back-translation is that "more data is better only to some level, after which it is harmful".

7 Conclusion

Back-translation is a widely used technique for creating synthetic parallel corpus, by translating monolingual data with a seed translation system to augment the amount of data available for training a new model. While back-translation has been found to be a valid technique for obtaining models with better performance, the evidence gathered in this paper indicates that the gains are not ever growing with the addition of more synthetic data. The performance peaks at some point, after which adding more back-translated data only hurts performance.

Our experiments addressed two methods for generating back-translated data, plain beam search and beam+noise, and were run for two language pairs representing different scenarios of resource availability, Chinese-Portuguese for a scenario of low resource availability, and English-German for a scenario of high resource availability. The finding that performance peaks before starting to drop was consistent throughout all experiments. We also confirm that the best method for generating back-translated data depends on the quality of the seed model doing the back-translation, with the beam+noise method being better suited for high resource scenarios, as this method outperforms beam search for the English-German highly resource language pair, but is outperformed by the latter for the under resourced Portuguese-Chinese language pair.

We conclude that back-translation, despite its strengths, is not an approach that should always be applied to ultimately arrive at a better system. The gains obtained from back-translation peak at some point, but the point where this peak occurs depends on various factors, such as the quality of the seed system and the method used for back-translation. As such, the use of back-translation should be carefully considered and monitored to eventually not reduce translation quality.

Acknowledgement. This research 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)
2. Barrault, L., et al.: Findings of the 2019 conference on machine translation (WMT19). In: Proceedings of the 4th Conference on MT, pp. 1–61 (2019)
3. 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 and Road: Language Resources and Evaluation", pp. 38–43 (2018)
4. 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)
5. 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)
6. 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)
7. Junczys-Dowmunt, M., et al.: Fast neural machine translation in C++. In: Proceedings of ACL 2018, System Demonstrations, pp. 116–121 (2018)
8. Lample, G., Conneau, A., Denoyer, L., Ranzato, M.: Unsupervised machine translation using monolingual corpora only. In: International Conference on Learning Representations (ICLR) (2018)
9. 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)
10. 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)
11. Santos, R., Silva, J., Branco, A., Xiong, D.: The direct path may not be the best: Portuguese-Chinese neural machine translation. In: Progress in Artificial Intelligence (EPIA 2019), pp. 757–768 (2019)
12. 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)
13. 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)
14. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: Neural Information Processing Systems, pp. 3104–3112 (2014)
15. Tiedemann, J.: Parallel data, tools and interfaces in OPUS. In: Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012), pp. 2214–2218 (2012)
16. Vaswani, A., et al.: Attention is all you need. In: Neural Information Processing Systems, pp. 5998–6008 (2017)
17. 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, pp. 993–1000 (2008)
18. Zhang, Z., Liu, S., Li, M., Zhou, M., Chen, E.: Joint training for neural machine translation models with monolingual data. In: 32nd AAAI Conference (2018)