The Direct Path May Not Be The Best: Portuguese-Chinese Neural Machine Translation

Rodrigo Santos[1(✉)], João Silva[1], António Branco[1], and Deyi Xiong[2]

[1] Department of Informatics, NLX—Natural Language and Speech Group, University of Lisbon, Lisbon, Portugal
{rsdsantos,jsilva,antonio.branco}@di.fc.ul.pt
[2] College of Intelligence and Computing, Tianjin University, Tianjin, China
dyxiong@tju.edu.cn

Abstract. Machine Translation (MT) has been one of the classic AI tasks from the early days of the field. Portuguese and Chinese are languages with a very large number of native speakers, though this does not carry through to the amount of literature on their processing, or to the amount of resources available to be used, in particular when compared with English. In this paper, we address the feasibility of creating a MT system for Portuguese-Chinese, using only freely available resources, by experimenting with various approaches to pairing source and target parallel data during training. These approaches are (i) using a model for each source-target language pair, (ii) using an intermediate pivot language, and (iii) using a single model that can translate from any language seen in the source side to any language seen on the target side. We find approaches whose performance is higher than that of the strong baseline consisting of an MT service provided by an IT industry giant for the pair Portuguese-Chinese.

Keywords: Neural Machine Translation · Portuguese · Chinese

1 Introduction

Human language is the prime vehicle we use for communication. In an increasingly globalized world, language differences pose a barrier to communication, reducing the number of people we can reach, or that can reach us. Artificial Intelligence, through Machine Translation (MT), appears as a viable solution to this problem by providing an automatic way of translating between languages.

Most of the research on MT concern English and some other language, often German or French, leaving other pairs of languages underrepresented in the literature. In this paper we are concerned with the translation between Portuguese (PT) and Chinese (ZH). Both languages have a very large number of native speakers, but despite this there are few available resources with which to build an MT system for these languages.

P. Moura Oliveira et al. (Eds.): EPIA 2019, LNAI 11805, pp. 757–768, 2019.
https://doi.org/10.1007/978-3-030-30244-3_62

In order to determine the feasibility of creating a state-of-the-art MT system for this pair of languages, we take the current best model for MT, the Transformer deep neural encoder-decoder and, using only freely available resources, experiment with three approaches to pairing source and target parallel data for training. The results we obtain show that it is possible to develop an MT system for Portuguese-Chinese with performance surpassing that of a very strong baseline consisting of an MT service, Google Translate, provided by an IT industry giant for the pair Portuguese-Chinese.

The paper is organized as follows. Section 2 presents related work on NMT, and Portuguese-Chinese NMT in particular. Section 3 describes the approaches for pairing languages during training, and Sect. 4 covers the corpora we used. Section 5 describes the NMT model and what was done to train the system. Section 6 presents the evaluation results. Finally, Sect. 7 provides concluding remarks.

2 Related Work

This Section introduces the current state-of-the-art for NMT and the existing literature on Portuguese-Chinese NMT.

2.1 Machine Translation Models

Machine Translation (MT) has been a perennial topic in Natural Language Processing since the early days of AI research. Over the years, many approaches have been attempted, from symbolic to statistical, with varying degrees of success. Recently, beginning with sequence to sequence models based on recurrent networks [10], deep neural models have become by far the most popular approach, buttressed by the availability of large amounts of training data and hardware capable of efficient parallel computation. A clear sign of this trend can be seen in the most recent Conference on Machine Translation, WMT 2018 [2], where 33 of the 38 participating systems in the News shared task used deep neural models.

All current top-performing neural MT (NMT) models employ some variant of an attention mechanism [1,7], which allows the model to assign different weights to the different words in the input sequence. The state-of-the-art approach, the Transformer model [13], relies solely on attention and completely does without the recurrent architectures of past NMT systems.

Given its state-of-the-art performance, we will use the Transformer model in this work. The model is described in Sect. 5.

2.2 Portuguese-Chinese Machine Translation

As mentioned above, most of the research on MT involves English as one of the languages, the other often being French or German, as much of the initial research targeted these pairs and subsequent studies continued the trend in order for the results obtained to be comparable.

There is very little literature on MT for the pair Portuguese-Chinese. To the best of our knowledge, for NMT in particular, only [3] and [6] address this pair. In both cases, the authors are presenting a new parallel corpus and a system is trained in order to assess the quality of the data and show that it is feasible to use it to train a NMT system. Since both works use different test sets, their results are hardly comparable and do not allow us to establish an expected performance score for the state-of-the-art of Portuguese-Chinese NMT. Nonetheless, we return to these publications in Sect. 6 when discussing our results.

3 Approaches to Training

Given that the Transformer is currently the uncontested state-of-the-art model, what remains to determine the feasibility of creating an MT system for the pair Portuguese-Chinese is how to best use the available resources. This Section describes three approaches to how existing parallel data can be used in training.

3.1 Using a Model for Each Source-Target Language Pair

The most straightforward solution for creating an MT system for a set of languages is to use a parallel corpus for each pair of languages.

In the particular case of the current study, a Portuguese-Chinese parallel corpus would allow us to create two models, one for each translation direction, that is a PT → ZH model and a ZH → PT model.

One might expect this approach to yield the best performance, as we are training separate models, each specific to a language pair and direction. The greatest disadvantages of this approach are that the number of models that are required grows quadratically with the number of languages, which is not an issue in this study, and that for some language pairs there is little parallel data available.

3.2 Using a Pivot Language

For some pairs of languages there are few available parallel corpora. In such cases it might be more advantageous for the translation to go through an intermediate third language, the *pivot* language (p), in a two-step process, as there might be more data available for the source-pivot and pivot-target pairs.

In the particular case of the current study, four models are required, two for each translation direction.[1] The PT → p → ZH direction requires models for PT → p and p → ZH, while the ZH → p → PT direction requires models for ZH → p and p → PT.

This approach allows using more training data, but this data may be more heterogeneous since it will originate from unrelated parallel corpora, and the

[1] If creating an MT system for many languages, this approach only requires two models per language; a much lower number than when using a model for each language pair.

⟨pt⟩ The quick brown fox jumps over the lazy dog *(translate to Portuguese)*
⟨zh⟩ The quick brown fox jumps over the lazy dog *(translate to Chinese)*

Fig. 1. Tagging the source sentence with the target language in the corpora for the many-to-many approach

two-step process is likely to introduce detrimental translation errors in the intermediate step. Experiments need to be carried out to assess whether the increase in training data is enough to mitigate or even overcome the problematic issues.

3.3 Using a Single Model for All Pairs (Many-to-Many)

Another approach that can be attempted is to gather all available parallel data into a single corpus. This approach draws its motivation from zero-shot translation [4], which revealed that any parallel data where the source language has been seen on the source side is useful for training and, likewise, any parallel data where the target language has been seen on the target side is also useful.

Using the language pairs mentioned in the above examples, this would mean gathering in a single corpus all the parallel data for PT-ZH, ZH-PT, PT-p, ZH-p, p-PT and p-ZH.

While at first blush this may seem complicated to manage, the way it is made to work is actually quite simple. For all source-target sentence pairs, the source sentence is prefixed with a special token indicating the language of the corresponding target sentence. After the model is trained, and when a sentence is to be translated, one needs only prefix that sentence with the special token for the desired target language, as exemplified in Fig. 1.

This approach greatly increases the amount of data that can be used for training and yields a *single model* that is able to translate from any of the languages that have been seen as source to any of the languages that have been seen as target, which provides much flexibility in its use. On the flip side, the model has to contend with what is presumably a much more difficult task, which might decrease its performance, and the large amounts of data will have a negative impact on the model training time.

4 The Corpora

As mentioned in Sect. 2, there has been very little work done on NMT for the pair Portuguese-Chinese, and the publications that exist ([3] and [6]) are geared towards presenting the parallel corpora that the authors had created rather than achieving good translation performance. Therefore, each one uses different corpora for training, development and evaluation, which makes the results of those systems non comparable.

In the present work, in order to allow future comparisons, we resort to News Commentary V11, a well known corpus with good quality, non-trivial translations, which is part of the OPUS collection of corpora [12]. This corpus exists

Table 1. Corpora (UM-PCorpus) for the direct approach

Domain	Sent.
News	146,095
Legal	173,420
Subtitles	250,000
Technology	250,000
General	250,000
Total	1,069,515

for multiple pairs of language, though the textual content for a pair does not overlap with the textual content for other pairs. We take the first 1000 sentences for the PT-ZH pair as the test set for evaluation.

4.1 Corpora for the Direct Approach

In order to address the lack of quality corpora for translation between Portuguese and Chinese, both [3] and [6] created new corpora. Of these, only the UM-PCorpus [3] has been made publicly available, and only partially so, as the corpus is reported to have 6 million sentences but the authors only release 1 million of them. Still, this portion of UM-PCorpus is currently the largest publicly available parallel corpus with acceptable quality for the creation of Portuguese-Chinese MT systems, and the one that we will use for the direct approach.

The released portion of UM-PCorpus is comprised of 1 million sentences from news, legal, subtitles, technology and general domains. A detailed breakdown of the number of sentences in each domain is given in Table 1, showing a rather balanced distribution of sentences over the various domains. Alongside this corpus, meant for training, the authors also make available an extra 5000 sentences for testing, 1000 from each domain. Since we have already established News Commentary V11 PT-ZH to be the test set, we set these 5000 sentences apart to use as a development set.

4.2 Corpora for the Pivot Approach

When opting for the pivot approach, the rationale is to take advantage of the fact of there being more training data available in the source-pivot and pivot-target pairs separately than there is for the source-target pair alone.

Not surprisingly, we use English (EN) as the pivot language, as there are several parallel corpora between English and both Portuguese and Chinese that are of good quality and of large enough size to support training MT systems.

Parallel corpora for Chinese-English is abundant. We resort to the OPUS [12] collection of corpora, from where we gather close to 10 million parallel sentences for the ZH-EN pair, to which we add an additional 2.2 million sentences from the UM-Corpus [11].[2]

[2] UM-Corpus [11], for ZH-EN, and UM-PCorpus [3], for ZH-PT, should not be confused.

Table 2. Corpora for the pivot approach

(a) ZH-EN pair

Corpus (domain)	Sent.
News Commentary V11 (News)	0.07M
Tanzil (Religious)	0.19M
UM-Corpus (Various)	2.22M
MultiUN (UN translations)	9.56M
total	12.04M

(b) PT-EN pair

Corpus (domain)	Sent.
Tanzil (Religious)	0.1M
JRC-Acquis (Law)	1.6M
Europarl (EU Parliament)	2.0M
Paracrawl V3 (Web crawl)	3.3M
total	7.0M

Corpora for Portuguese-English is not as abundant. By again resorting to OPUS we were able to gather close to 4 million parallel sentences for the PT-EN pair, which we extend with 3.3 million sentences from version 3 of Paracrawl.[3]

The Paracrawl corpus results from a Web crawl. Consequently, its quality is not the best, and some clean-up and filtering were needed before the sentences could be added to the corpus. The sentences in Paracrawl are annotated with extra information that we use to guide the filtering. Our filtering criteria removed all sentence pairs where: (i) either sentence was shorter than 3 tokens; (ii) sentences had arabic numerals that did not match; (iii) both sentences were equal; (iv) either sentence had only numbers or symbols; and (v) where the length ratio between the sentences was larger than 3:2.

Table 2 gives a more detailed breakdown of the constituent parts of the training corpora used in the pivot approach.

For development, 5000 sentences of News Commentary V11 PT-EN were used for the PT-EN pair and the UM-Corpus test set, with also 5000 sentences, was used for the ZH-EN pair.

4.3 Corpora for the Many-to-Many Approach

For this approach, we gather in a single corpus all of the parallel corpora used for the two other approaches, that is the 1 million sentences of PT-ZH used in direct approach, and the two corpora used in the pivot approach, namely 7 million sentences of PT-EN and 12 million sentences of ZH-EN.

Note that, in the many-to-many approach, each parallel corpus can contribute to training twice, once for each translation direction. For instance, the 1 million sentences of PT-ZH can be used as an additional 1 million sentences of ZH-PT data. This results in a corpus of 40 million sentences, which turned out to be such a large amount of data as to make training unfeasible. To overcome this, we halved the amount of data.[4] For each parallel corpora, odd numbered lines are used for one translation direction while even numbered lines are used for the other translation direction (cf. Table 3), resulting in a parallel corpus with 20 million sentences, where all the sentences available to us still occur in some translation direction.

[3] https://paracrawl.eu/.

[4] Despite this 50% reduction in the size of the corpus, training the many-to-many model took around 808 GPU hours (more than 33 days) to converge.

Table 3. Corpora for the many-to-many approach

	$p \rightarrow$ PT	$p \rightarrow$ ZH	$p \rightarrow$ EN
PT $\rightarrow p$	—	0.5M (Even)	3.5M (Odd)
ZH $\rightarrow p$	0.5M (Odd)	—	6.0M (Even)
EN $\rightarrow p$	3.5M (Even)	6.0M (Odd)	—

5 The NMT System

This Section describes the NMT architecture, implementation framework and pre-processing steps, which were common to all approaches.

5.1 Transformer Model

The Transformer model [13] is rather recent, but it has quickly established itself as the state-of-the-art for NMT.[5] The model still follows the standard deep encoder-decoder architecture to learn a mapping between a source sequence and a target sequence. Its main innovation is in how it relies only on multiple heads of attention [1,7] and self-attention, without any of the recurrent modules of previous architectures. A high level overview of the model is shown in Fig. 2.

As per usual in neural approaches to text processing, the symbols in the source and target sequences are represented in an embedding space. Since the Transformer does not use a recurrent mechanism, information about the position of the symbols in the sequences is explicitly added through sinusoidal positional embeddings (not shown in the Figure).

The source and target sequences are then fed to a stack of encoder and decoder blocks (6 of each, in the Figure), which is the common procedure in all deep encoder-decoder architectures. These blocks begin by applying multi-head self-attention to their inputs (8 heads, in the Figure), concatenating the output of each head and running the result through a feed forward layer. Note that, for the decoder blocks, self-attention is masked in order to hide the symbols that occur after the symbol currently being predicted. For each decoder block, there is an additional multi-head attention component, this one weighing the output states of the corresponding encoder block. The output of the final decoder block is fed to linear and softmax layers in order to predict the output symbol.

The Transformer paper [13] presents variations of hyper-parameters for the model. We use the hyper-parameters of the "base" variant, with 6 encoder and decoder layers, 8 attention heads and an embedding dimension of 512. The models are trained until they converge, which is determined using a cross-entropy patience of 10 on the development corpus. That is, training stops after 10 iterations, of 5000 update steps each, with no improvement to cross-entropy.

[5] In the most recent WMT 2018 [2], 33 of the 38 systems used deep neural models, and 29 of these 33 were based on the Transformer model.

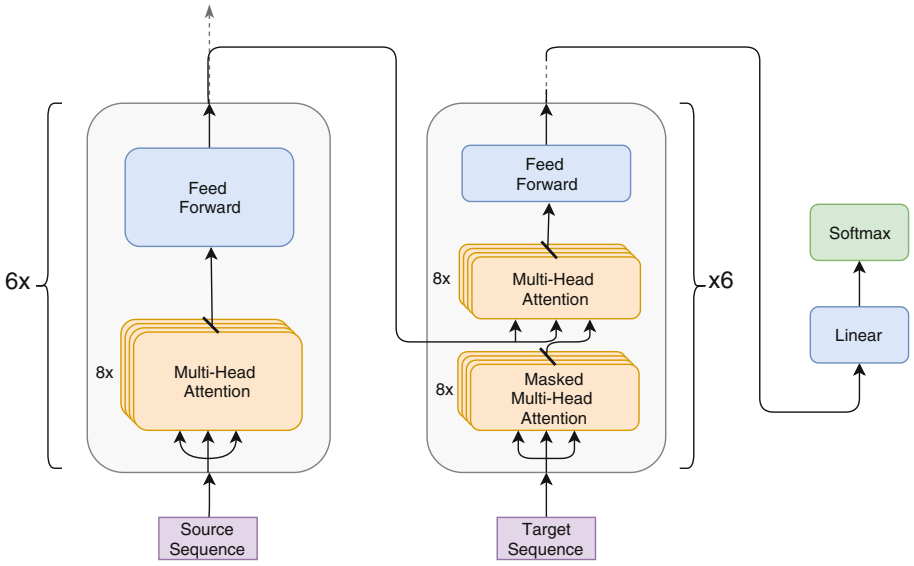

Fig. 2. The transformer - model architecture.

5.2 Training the System

To implement the system, we use the Transformer implementation that is part of the Marian framework [5], which offers an efficient runtime, good documentation and an easy API with which we built an online translation service.[6] Training was performed on a single Nvidia Tesla K40m GPU.

All data is pre-processed before being fed to the model. The pre-processing steps consist of tokenization/segmentation and sub-word vocabulary creation.

Chinese sentences were segmented using Jieba[7] and both Portuguese and English sentences were tokenized using the sacremoses[8] implementation of the Moses tokenizer.

Current NMT systems do not process texts at the word level. Since NMT models have a finite vocabulary, working at the level of words would mean that eventually the model would be faced with words that it has not encountered before, be them rare terms, neologisms or misspellings. These are known as out-of-vocabulary (OOV) words, and they have a negative impact on performance. To tackle this problem, words are split into sub-word sequences. This sub-word vocabulary is built in such a way that, for any input word, there is always some sequence of sub-words that forms the input word [9], thus eliminating the possibility of occurrence of OOV symbols while maintaining a finite vocabulary,

[6] https://portulanclarin.net/workbench/lx/translator.
[7] https://github.com/fxsjy/jieba.
[8] https://github.com/alvations/sacremoses.

Table 4. BLEU evaluation scores

Model	ZH → PT	PT → ZH
Google Translate	12.23	13.69
Our systems		
direct	13.38	10.72
pivot	17.79	14.84
many-to-many	14.04	11.99

at the cost of increasing the length of the sequences. For all models, a vocabulary of 32,000 sub-words was created using the implementation[9] from [9].

6 Results

This Section summarizes and discusses the results achieved in the work described in this paper. All translation quality scores are reported using the BLEU [8] metric implemented by the multi-bleu.perl script in Moses. Recall that the test corpus we use for all experiments corresponds to the first 1000 sentences from News Commentary V11 PT-ZH. The results are summarized in Table 4.

As mentioned in Sect. 2, the only other works on Portuguese-Chinese NMT are [3] and [6]. The authors of the former release a test set, which provides some common ground on which to perform a comparison, but only part (1 million out of 6 million sentences) of the training corpus they used is made available. The authors of the latter do not release the system or the corpus, making it unfeasible to compare their results to ours.

To have a competing system that we could run on the News Commentary V11 PT-ZH test set, we resorted to a popular online translation services, Google Translate.[10] To run these tests, we used an option that allows submitting to the service a file to be translated, taking care to split the work into batches as there is an unstated limit on the size of the text that the service accepts.

Google Translate is a proprietary system, but existing publications point towards it also using deep NMT, and likely a many-to-many approach [4]. Also, given the resources available to Google, it is reasonable to expect their system to have good performance and set a competitive baseline.

Evaluating the output of Google Translate raises an issue that is specific to the PT → ZH translation direction. The BLEU metric takes into account the n-gram overlap between the automatic and the reference translation, but the output of the system, being Chinese, does not separate words with whitespace. As such, it has to be segmented prior to evaluation. For this, we again use the Jieba tool, the same used for pre-processing data for our models (cf. Sect. 5.2).

[9] https://github.com/rsennrich/subword-nmt.
[10] https://translate.google.com/.

6.1 Direct Approach

Our direct approach gets better results than Google Translate on ZH → PT (13.38 vs. 12.23) but a worse score on PT → ZH (10.72 vs. 13.69).

Our results suggest that, when having a single parallel corpus in a direct approach, where the models are trained on the same amount of data for each direction, the PT → ZH direction performs worse than the ZH → PT direction. This is supported by the two other studies in Portuguese-Chinese NMT, namely [3] and [6], which find the same trend using different corpora and NMT systems.

Google Translate likely uses a many-to-many approach [4], and the fact that the PT → ZH direction outperforms the ZH → PT direction can be explained by there being more parallel data available where Chinese is one of the languages in the pair than there is parallel data involving Portuguese.

6.2 Pivot Approach

The results we obtained for the pivot approach corroborate the assumption that, given enough data, going through a pivot language, even if doing so introduces errors in the intermediate step, can outperform a direct approach trained with fewer data. Our system, when using the pivot approach, clearly outperform the direct approach and the Google Translate baseline on either direction, achieving a score of 17.79 for ZH → PT, against the 13.38 of the direct approach and the 12.23 of Google Translate, and 14.84 for ZH → PT, against the 10.72 of the direct approach and the 13.69 of Google Translate.

We find a similar experiment in [6], where the authors also compared a direct approach, using their (unreleased) parallel corpus, with doing a pivot approach using a greater amount of data. The results obtained, however, show the direct approach performing better than the pivot approach. This mismatch between their findings and ours can be explained by the amount of data that was used in each of the two steps of the pivot approach. For the direct approach, [6] use a corpus with nearly 1 million sentences, which is of similar size to the one we use. For the pivot approach, they use 25 million sentences for ZH-EN, which is quite larger than the 12 million we use in our work, but their corpus for PT-EN has only 2 million sentences, while ours has 7 million. The pivot approach relies on both steps having good performance, and in [6] there is not enough data for the PT-EN pair to ensure a good translation quality for the whole two-step process.

6.3 Many-to-Many Approach

Finally, the last experiment served to assess whether it was feasible to achieve good performance with the many-to-many approach given the amount of data that we used for training.

The results we obtained place the performance of the many-to-many approach between that of the direct approach and that of the pivot approach, for either direction, and above that of the Google Translate baseline for ZH → PT.

These results again highlight the extent to which neural models rely on large amounts of data to obtain good performance. The many-to-many approach gives us a single model can translate in any direction between Portuguese and Chinese (and also English, though we have not evaluated that), with better performance than that of the two direction-specific models of the direct approach. It is unable, however, to reach the performance of the pivot approach, but we believe this to be due to us having had to halve the amount of data used in training.

Although the pivot approach has to go through a presumably noisy intermediate step, its models have been trained over 7 million sentences where Portuguese is either source (PT → EN) or target (EN → PT), and 12 million sentences where Chinese is either source (ZH → EN) or target (EN → ZH); while in the many-to-many approach the amount of data for each language is much smaller (cf. Table 3).

7 Conclusion

This paper presented a study on the feasibility of creating a state-of-the-art NMT system for Portuguese-Chinese using only freely available resources. Using the state-of-the-art Transformer model, we experimented with three approaches to pairing source and target parallel data for training the system. These are (i) the direct approach, using a model for each source-target language pair; (ii) the pivot approach, translating through an intermediate language for which there is more data available; and (iii) the many-to-many approach, using a single model, trained on all available data, that can translate from any language seen as source to any language seen as target. We compare our systems against Google Translate, which we outperform when using the pivot approach for both directions, and when using the direct and many-to-many approaches for the ZH → PT direction.

The pivot approach had the best performance, but it is reasonable to expect that the many-to-many approach, given enough computational power, will be able to surpass it, as it can always make use of more data than the other two approaches since any parallel corpora where a language occurs either on the source/target sides can be used in the training of a many-to-many model that is able to translate from/to that language.

For future work, and still in keeping with the rationale of using only freely available resources, we will experiment with augmenting the training data, for all three approaches, with backtranslated monolingual texts.

An online service demonstrating the system, as well as its supporting model, may be found at https://portulanclarin.net/workbench/lx/translator.

Acknowledgements. The research results presented here were supported by FCT—Foundation for Science and Technology of Portugal, MOST—Ministry of Science and Technology of China, through the project Chinese-Portuguese Deep Machine Translation in eCommerce Domain (441.00 CHINA-BILATERAL), the PORTULAN CLARIN Infrastructure for the Science and Technology of Language, the National Infrastructure for Distributed Computing (INCD) of Portugal, and the ANI/3279/2016 grant.

Deyi Xiong was supported by National Natural Science Foundation of China (Grants No. 61622209 and 61861130364).

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). arXiv preprint arXiv:1409.0473
2. Bojar, O., et al.: Findings of the 2018 conference on machine translation (WMT18). In: Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers, pp. 272–307 (2018)
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 & Road: Language Resources and Evaluation", pp. 38–43 (2018)
4. Johnson, M., et al.: Google's multilingual neural machine translation system: enabling zero-shot translation. Trans. Assoc. Comput. Linguist. **5**, 339–351 (2017)
5. Junczys-Dowmunt, M., et al.: Marian: fast neural machine translation in C++. In: Proceedings of ACL 2018, System Demonstrations, pp. 116–121 (2018)
6. 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)
7. Luong, M.T., Pham, H., Manning, C.D.: Effective approaches to attention-based neural machine translation. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 1412–1421 (2015)
8. 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)
9. 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)
10. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: Neural Information Processing Systems, pp. 3104–3112 (2014)
11. Tian, L., Wong, D.F., Chao, L.S., Quaresma, P., Oliveira, F., Yi, L.: UM-Corpus: a large English-Chinese parallel corpus for statistical machine translation. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014), pp. 1837–1842 (2014)
12. 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)
13. Vaswani, A., et al.: Attention is all you need. In: Neural Information Processing Systems, pp. 5998–6008 (2017)