Portuguese-Chinese Neural Machine Translation

Rodrigo Soares dos Santos

Faculty of Sciences, University of Lisbon
rsdsantos@di.fc.ul.pt

Abstract. This work reports on a study addressing Neural Machine Translation for the language pair Portuguese ↔ Chinese where the development of a state-of-the-art Machine Translation system for this pair was undertaken by using only freely available resources.

There are two main challenges to this work: (i) the gathering of corpora with good enough quality and quantity, which for this under-resourced pair of languages is complicated issue; and (ii) the adoption of a suitable architecture that is able to make the most of these corpora.

Three approaches are experimented with, one of them outperforming a baseline consisting of the Google Translate service for this language pair, in both translation directions. All the implemented systems make use of deep learning, namely by resorting to the Transformer architecture.

An online translation service was also developed, showcasing the two translation directions, and is freely available online.[1]

Part of the work presented in this dissertation was peer reviewed and published in EPIA 2019 (Santos et al., 2019).

Keywords: Neural Machine Translation · Portuguese · Chinese

1 Introduction

Language is the prime vehicle for human communication and, since the early days of Artificial Intelligence, it has been the subject of study in the subdomain of Natural Language Processing (NLP). As an application of NLP, Machine Translation (MT) contributes towards diluting the communication barriers among humans that have mastered different natural languages, which in an increasingly globalized world are obstacles for mutual understanding.

Literature on MT revolves mostly around pairs involving English and a few other languages, leaving some important and widely spoken languages comparatively under-represented. Both Chinese and Portuguese have strong positions world wide and a large number of speakers, being the 1^{th} and 6^{th} most spoken languages in the world, respectively,[2] but, research on Neural Machine Translation (NMT) for this language pair has been residual. Against this background, this work addresses the challenge of determining how far one is presently able to

[1] https://portulanclarin.net/workbench/lx/translator/

[2] Data from ethnologue.com.

go when developing NMT solutions for both directions of the Portuguese ↔ Chinese (PT ↔ ZH) language pair making use only of freely available resources.

2 NMT Background

In this Section I provide a brief overview of the relevant NMT background for this work.

2.1 Sequence-to-Sequence Encoder-Decoder

Sutskever et al. (2014) was the first to use Deep Neural Networks to map sequences to sequences (Seq2Seq), which is the core of a translation system where a source sequence is mapped to a target sequence. The idea is that these systems learn to map a source sentence to a target sentence directly, in an end-to-end fashion, given enough training data exist from a large parallel corpus.

In order to obtain a representation from a sentence, one needs a method that can process sequences of words of variable size. An idea that was successful with speech recognition was the use of Recurrent Neural Networks (RNN). When applied to MT these networks have recurrent units that process one word at each time step in a recurrent way, keeping an internal state between time steps. By using LSTM units, a type of recurrent units, Sutskever et al. (2014) devised a method to encode the input sequence, regardless of its length, into a vector of fixed dimensionality, and then use another LSTM to decode the target sequence from that vector, hence the name encoder-decoder.

2.2 Attention Mechanism

In the Seq2Seq encoder-decoder architecture mentioned above, the encoder has to pack the representation of the whole input sequence into a single vector that is passed on to the decoder, which places a great burden on the model. The mechanism of attention, introduced in the seminal paper of Bahdanau et al. (2015), releases the encoder of this burden by, instead of passing a single vector from the encoder to the decoder, allowing the decoder to access *all encoder states*, each contributing a different amount to the final vector representation of the input.

The attention mechanism brought large improvements to all encoder-decoder architectures and has since become a staple of all NMT systems.

2.3 Transformer

The Transformer (Vaswani et al., 2017) is a rather recent architecture, but it has quickly established itself as the state of the art for NMT. It follows the standard encoder-decoder architecture to learn a mapping between a source and a target sequence.

The main innovations of the Transformer model are in (i) how it relies solely on attention, dispensing with any of the recurrent modules of previous architectures; and (ii) how it resorts to multiple heads of attention and self-attention.

However, since the Transformer does not use a recurrent mechanism, information about the position of the words in the sequences needs to be explicitly added to the input source and target sequences. In (Vaswani et al., 2017), this is done through sinusoidal positional embeddings.

Note that, as an additional benefit, not having recurrent modules allows to greatly accelerate training of the model since its layers are almost only feed forward layers and do not have temporal dependencies between them.

3 Approaches to Training

A core issue in MT is how to make the best use of the available parallel data. Hence, in the present work I experiment with three different approaches to training an NMT system. These approaches are described below.

3.1 Using a Different Model for Each Direction (Direct)

A straightforward option to create an MT system for a pair of languages is to use a parallel corpus of these languages. For the language pair under study, a single PT ↔ ZH parallel corpus will allow 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 given separate models are trained, each specific to a language pair and direction. This is the way most of the literature tackles the problem of translation, and the approach that normally sets the state of the art for most language pairs.

As neural network models need large amounts of data, underperformance with this approach is encountered for languages for which there is little parallel corpora available.

I refer to this solution as the *direct approach*.

3.2 Using a Pivot Language (Pivot)

For some pairs of languages, there are few parallel corpora available. The pair PT ↔ ZH is one such case (Chao et al., 2018). In this circumstance, it might be more advantageous for the translation to go through an intermediate third language, the pivot language, in a two-step process, as there might be more data available for the source-pivot and pivot-target pairs than there is for the source-target pair. This may permit to train two systems where concatenation delivers better performance than a direct approach with fewer data, in spite of the accumulated losses in the two steps.

The first system starts by translating from Portuguese or Chinese to the pivot language and then, the second system, translates from the pivot language to Chinese or Portuguese, respectively. So, all in all, four models are needed in

Source	Target
<pt> What is your name?	Qual é o teu nome?
<zh> What is your name?	你叫什么名字?
<en> 你叫什么名字?	What is your name?
<pt> 你叫什么名字?	Qual é o teu nome?
<zh> Qual é o teu nome?	你叫什么名字?
<en> Qual é o teu nome?	What is your name?

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

order to accomplish the translation in both directions. The data used are parallel corpora for Portuguese ↔ pivot and Chinese ↔ pivot. Note that there is no direct translation between Portuguese and Chinese in this approach.

This approach is referred to in this work as the *pivot approach.*

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

Following the ideas from Johnson et al. (2017), the so called zero-shot machine translation seems to be a useful approach for NMT between under-resourced languages. This consists in giving more language pairs to a model for training than those available under the direct approach in order to improve translation quality of under-resourced pairs, and even translate between pairs that are not seen in training.

To this extent, a system consisting of a single model was created from a corpus composed by all the data used for the direct and pivot approaches. In order to know to which language the system should translate to, a special token is appended to the beginning of the source sentence denoting the language of the target sentence, as exemplified in Figure 1.

An advantage of the many-to-many system is there being more data available than for either of the two previous approaches, as it is capable of using all their corpora for training, and in this way provide the model with more data. On the flip side, the model has to contend with a more difficult task that may decrease its potential performance.

I refer to this solution as the *many-to-many approach.*

4 Corpora

The three approaches indicated above require or benefit from different types of parallel corpora, which are discussed in this section.

4.1 Direct Approach Corpora

Parallel data for the PT ↔ ZH pair is scarce (Chao et al., 2018). Existing corpora are normally of low quality and/or low quantity, which leads to training

Table 1: Pivot corpus distribution

(a) PT ↔ EN pair

Corpus (Domain)	**Sent.**
Tanzil (Religious)	0.12M
JRC-ACQUIS (EU Law)	1.63M
Europarl (EU Parliament)	1.96M
Paracrawl (Web Crawl)	3.25M
Total	**6.96M**

(b) ZH ↔ EN pair

Corpus (Domain)	**Sent.**
News Commentary v11 (News)	0.07M
Tanzil (Religious)	0.19M
UMCorpus (Various)	2.22M
MultiUN (United Nations)	9.56M
Total	**12.04M**

sub-optimal neural networks for MT. This happens because the PT ↔ ZH language pair has not been the focus of much research and, as such, there are few suitable corpora for training a NMT model.

Chao et al. (2018) try to tackle this problem, and create a PT ↔ ZH parallel corpus. Despite indicating that the corpus has around 6 million sentences, Chao et al. (2018) only make available for public use a subcorpus with 1 million PT ↔ ZH parallel sentences. These 1 million sentences, which include texts from 5 domains (news, legal, technology, subtitles, and general), are used as training set for the direct approach.

4.2 Pivot Approach Corpora

For the pivot approach, there was the need to find parallel data involving both Portuguese, Chinese and a pivot language. The pivot language chosen was English (EN) given the availability of parallel language data between English and both Portuguese and Chinese, and given the quality and quantity of those data.

The corpus used for the pair PT ↔ EN resulted from the concatenation of four corpora. These four corpora were taken from the OPUS repository (Tiedemann, 2012). In the same fashion, for the ZH ↔ EN directions, four corpora were gathered, with around 12 million sentences in total. The corpora for the pivot approach are summarized in Table 1.

4.3 Many-to-many Corpora

The many-to-many approach benefits from being supported by more corpora than the other two approaches. It benefits from all kinds of parallel corpora where one of the languages of interest occurs, in our case Portuguese or Chinese.

The final corpus consisted of all the data used by the previous two approaches, i.e. the 1 million sentence pairs from the direct approach, the 7 million sentence pairs used in the pivot approach for the PT ↔ EN directions, and the 12 million pairs also used in the pivot approach for the ZH ↔ EN directions.

All the data was duplicated, and by means of appropriate prefixation (as described in Section 3.3), every sentence pair was given to the model in both directions, resulting in a corpus with 40 million sentence pairs.

5 Evaluation

In order to have a baseline against which the performance of the various systems that I developed in the present study can be compared, I resort to the online service Google Translate.[3] To obtain the relevant baseline score, I evaluated this service on the same test set as the various approaches studied here, the first 1,000 sentences of the PT ↔ ZH News Commentary corpus. This established a very strong baseline to be challenged by my systems.

With Google Translate being one of the most used translation services around the world, any score near this baseline would be praiseworthy, taking into account the dimension of the company and the resources available to its MT team, in terms of qualified expert human resources, data and computational power, and them not being restricted to freely available resources. Evaluation against such an industry giant was made easier because Google Translate allows uploading documents to be translated.

Table 2 summarizes the BLEU scores obtained by the baseline and the three approaches I developed. Following the common practice in the literature, I evaluate every trained model using the BLEU metric (Papineni et al., 2002), in this work implemented by the `multi-bleu.perl` script, part of the Moses[4] toolkit.

Automatic Evaluation. Considering the three studied approaches, the direct approach was the one with the lowest scores. However, it was still able to surpass the Google Translate baseline for the ZH → PT translation direction by more than 1 BLEU point. For the ZH → PT translation direction, the direct approach achieved 13.38 BLEU points against the 12.23 points from the baseline. This is a very satisfactory result, considering that this approach is the most straightforward approach studied here and the one that used the least amount of training data.

The second approach to train a MT system studied in this dissertation was the pivot approach. This approach achieved the best results, outperforming the baseline for both translation directions. When translating from Portuguese to Chinese, the pivot approach achieves 15.35 points BLEU, an improvement of around 1 BLEU points over the baseline (14.29). For the other translation direction (ZH → PT) the pivot approach achieves 17.79 BLEU points, an impressive improvement of over 5 BLEU points on the score of the baseline.

Finally, the many-to-many approach fared between the other two approaches. For the PT → ZH direction, it falls slightly behind the baseline, with 13.98 BLEU points against 14.29. For the other translation direction (ZH → PT), this approach achieves 16.22 BLEU points, another impressive improvement, with 4 points above the baseline.

[3] `https://translate.google.com/`

[4] `https://www.statmt.org/moses/`

Table 2: Summary of BLEU scores

(a) ZH → PT direction

Corpus	BLEU
Google Translate baseline	12.23
Direct approach	13.38
Pivot approach	**17.79**
Many-to-Many approach	16.22

(b) PT → ZH direction

Corpus	BLEU
Google Translate baseline	14.29
Direct approach	11.05
Pivot approach	**15.25**
Many-to-Many approach	13.98

6 Future Work Exploration

While the results obtained in this dissertation are very satisfactory, with performance above of that obtained by a technology giant, research on this language pair for NMT is far from over. One path to improve translation quality is by having more parallel corpora. Back-translation addresses this issue. It is a method for extending corpora where a previously trained MT system on a chosen pair of languages is used to translate monolingual corpora in one of these languages to the other, hence creating additional “synthetic” parallel data for those two languages.

As an initial exploratory experiment in this topic, I trained a system using the direct approach on the concatenation of a back-translated “synthetic” corpus, generated via back-translation, with the UMPCorpus (Chao et al., 2018). The resulting system outperforms the direct approach (cf. Table 2) in both directions, with 15.29 BLEU for the ZH → PT direction and 13.17 for the PT → ZH direction, an improvement of 2 BLEU points in each direction.

Back-translation is not the only path to improve translation quality, other research directions like unsupervised NMT (MT that uses only monolingual data), or multilingual NMT (incorporating information from other languages/language pairs) appear a valid option to increase translation quality for PT ↔ ZH in future work.

7 Conclusion

The main objective of this work was to address the challenge of determining how far one is presently able to go when developing MT solutions for both directions of the Portuguese ↔ Chinese language pair making use only of freely available resources, and ultimately to develop a state of the art MT system that is able to translate from Portuguese to Chinese, and from Chinese to Portuguese. This objective has been successfully completed, achieving better performance than the Google Translate service, which has been created by a tech giant with access to a very large supply of expert human resources, of parallel data and of computational resources.

An online translation service was also developed, showcasing one of the three approaches studied in this document for the two translation directions, and is freely available online (see below for link).

Part of the work presented in this dissertation was already peer reviewed and was published in EPIA2019 (Santos et al., 2019).

Relevant links

CV: `http://nlx-server.di.fc.ul.pt/~rodrigo/CV.pdf`
Dissertation: `http://nlx-server.di.fc.ul.pt/~rodrigo/MScDiss.pdf`
NMT service: `https://portulanclarin.net/workbench/lx/translator/`

Acknowledegements

The dissertation was supported by the FCT–Foundation for Science and Technology of Portugal (ANI/3279/2016) and the PORTULAN CLARIN Research Infrastructure for the Science and Technology of Language.

Bibliography

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio (2015). Neural machine translation by jointly learning to align and translate. In *Proceedings of the International Conference on Learning Representations (ICLR)*. Available as arXiv preprint arXiv:1409.0473.

Chao, Lidia S., Derek F. Wong, Chi Hong Ao, and Ana Luísa Leal (2018). UM-PCorpus: A large Portuguese-Chinese parallel corpus. In *Proceedings of the LREC 2018 Workshop*, pages 38–43.

Johnson, Melvin, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean (2017). Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, pages 339–351.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Santos, Rodrigo, João Silva, António Branco, and Deyi Xiong (2019). The direct path may not be the best: Portuguese-chinese neural machine translation. In *Progress in Artificial Intelligence (EPIA 2019)*, pages 757–768.

Sutskever, Ilya, Oriol Vinyals, and Quoc V Le (2014). Sequence to sequence learning with neural networks. In *NIPS*, pages 3104–3112.

Tiedemann, Jörg (2012). Parallel data, tools and interfaces in OPUS. In *Proceedings of the LREC 2012*, pages 2214–2218.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin (2017). Attention is all you need. In *NIPS*, pages 5998–6008.