Multilingual Neural Machine Translation: Case-study for Catalan, Spanish and Portuguese Romance Languages

Pere Vergés Boncompte and Marta R. Costa-jussà
TALP Research Center
Universitat Politècnica de Catalunya, Barcelona
pere.verges@est.fib.upc.edu, marta.ruiz@upc.edu

Abstract

In this paper, we describe the TALP-UPC participation in the WMT Similar Language Translation task between Catalan, Spanish, and Portuguese, all of them, Romance languages. We made use of different techniques to improve the translation between these languages. The multilingual shared encoder/decoder has been used for all of them. Additionally, we applied back-translation to take advantage of the monolingual data. Finally, we have applied fine-tuning to improve the in-domain data. Each of these techniques brings improvements over the previous one.

In the official evaluation, our system was ranked 1st in the Portuguese-to-Spanish direction, 2nd in the opposite direction, and 3rd in the Catalan-Spanish pair.

1 Introduction

Research in the field of Machine Translation (MT) has been growing during these last years. From statistical approaches (Koehn et al., 2003) to neural ones (Bahdanau et al., 2015), the progress has been impressive. Even after having achieved exceptional results based only on attention mechanisms (Vaswani et al., 2017), there are still many challenges and improvements remaining, for instance, multilingual translation from languages other than English, which have lower resources, and domain adaptation.

In order to tackle these challenges, the Similar Language Task organized in the context of the Conference on Machine Translation (WMT 2020) has provided an appropriate setting for them. Within this task, the focus is the translation between languages that are different from English, and more specifically, the focus consists of translating languages that are from the same family. The families included are the following: South-Slavic, Indo-Aryan, and Romance.

In our case, we have devoted the research to Romance languages, which include Spanish, Portuguese, and Catalan. The evaluation comprised all translation directions, but only provided parallel training data for Spanish-Portuguese and Spanish-Catalan. We approached the Portuguese-Catalan pair both from a pivot-based and zero-shot perspective.

In this paper, we make use of the well-known multilingual shared encoder/decoder and we show its effectiveness when applied to languages of the same linguistic family. Additionally, we benefited from back-translation and fine-tuning.

2 Background

In this section, we show an overview of neural-based multilingual machine translation and domain adaptation using fine-tuning.

2.1 Multilingual translation

When having multiple languages, there is the opportunity to use several NMT architectures, based in the Transformer (Vaswani et al., 2017). Among the alternatives, we can share encoders and decoders (Johnson et al., 2017) or have specific encoders and decoders for each language (Escolano et al., 2020). In this paper, we are using the shared approach and we are leaving as further work to compare with other ones.

Shared encoder-decoder One direct approach is using a single encoder/decoder shared for all languages (Johnson et al., 2017). In this case, parameters and vocabulary are shared among all language pairs and it helps the generalization across languages improving the translation for the low resource language pairs (Aharoni et al., 2019). Additionally, the shared encoder/decoder allows using zero-shot easily, only by adding a tag in the source sentence. The source sentence has to contain the

Proceedings of the 5th Conference on Machine Translation (WMT), pages 447–450

language abbreviation of the target language. So, when translating from Catalan to Spanish, we have to include the *<2es>* tag at the beginning of the Catalan source sentence, which means that we are translating into Spanish.

```
<2es> Bon dia -> Buenos días
```

Therefore, it is necessary to add the tag to indicate the target language, followed by the sentence to be translated. This is necessary both in training and inference.

2.2 Monolingual corpus selection for back-translation

There is a large amount of monolingual data available for this task. Monolingual data can improve the system by using back-translation (Sennrich et al., 2016). However, back-translation is a process that consumes a lot of resources, so we decided to select the monolingual data within the target domain. The selection criterion has been the TF-IDF (Term Frequency – Inverse Document Frequency), which defines the relevance of the words in a document. Using this criterion, we compared all the available monolingual data against the development set and only kept the files that had a higher score among all.

2.3 Domain adaptation

One approach to improve the translation of a specific language domain is to make use of fine-tuning techniques. Fine-tuning consists of retraining a model that has already been trained with out-of-domain data, with in-domain data. The disadvantage of fine-tuning is that it tends to overfit, due to the small amount of in-domain data used, compared to the out-of-domain data. Sometimes the final model might fall into the problem of catastrophic forgetting (French, 1999).

One approach to avoid over-fitting and catastrophic forgetting is to do mixed fine-tuning, which consists of shuffling the in-domain with the out-of-domain data, and then train normally on this combined data (Chu and Dabre, 2019).

3 Experimental Framework

In this section, we describe the datasets used for the task, the data preprocessing, the training, and the evaluation of the bilingual and multilingual systems.

3.1 Data and Preprocessing

Data Selection All the data used in our experiments has been provided by the organizers, so we did not make use of any additional parallel nor monolingual data. For the Catalan-Spanish and Spanish-Portuguese translation, we used all the parallel data available, which is about 11.3 million sentences for the Catalan-Spanish translation and 4.1 million sentences for the Spanish-Portuguese. For the Catalan-Portuguese we did not have any parallel data. We have also used monolingual data for back-translation purposes. Two million sentences have been used from the *CaWaC* file for Catalan, about 1.1 million sentences from *News-commentary-v15* and *News-crawl-2019* files for Portuguese, and 1.5 million sentences from *News-commentary-v15* and *News-crawl-2015* for Spanish. The multilingual model has been trained using all the parallel data, and with pseudo-parallel data that has been obtained by applying back-translation. To achieve the back-translation we used our best system at the moment to perform the translation of the monolingual data, obtaining the pseudo-parallel corpus. As said in Section 2.2, the monolingual data has been selected using TF-IDF as the measure for text similarity [1]. We used 2/3 of the development set for fine-tuning purposes and 1/3 of the development set as a test set.

Preprocessing We followed the standard procedure for preparing the data, which consists of normalizing, tokenizing, truecasing, and cleaning (limiting sentences from 1 to 50 words). To perform these actions we made use of the *Moses*[2] scripts. We extracted the joint subwords with byte-pair encoding (BPE)[3].

3.2 Parameter Details

The bilingual and multilingual models are both based on the Transformer architecture, implemented with *fairseq* toolkit [4]. We assigned six attention layers for the encoder and the decoder, each having four attention heads per layer, with an embedding dimension of 512. Additionally, all the models shared the source and target embeddings. The multilingual model shared the embeddings among all language pairs. Each batch was

[1] https://github.com/BhargavaRamM/Document-Similarity

[2] https://github.com/moses-smt/mosesdecode

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

[4] https://github.com/pytorch/fairseq

assigned to have a maximum number of tokens of 2048. The optimizer used was Adam, setting the betas to β_1 = 0.9 and β_2 = 0.98, with a learning rate of 5e-4 varied with the inverse square root of the step number. The warm-up steps were set equal to 4000, a dropout of 0.1, and a weight decay and gradient clipping norm set to 0.

4 Results

The results show the improvements obtained by applying multilinguality, back-translation, and fine-tuning techniques. For the pair Catalan-Portuguese (CA-PT), in which there was no training data available. We have used the cascade technique, which consists of concatenating the translation of Catalan-to-Spanish and Spanish-to-Portuguese systems, and the other way around for the opposite direction. Also, we have used the multilingual system to obtain zero-shot translation for this pair.

Directions	BI	MULT	+BACK	+FT
ES→CA	64.23	73.12	70.59	71.21
CA→ES	60.64	69.56	73.01	74.05
ES→PT	27.20	27.62	28.80	29.55
PT→ES	29.70	30.57	30.89	32.12
CA→PT	20.99	24.94	25.52	26.94
PT→CA	25.21	28.00	27.97	29.18
CA→PT ZS	-	12.47	13.56	16.05
PT→CA ZS	-	17.67	19.64	19.56

Table 1: BLEU results for all the systems evaluated in the development of this study. BI = Bilingual, MULT = Multilingual, BACK = Multilingual with Backtranslation, FT = Multilingual with back-translation and Fine-tuning, ZS = zero-shot.

Table 1 shows that the multilingual model outperforms the bilingual model in all cases. Zero-shot performs worse than the cascade method. Applying back-translation to the multilingual model improves for most language pairs and directions. Finally, when applying fine-tuning to the back-translation model, we see an improvement in all pairs and directions, except for the PT→CA direction with zero-shot.

4.1 Official evaluation results

Here we report the official evaluation. We participated with our best system which was the multilingual model with back-translation and fine-tuning. For the CA-PT directions, we translated using the cascade technique, Table 2 reports the results on the evaluation test set. Our system was ranked 1st in the Portuguese-to-Spanish direction, 2nd in the opposite direction, and 3rd in the Catalan-Spanish pair. For the Catalan-Portuguese directions, the results were not released.

Directions	BLEU
ES-CA	60.50
CA-ES	68.84
ES-PT	32.33
PT-ES	33.82
CA-PT	32.80
PT-CA	34.40

Table 2: Official BLEU scores for the evaluation of the final test set.

5 Discussion

We will now discuss the results obtained for each system we have trained, comparing one against the others.

Bilingual model compared to the Multilingual model We have shown that the multilingual model outperforms the bilingual model in all translations directions, with an improvement that varies from +0.4 to +6.9 BLEU. The multilingual model allows for a better generalization by sharing the vocabulary among all the languages. Additionally, the multilingual model allows for zero-shot translation.

Back-translation This technique allows us to make use of monolingual data. The improvement with this technique varies from +0.5 to +3.4 BLEU, except when using the monolingual Catalan data (ES→CA and PT→CA directions). This deterioration is probably due to the lower resemblance (estimated using the TF-IDF score) of the *CaWaC* dataset compared to the target domain.

Fine-tuning We have applied fine-tuning to perform the domain adaptation. To do so, we added 2/3 of the development data set to the already trained model, which is the multilingual model with back-translation, since it was the best model we had so far. After doing so, we had to retrain the model from the last checkpoint, preventing it from overfitting. By applying fine-tuning, we were able to achieve improvements between +0.6 and +2.5 BLEU points (except in zero-shot). This fine-tuning improvement is achieved by using very few resources (1500 sentences) and less time compared to back-translation, which requires more resources and time.

6 Conclusion

We have observed how using a multilingual shared encoder/decoder in languages from the same family improves bilingual translation. This is due to a positive transfer among these languages while sharing vocabulary and embeddings. Additionally, this multilingual shared system has been improved with both back-translation and fine-tuning methods.

Acknowledgments

We are grateful to Carlos Escolano for his comments, corrections, and help throughout the investigations. This work is supported in part by the Spanish Ministerio de Ciencia e Innovación, through the postdoctoral senior grant Ramón y Cajal and by the Agencia Estatal de Investigación through the projects EUR2019-103819, PCIN-2017-079 and PID2019-107579RB-I00 / AEI / 10.13039/501100011033

References

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3874–3884, Minneapolis, Minnesota. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Chenhui Chu and Raj Dabre. 2019. Multilingual multi-domain adaptation approaches for neural machine translation. *CoRR*, abs/1906.07978.

Carlos Escolano, Marta R. Costa-jussà, José A. R. Fonollosa, and Mikel Artetxe. 2020. Multilingual machine translation: Closing the gap between shared and language-specific encoder-decoders. abs/2004.06575.

Robert M. French. 1999. Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, 3(4):128 – 135.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proc. of the Conference of the NAACL*, pages 48–54.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. 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)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.