Neural Machine Translation with the Transformer and Multi-Source Romance Languages for the Biomedical WMT 2018 task

Brian Tubay and Marta R. Costa-jussà
TALP Research Center, Universitat Politcnica de Catalunya, Barcelona
brian.alcides.tubay.alvarez@alu-etsetb.upc.edu,marta.ruiz@upc.edu

Abstract

The Transformer architecture has become the state-of-the-art in Machine Translation. This model, which relies on attention-based mechanisms, has outperformed previous neural machine translation architectures in several tasks. In this system description paper, we report details of training neural machine translation with multi-source Romance languages with the Transformer model and in the evaluation frame of the biomedical WMT 2018 task. Using multi-source languages from the same family allows improvements of over 6 BLEU points.

1 Introduction

Neural Machine Translation (NMT) (Bahdanau et al., 2015) proved to be competitive with the encoder-decoder architecture based on recurrent neural networks and attention. After this architecture, new proposals based on convolutional neural networks (Gehring et al., 2017) or only attention-based mechanisms (Vaswani et al., 2017) appeared. The latter architecture has achieved great success in Machine Translation (MT) and it has already been extended to other tasks such as Parsing (Kaiser et al., 2017), Speech Recognition [1], Speech Translation (Cros et al., 2018), Chatbots (Costa-jussà et al., 2018) among others.

However, training with low resources is still a big drawback for neural architectures and NMT is not an exception (Koehn and Knowles, 2017). To face low resource scenarios, several techniques have been proposed, like using multi-source (Zoph and Knight, 2016), multiple languages (Johnson et al., 2017) or unsupervised techniques (Lample et al., 2018; Artetxe et al., 2018), among many others.

[1] https://tensorflow.github.io/tensor2tensor/tutorials\/asr$_$with$_$transformer.html

In this paper, we use the Transformer enhanced with the multi-source technique to participate in the Biomedical WMT 2018 task, which can be somehow considered a low-resourced task, given the large quantity of data that it is required for NMT. Our multi-source enhancement is done only with Romance languages. The fact of using similar languages in a multi-source system may be a factor towards improving the final system which ends up with over 6 BLEU points of improvement over the single source system.

2 The Transformer architecture

The Transformer model is the first NMT model relying entirely on self-attention to compute representations of its input and output without using recurrent neural networks (RNN) or convolutional neural networks (CNN).

RNNs read one word at a time, having to perform multiple steps before generating an output that depends on words that are far away. But it has been demonstrated that the more steps required, the harder it is to the network to learn how to make these decisions (Bahdanau et al., 2015). In addition, given the sequential nature of the RNNs, it is difficult to fully take advantage of modern computing devices such as Tensor Processing Units (TPUs) or Graphics Processing Units (GPUs) which rely on parallel processing. The Transformer is an encoder-decoder model that was conceived to solve these problems.

The encoder is composed of three stages. In the first stage input words are projected into an embedded vector space. In order to capture the notion of token position within the sequence, a positional encoding is added to the embedded input vectors. Without positional encodings, the output of the multi-head attention network would be the same for the sentences “I love you more than her”

Proceedings of the Third Conference on Machine Translation (WMT), Volume 2: Shared Task Papers, pages 667–670

https://doi.org/10.18653/v1/W18-64076

and "I love her more than you". The second stage is a multi-head self-attention. Instead of computing a single attention, this stage computes multiple attention blocks over the source, concatenates them and projects them linearly back onto a space with the initial dimensionality. The individual attention blocks compute the scaled dot-product attention with different linear projections. Finally a position-wise fully connected feed-forward network is used, which consists of two linear transformations with a ReLU activation (Vinod Nair, 2010) in between.

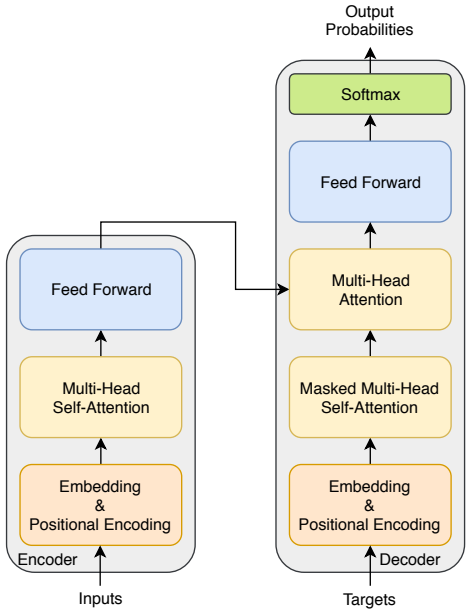

Figure 1: Simplified diagram of the Transformer model

The decoder operates similarly, but generates one word at a time, from left to right. It is composed of five stages. The first two are similar to the encoder: embedding and positional encoding and a masked multi-head self-attention, which unlike in the encoder, forces to attend only to past words. The third stage is a multi-head attention that not only attends to these past words, but also to the final representations generated by the encoder. The fourth stage is another position-wise feed-forward network. Finally, a softmax layer allows to map target word scores into target word probabilities. For more specific details about the architecture, refer to the original paper (Vaswani et al., 2017).

3 Multi-Source translation

Multi-source translation consists in exploiting multiple text inputs to improve NMT (Zoph and Knight, 2016). In our case, we are using this approach in the Transformer architecture described above and using only inputs from the same language family.

4 Experiments

In this section we report details on the database, training parameters and results.

4.1 Databases and Preprocessing

The experimental framework is the Biomedical Translation Task (WMT18)[2]. The corpus used to train the model are the one provided for the task for the selected languages pairs: Spanish-to-English (es2en), French-to-English (fr2en) and Portuguese-to-English (pt2en). Sources are mainly from Scielo and Medline and detailed in Table 3.

Training	Scielo	Medline	Total
es2en	713127	285358	998485
fr2en	9127	612645	621772
pt2en	634438	74267	708705
all2en	1356692	972270	2328962

Table 3: Corpus Statistics (number of segments).

Validation sets were taken from *Khresmoi development data*[3], as recommended in the task description. Each validation dataset contains 500 sentence pairs. Test sets were the ones provides by the task for the previous year competition (WMT17[4]).

Preprocessing relied on three basic steps: tokenization, truecasing and limiting sentence length to 80 words. Words were segmented by means of Byte-Pair Encoding (BPE) (Sennrich et al., 2015).

4.2 Parameters

The system was implemented using OpenNMT in PyTorch (Klein et al., 2017) with the hyperparameters suggested in the website [5]. Other parameters used in training are defined in Table 4. Both single-language systems and multi-source system

[2] http://www.statmt.org/wmt18/biomedical-translation-task.html
[3] https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-2122
[4] http://www.statmt.org/wmt17/biomedical-translation-task.html
[5] http://opennmt.net/OpenNMT-py/FAQ.html

System	es2en		pt2en		fr2en	
	WMT17	WMT18	WMT17	WMT18	WMT17	WMT18
Best performing system	37.49	**43.31**	43.88	**42.58**	-	**25.78**
Single-Language	39.35	39.06	44.31	38.54	31.75	19.42
Multi-Language	**40.11**	40.49	**45.55**	39.49	**38.31**	**25.78**

Table 1: Trained systems results for WMT17 and WMT18 official test sets.

Spanish	Utilizando la base de datos Epistemonikos, la cual es mantenida mediante bsquedas realizadas en 30 bases de datos, identificamos seis revisiones sistemticas que en conjunto incluyen 36 estudios aleatorizados pertinentes a la pregunta.
Single-Language	Using the Epistemonikos database, which is maintained through searches in 30 databases, we identified six systematic reviews including 36 randomized studies relevant to the question.
Multi-Language	Using the Epistemonikos database, which is maintained through searches in 30 databases, we identified six systematic reviews that altogether include 36 randomized studies relevant to the question.
Portuguese	Os resultados dos modelos de regresso mostraram associao entre os fatores de correo estimados e os indicadores de adequao propostos
Single-Language	Regression models showed an association between estimated correction factors and the proposed adequacy indicators.
Multi-Language	The results of the regression models showed an association between the estimated correction factors and the proposed adequacy indicators.
French	(Traduit par Docteur Serge Messier).
Single-Language	[Doctor Serge Messier].
Multi-Language	[(Translated by Doctor Serge Messier)].

Table 2: Spanish/Portuguese/French to English examples for WMT18

were trained with same architecture and parameters.

Hparam	Text-to-Text
Encoder layers	6
Decoder layers	6
Batch size	4096
Adam optimizer	$\beta_1 = 0.9 \quad \beta_2 = 0.998$
Attention heads	8

Table 4: Training parameters.

We trained three single-language systems, one for each language pair. We required 14 epochs for the Spanish-to-English system (7 hours of training), 16 epochs for the French-to-English system (9 hours of training), and 17 epochs for the Portuguese-to-English system (7 hours of training). For the multi-source system, which concatenated the three parallel corpus together, we required 11 epochs (23 hours of training). We stopped training when the validation accuracy did not increase in two consecutive epochs.

4.3 Results

Best ranking systems from WMT17 and WMT18 are shown in Table 1, except for French-to-English WMT17 since the references for this set are not available. For this pair, we used 1000 sentences from the Khresmoi development data. Table 1 shows BLEU results for the baseline systems, the single-language and multi-source approaches.

The Transformer architecture outperforms WMT17 best system. Results become even better with the system is trained with the common corpus of Romance languages, what we call the multi-source approach. The latter is consistent with the universal truth that more data equals better results, even if the source language is not the same.

Finally, Table 2 shows some examples of the output translations.

5 Conclusions

The main conclusions of our experiments are that the multi-source inputs of the same family applied to the Transformer architecture can improve the single input. Best improvements achieve an increase of 6 BLEU points in translation quality.

Acknowledgments

Authors would like to thank Noe Casas for his valuable comments. This work is supported in

part by the Spanish Ministerio de Economía y Competitividad, the European Regional Development Fund and the Agencia Estatal de Investigación, through the postdoctoral senior grant Ramón y Cajal, the contract TEC2015-69266-P (MINECO/FEDER,EU) and the contract PCIN-2017-079 (AEI/MINECO).

References

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. Unsupervised neural machine translation. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

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

Marta R. Costa-jussà, Álvaro Nuez, and Carlos Segura. 2018. Experimental research on encoder-decoder architectures with attention for chatbots. *Computación y Sistemas*.

Laura Cros, Carlos Escolano, José A. R. Fonollosa, and Marta R. Costa-jussà. 2018. End-to-end speech translation with the transformer. *Submitted to IberSpeech*.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. 2017. Convolutional sequence to sequence learning. *CoRR*, abs/1705.03122.

Melvin Johnson, 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*, 5:339–351.

Lukasz Kaiser, Aidan N Gomez, Noam Shazeer, Ashish Vaswani, Niki Parmar, Llion Jones, and Jakob Uszkoreit. 2017. One model to learn them all. *arXiv preprint arXiv:1706.05137*.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72. Association for Computational Linguistics.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proc. of the 1st Workshop on Neural Machine Translation*, pages 28–39, Vancouver.

Guillaume Lample, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Unsupervised machine translation using monolingual corpora only. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6000–6010. Curran Associates, Inc.

Geoffrey E. Hinton Vinod Nair. 2010. Rectified linear units improve restricted boltzmann machines. In *27th International Conference on Machine Learning*.

Barret Zoph and Kevin Knight. 2016. Multi-source neural translation. In *NAACL-HLT 2016*, pages 30–34.