UoS Participation in the WMT20 Translation of Biomedical Abstracts

Felipe Soares
University of Sheffield - NLP Group
fs@felipesoares.net

Delton de Andrade Vaz
UFRGS
University of Montpellier
delton.vaz@gmail.comn

Abstract

This paper describes the machine translation systems developed by the University of Sheffield (UoS) team for the biomedical translation shared task of WMT20. Our system is based on a Transformer model with TensorFlow Model Garden toolkit. We participated in ten translation directions for the English/Spanish, English/Portuguese, English/Russian, English/Italian, and English/French language pairs. To create our training data, we concatenated several parallel corpora, both from in-domain and out-of-domain sources.

1 Introduction

In this paper, we present the system developed by the University of Sheffield for the Biomedical Translation shared task in the Fifth Conference on Machine Translation (WMT20), which consists in translating scientific texts from the biological and health domain.

Our participation in this task considered the English/Portuguese, English/Spanish, English/Russian, English/Italian, and English/French language pairs with translations in both directions. For that matter, we developed a machine translation (MT) system based on neural machine translation (NMT), using Google's TensorFlow Model Garden. [1]

2 Related Works

Previous participation in biomedical translation tasks include the works of Costa-Jussà et al. (2016) which employed Moses Statistic Machine Translation (SMT) to perform automatic translation integrated with a neural character-based recurrent neural network for model re-ranking and bilingual word embeddings for out of vocabulary (OOV) resolution. Given the 1000-best list of SMT translations, the RNN performs a re-scoring and selects the translation with the highest score. The OOV resolution module infers the word in the target language based on the bilingual word embedding trained on large monolingual corpora. Their reported results show that both approaches can improve BLEU scores, with the best results given by the combination of OOV resolution and RNN re-ranking. Similarly, Ive et al. (2016) also used the n-best output from Moses as input to a re-ranking model, which is based on a neural network that can handle vocabularies of arbitrary size.

More recently, Tubay and Costa-Jussà (2018) employed multi-source language translation using romance languages to translate from Spanish, French, and Portuguese to English. They used data from SciELO and Medline abstracts to train a Transformer model with individual languages to English and also with all languages concatenated to English.

In the last two WMT biomedical translation challenges (WMT18 and WMT19) (Neves et al., 2018; Bawden et al., 2019), the submissions that achieved the best BLEU scores for the ES/EN and PT/EN, in both directions (Soares and Becker, 2018; Tubay and Costa-Jussà, 2018; Carrino et al., 2019; Saunders et al., 2019; Soares and Krallinger, 2019), used the Transformer architecture with enhancements such as handling of terminology during tokenization (Carrino et al., 2019), multi-domain inference (Saunders et al., 2019) and exploitation of additional linguistic resources (Soares and Becker, 2018; Soares and Krallinger, 2019).

3 Resources

In this section, we describe the language resources used to train both models.

[1] https://github.com/tensorflow/models

Proceedings of the 5th Conference on Machine Translation (WMT), pages 870–874
Online, November 19–20, 2020.

3.1 Corpora

We used both in-domain and general domain corpora to train our systems. For general domain data, we used the ParaPat patent corpus (Soares et al., 2020), which is available for several languages, included the ones we explored in our systems. As for in-domain data, we included several different corpora:

- The corpus of full-text scientific articles from SciELO (Soares et al., 2018a), which includes articles from several scientific domains in the desired language pairs, but predominantly from biomedical and health areas.
- A subset of the UFAL medical corpus[2], containing the Medical Web Crawl data for the English/Spanish language pair.
- The EMEA corpus (Tiedemann, 2012), consisting of documents from the European Medicines Agency.
- A corpus of theses and dissertations abstracts (BDTD) (Soares et al., 2018b) from CAPES, a Brazilian governmental agency responsible for overseeing post-graduate courses. This corpus contains data only for the English/Portuguese language pair.
- A corpus from Virtual Health Library[3] (BVS), containing also parallel sentences for the language pairs explored in our systems.
- A corpus from SciELO (Neves et al., 2016), containing also parallel sentences from abstracts in English/Portuguese, English/Spanish, and English/French.

A new crawl of MEDLINE using the Ebot provided by the National Library of Medicine.[4]

Table 1 depicts the original number of parallel segments according to each corpora source. In Section 4.1, we detail the pre-processing steps performed on the data to comply with the task evaluation.

[2] https://ufal.mff.cuni.cz/ufal_medical_corpus

[3] http://bvsalud.org/

[4] https://www.ncbi.nlm.nih.gov/Class/PowerTools/eutils/ebot/ebot.cgi

4 Experimental Settings

In this section, we detail the pre-processing steps employed as well as the architecture of the Transformer.

4.1 Pre-processing

As detailed in the description of the biomedical translation task, the evaluation is based on texts extracted from MEDLINE. Since two of our corpora, the one comprised of full-text articles from SciELO and the new crawl from PubMed, may contain a considerable overlap with MEDLINE data, we decided to employ a filtering step in order to avoid including such data.

The first step in our filter was to download the parallel data from PubMed articles in Russian, French, and Italian. For that matter, we used the Ebot utility[5] provided by NLM using the queries *ITA[la]*, *FRE[la]*, and *RUS[la]*, retrieving all results available. Once downloaded, we performed sentence alignment using LF-Aligner[6]. To perform the filtering, we decided to use simple case insensitive string matching with *grep* supplying the option *-xvf* and the test set in English.

4.2 NMT System

As for the NMT system, we employed the official Google's implementation of the Transformer architecture (Vaswani et al., 2017) to train ten MT systems for the five language pairs. Tokenization was performed using the WordPiece unsupervised tokenizer with a vocabulary size of 32,000 on the initial training data, with a shared vocabulary between source and target.

For systems where the target language was English, back-translation was used with a number of sentences equals to the initial training system where English was the source. For the Spanish/English language pair, the system used to produce the artificial parallel sentences was the one developed by Soares and Krallinger (2019), while for the other language pairs we used the same systems trained by our team.

The parameters of our network for all language pairs excluding English/Portuguese are as follows. Encoder and Decoder: Transformer; Word vector size: 512; Layers for encoder and decoder:

[5] https://www.ncbi.nlm.nih.gov/Class/PowerTools/eutils/ebot/ebot.cgi

[6] https://sourceforge.net/projects/aligner/

Corpus	Sentences				
	EN/ES	EN/PT	EN/FR	EN/RU	EN/IT
ParaPat	-	-	-	3.28M	-
UFAL	286,779	-	1.6M	-	-
Abstract SciELO	767,069	669,629	-	-	-
Full-text SciELO	425,631	2.86M	-	-	-
EMEA	1.01M	1.08M	609,852	-	1.08M
CAPES-BDTD	-	950,252	-	-	-
BVS	-	931,946	10,812	-	-
MEDLINE (titles and abstracts)	-	-	582,007	11,271	1,298
Total	2.48M	6.49M	2.25M	3.28M	1.08M

Table 1: Original size of individual corpora used in our experiments

6; Attention heads: 16; RNN size: 512; Hidden transformer feed-forward: 2048; Batch size: 8196. For the English/Portuguese language pair, due to the large training set, we employed a bigger network as follows. Word vector size: 1024; Layers for encoder and decoder: 6; Attention heads: 16; RNN size: 1024; Hidden transformer feed-forward: 4096; Batch size: 8192.

To train our systems, we used 5 Tensor Processing Units (TPUs) v3, with a number of 250,000 steps (for all systems with exception of Russian, which was trained with fewer steps). The models with the best perplexity value were chosen as final models.

For the English/Russian language pair, incremental training was performed, since the size of the in-domain dataset was reduced. For such, we first trained our system in the out-of-domain data from patents for 100,000 steps. We then proceeded with additional training for 25,000 steps with in-domain data.

5 Results

We now detail the results achieved by our Transformer systems on the official test data used in the shared task regarding automatic evaluation. Table 2 shows the BLEU scores (Papineni et al., 2002) for our systems for the 10 language pairs we participated. For the Spanish and Portuguese language pairs we achieve high competitive results. For ES/EN, the best system (NLE) achieved BLEU of 0.5075, while the second best achieved BLEU of 0.4662 (TRAMECAT), very close to our result of 0.4624. For the opposite direction, EN/ES, the best system (UCAM) achieved 0.4662, the second best (Elhuyar_NLP) 0.4498, while our system scored 0.4493.

Language Pair	**BLEU**
EN/PT	0.4744
PT/EN	0.5334
EN/ES	0.4493
ES/EN	0.4624
EN/FR	0.3049
FR/EN	0.3514
EN/RU	0.2573
RU/EN	0.2936
EN/IT	0.2073
IT/EN	0.2276

Table 2: Official BLEU scores for the language pairs we submitted systems. These scores are evaluated on the "OK" aligned sentences.

For the Portuguese language, in both directions we achieved the best scores, with an EN/PT BLEU of 0.4744 and PT/EN of 0.5334. The second team in both languages (UNICAMP_DL) achieved scores of 0.4095 and 0.4988, respectively.

As for the Russian, French, and Italian languages, our scores were not as competitive as the best systems, with the exception of FR/EN, which we stood as 3 out of 5 teams. After carefully checking our training data, we found encoding issues with the different gathered data for those languages, especially with the encoding and tokenization of words containing apostrophes in French and Italian, as well as the Cyrillic Kha.

6 Conclusions

We presented the University of Sheffield (UoS) machine translation system for the biomedical translation shared task in WMT20. For our submission, we trained ten Transformers NMT systems, employing different corpora for each language pair. In addition, for systems with English as target language, back-translation was used, and for the Russian language, incremental training from Patent abstracts was used.

For model building, we included several corpora from biomedical and health domain, and from out-of-domain data that we considered to have similar textual structure, such as books and patents. Prior training, we also pre-processed our corpora to ensure that we did not include any sentence from the released test set, which could produce biased models.

Regarding future work, we are planning on optimizing our systems by performing pre-selection of out-of-domain data, aiming at selecting only the most similar sentences to the in-domain data. In addition, we plan to explore the potential use of domain-specific decoding, as proposed in Saunders et al. (2019).

Acknowledgements

This work was supported by Amazon AWS Cloud Credits for Research, which were used for corpora processing and gathering, and by Google TensorFlow Research Cloud credits, which were used for model training and inference.

References

Rachel Bawden, Kevin Bretonnel Cohen, Cristian Grozea, Antonio Jimeno Yepes, Madeleine Kittner, Martin Krallinger, Nancy Mah, Aurelie Neveol, Mariana Neves, Felipe Soares, Amy Siu, Karin Verspoor, and Maika Vicente Navarro. 2019. Findings of the wmt 2019 biomedical translation shared task: Evaluation for medline abstracts and biomedical terminologies. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 31–55, Florence, Italy. Association for Computational Linguistics.

Casimiro Pio Carrino, Bardia Rafieian, Marta R. Costa-jussà, and JosÃ© A. R. Fonollosa. 2019. Terminology-aware segmentation and domain feature for the wmt19 biomedical translation task. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 153–157, Florence, Italy. Association for Computational Linguistics.

Marta R Costa-Jussà, Cristina España-Bonet, Pranava Madhyastha, Carlos Escolano, and José AR Fonollosa. 2016. The talp–upc spanish–english wmt biomedical task: Bilingual embeddings and char-based neural language model rescoring in a phrase-based system. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, volume 2, pages 463–468.

Julia Ive, Aurélien Max, and François Yvon. 2016. Limsi's contribution to the wmt'16 biomedical translation task. In *First Conference on Machine Translation*, volume 2, pages 469–476.

Mariana Neves, Antonio Jimeno Yepes, Aurélie Névéol, Cristian Grozea, Amy Siu, Madeleine Kittner, and Karin Verspoor. 2018. Findings of the wmt 2018 biomedical translation shared task: Evaluation on medline test sets. In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 328–343, Belgium, Brussels. Association for Computational Linguistics.

Mariana Neves, Antonio Jimeno Yepes, and Aurélie Névéol. 2016. The scielo corpus: a parallel corpus of scientific publications for biomedicine. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

Kishore Papineni, 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 on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Danielle Saunders, Felix Stahlberg, and Bill Byrne. 2019. Ucam biomedical translation at wmt19: Transfer learning multi-domain ensembles. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 171–176, Florence, Italy. Association for Computational Linguistics.

Felipe Soares and Karin Becker. 2018. Ufrgs participation on the wmt biomedical translation shared task. In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 673–677, Belgium, Brussels. Association for Computational Linguistics.

Felipe Soares and Martin Krallinger. 2019. Bsc participation in the wmt translation of biomedical abstracts. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 177–180, Florence, Italy. Association for Computational Linguistics.

Felipe Soares, Viviane Moreira, and Karin Becker. 2018a. A Large Parallel Corpus of Full-Text Scientific Articles. In *Proceedings of the Eleventh International Conference on Language Resources and*

Evaluation (LREC 2018), Miyazaki, Japan. European Language Resources Association (ELRA).

Felipe Soares, Mark Stevenson, Diego Bartolome, and Anna Zaretskaya. 2020. ParaPat: The multi-million sentences parallel corpus of patents abstracts. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 3769–3774, Marseille, France. European Language Resources Association.

Felipe Soares, Gabrielli Yamashita, and Michel Anzanello. 2018b. A parallel corpus of theses and dissertations abstracts. In *The 13th International Conference on the Computational Processing of Portuguese (PROPOR 2018)*, Canela, Brazil. Springer International Publishing.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Brian Tubay and Marta R. Costa-Jussà. 2018. Neural machine translation with the transformer and multi-source romance languages for the biomedical wmt 2018 task. In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 678–681, Belgium, Brussels. 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.