BSC Participation in the WMT Translation of Biomedical Abstracts

Felipe Soares
Barcelona Supercomputing Center (BSC)
felipe.soares@bsc.es

Martin Krallinger
Centro Nacional de Investigaciones Oncológicas (CNIO)
Barcelona Supercomputing Center (BSC)
martin.krallinger@bsc.es

Abstract

This paper describes the machine translation systems developed by the Barcelona Supercomputing (BSC) team for the biomedical translation shared task of WMT19. Our system is based on Neural Machine Translation unsing the OpenNMT-py toolkit and Transformer architecture. We participated in four translation directions for the English/Spanish and English/Portuguese language pairs. To create our training data, we concatenated several parallel corpora, both from in-domain and out-of-domain sources, as well as terminological resources from UMLS.

1 Introduction

In this paper, we present the system developed at the Barcelona Supercomputing Center (BSC) for the Biomedical Translation shared task in the Fourth Conference on Machine Translation (WMT19), which consists in translating scientific texts from the biological and health domain.

Our participation in this task considered the English/Portuguese and English/Spanish language pairs, with translations in both directions. For that matter, we developed a machine translation (MT) system based on neural machine translation (NMT), using OpenNMT-py (Klein et al., 2017).

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 rescoring 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-jussÃ (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 WMT biomedical translation challenge (2018) (Neves et al., 2018), the submission that achieved the best BLEU scores for the ES/EN and PT/EN, in both directions, were the ones submited by the UFRGS team (Soares and Becker, 2018), followed by the TALP-UPC (Tubay and Costa-jussÃ, 2018) in the ES/EN direction and the UHH-DS in the EN/PT directions.

3 Resources

In this section, we describe the language resources used to train both models, which are from two main types: corpora and terminological resources.

3.1 Corpora

We used both in-domain and general domain corpora to train our systems. For general domain data, we used the books corpus (Tiedemann, 2012), which is available for several languages, included the ones we explored in our systems, and the JRC-Acquis (Tiedemann, 2012). As for in-domain data, we included several different corpora:

Proceedings of the Fourth Conference on Machine Translation (WMT), Volume 3: Shared Task Papers (Day 2) pages 177–180
Florence, Italy, August 1-2, 2019.

- 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[1], 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[2] (BVS), containing also parallel sentences for the language pairs explored in our systems.

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.

Corpus	Sentences	
	EN/ES	EN/PT
Books	93,471	-
UFAL	286,779	-
Full-text Scielo	425,631	2.86M
JRC-Acquis	805,757	1.64M
EMEA	-	1.08M
CAPES-BDTD	-	950,252
BVS	737,818	631,946
Total	2.37M	7.19M

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

3.2 Terminological Resources

Regarding terminological resources, we extracted parallel terminologies from the Unified Medical Language System[3] (UMLS). For that matter, we used the MetamorphoSys application provided by U.S. National Library of Medicine (NLM) to subset the language resources for our desired language pairs. Our approach is similar to what was proposed by Perez-de Viñaspre and Labaka (2016).

Once the resource was available, we imported the MRCONSO RRF file to an SQL database to split the data in a parallel format in the two language pairs. Table 2 shows the number of parallel concepts for each pair.

Language Pair	Concepts
EN/ES	14,399
EN/PT	26,194

Table 2: Number of concepts from UMLS for each language pair

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 one of our corpora, the one comprised of full-text articles from Scielo, 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 metadata from Pubmed articles in Spanish and Portuguese. For that matter, we used the Ebot utility[4] provided by NLM using the queries *POR[la]* and *ESP[la]*, retrieving all results available. Once downloaded, we imported them to an SQL database which already contained the corpora metadata. To perform the filtering, we used the *pii* field from Pubmed to match the Scielo unique identifiers or the title of the papers, which would match documents not from Scielo.

Once the documents were matched, we removed them from our database and partitioned the data in training and validation sets. Table 3 contains the final number of sentences for each language pair and partition.

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

[2] http://bvsalud.org/

[3] https://www.nlm.nih.gov/research/umls/

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

Language	Train	Dev
EN/ES	2.35M	22,670
EN/PT	7.17M	24,206

Table 3: Final corpora size for each language pair

4.2 NMT System

As for the NMT system, we employed the OpenNMT-py toolkit (Klein et al., 2017) to train three MT systems, one for (Spanish,Portuguese)→English, another one for (English,Spanish)→Portuguese and a third one for (English,Portuguese)→Spanish. Tokenization was performed using the SentecePiece[5] unsupervised tokenizer with a vocabulary size of 32,000. The tokenization was done for each MT system (e.g. concatenated English, Spanish and Portuguese to generate one of the models).

The parameters of our network are as follows. Encoder and Decoder: Transformer; Word vector size: 1024; Layers for encoder and decoder: 6; Attention heads: 16; RNN size: 1024; Hidden transformer feed-forward: 4096; Batch size: 4096.

To train our system, we used the an IBM cluster with 2 Power-9 CPUs and with four NVIDIA Tesla V100 GPUs. The models with the best perplexity value were chosen as final models. During translation, OOV words were replace by their original word in the source language, all other OpenNMT-py options for translation were kept as default.

5 Results

We now detail the results achieved by our Transformer systems on the official test data used in the shared task. Table 4 shows the BLEU scores (Papineni et al., 2002) for our systems and for the submissions made by other teams. For the ES/EN language pair, we figured in 5 out of 11, while for EN/ES in 4 out of 8.

However, one should also take in account the confidence interval of the average of the results. By performing a t-test on the ES/EN results, we found out that the mean of the BLEU scores is 0.4366 (p-value < 0.01 with confidence interval (95%) between 0.4145 and 0.4857. This means that only the submissions from UCAM can be said to be better than the average. Similarly, the team from UHH-DS is has statistically lower performance than the average. Meanwhile, all other teams, including ours, are statistically tied around the mean, meaning that there is no sufficient information to difference the performance from one system to another.

Similarly, for the EN/ES language pair, we performed the same statistical test and achieved p-value < 0.01. The reported mean is 0.4675, with confidence interval (95%) between 0.4489 and 0.4861. Thus, Only submissions 2 and 3 from UCAM can be said to be better than average, while the submission from MT-UOC-UPF performed worse than the average. All other teams, including ours, are statistically tied around the mean, without evidence that there is any significant difference among the systems.

Unfortunately, no other team participated on the PT/EN and EN/PT language pairs.

6 Conclusions

We presented the BSC machine translation system for the biomedical translation shared task in WMT19. For our submission, we trained three Transformers NMT systems with multilingual implementation for the English/Spanish and English/Portuguese language pairs.

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 JRC-Acquis and books. Prior training, we also pre-processed our corpora to ensure, or at least minimize the risk, of including Medline data in our training set, which could produce biased models, since the evaluation was carried out on texts extracted from Medline.

Regarding future work, we are planning on optimizing our systems by studying the use of synthetic data from back-translation of monolingual to increase NMT performance (Sennrich et al., 2016) by providing additional training data.

Acknowledgements

This work was supported by the Encargo de Gestion SEAD-BSC of the Spanish National Plan for the Advancement of Language technologies, the ICTUSnet INTERREG Sudoe programme, the European Union Horizon2020 eTransafe (grant agreemed 777365) project, and the Amazon AWS Cloud Credits for Research.

[5] https://github.com/google/sentencepiece

Teams	Runs	ES/EN	EN/ES	PT/EN	EN/PT
BSC	1	0.4356	0.4701	0.3990	0.4811
MT-UOC-UPF	1	0.4159	0.4219	-	-
Talp_upc	1	0.4509	0.4568	-	-
Talp_upc	2	0.4355	0.4609	-	-
Talp_upc	3	0.4270	0.4683	-	-
UCAM	1	0.4770	0.4834	-	-
UCAM	2	**0.4833**	0.4891	-	-
UCAM	3	0.4811	**0.4896**	-	-
UHH-DS	1	0.3969	-	-	-
UHH-DS	2	0.3999	-	-	-
UHH-DS	3	0.3997	-	-	-

Table 4: Official BLEU scores for the English/Spanish and English/Portuguese language pairs in both translation directions for the well aligned sentences of the test set. Bold numbers indicate the best result for each direction.

References

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.

G. Klein, Y. Kim, Y. Deng, J. Senellart, and A. M. Rush. 2017. OpenNMT: Open-Source Toolkit for Neural Machine Translation. *ArXiv e-prints*.

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.

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.

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)*, volume 1, pages 86–96.

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, 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, 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-jussÃ. 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.

Olatz Perez-de Viñaspre and Gorka Labaka. 2016. Ixa biomedical translation system at wmt16 biomedical translation task. In *Proceedings of the First Conference on Machine Translation*, pages 477–482, Berlin, Germany. Association for Computational Linguistics.