UFRGS Participation on the WMT Biomedical Translation Shared Task

Felipe Soares
Instituto de Informtica - UFRGS
Porto Alegre - RS - Brazil
felipe.soares@inf.ufrgs.br

Karin Becker
Instituto de Informtica - UFRGS
Porto Alegre - RS - Brazil
karin.becker@inf.ufrgs.br

Abstract

This paper describes the machine translation systems developed by the Universidade Federal do Rio Grande do Sul (UFRGS) team for the biomedical translation shared task. Our systems are based on statistical machine translation and neural machine translation, using the Moses and OpenNMT toolkits, respectively. 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. Our systems achieved the best BLEU scores according to the official shared task evaluation.

1 Introduction

In this paper, we present the system developed at the Universidade Federal do Rio Grande do Sul (UFRGS) for the Biomedical Translation shared task in the Third Conference on Machine Translation (WMT18), which consists in translating scientific texts from the biological and health domain. In this edition of the shared task, six language pairs are considered: English/Chinese, English/French, English/German, English/Portuguese, English/Romanian, and English/Spanish.

Our participation in this task considered the English/Portuguese and English/Spanish language pairs, with translations in both directions. For that matter, we developed two machine translation (MT) systems: one based on statistical machine translation (SMT), using Moses (Koehn et al., 2007), and one using neural machine translation (NMT), using OpenNMT (Klein et al., 2017).

This paper is structured as follows: Section 3 details the language resources used to train our translation models. Section 4 contains the description of the experimental settings of our SMT and NMT models, including the pre-processing step performed to comply with the shared task guidelines. In Section 5 we present the results and briefly discuss the main findings. Section 6 contains the conclusions and directions of future works to improve our models.

2 Related Works

Most of related works in biomedical machine translation used SMT models to perform automatic translation. Aires et al. (2016) developed a phrase-based SMT that differs significantly from the usual Moses toolkit, especially by not analyzing phrases at word level and adopting a translation score that is a tuned weighted average between the translation model and the language model, instead of the traditional log-linear approach.

Costa-Jussà et al. (2016) employed Moses 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.

In the last WMT biomedical translation chal-

Proceedings of the Third Conference on Machine Translation (WMT), Volume 2: Shared Task Papers, pages 662–666
Belgium, Brussels, October 31 - Novermber 1, 2018. ©2018 Association for Computational Linguistics
https://doi.org/10.18653/v1/W18-64075

lenge (2017) (Yepes et al., 2017), the submission that achieved the best BLEU scores for the FR/EN language pair on the EDP dataset, in both directions, was based on NMT models developed in the University of Kyoto (Cromieres, 2016). For the other datasets, the submission from the University of Edinburgh (Sennrich et al., 2017) achieved the best BLEU scores with their NMT models based on the Nematus implementation with BPE tokenization and the use of parallel and backtranslated data.

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:

- 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.

[1] https://ufal.mff.cuni.cz/ufal_medical_corpus
[2] http://bvsalud.org/

Table 1 depicts the original number of parallel segments according to each corpora source. In Section 3.1, 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 SMT and NMT systems.

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

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.

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 SMT System

We used the popular Moses toolkit (Koehn et al., 2007) to train our SMT system for the two language pairs. As training parameters, we followed the Moses baseline steps[5] to train four MT systems (i.e. one for each translation direction).

Regarding training, we used the Amazon AWS spot virtual machines with 24 cores and 60GB of RAM, and used parallelization as much as possible to reduce training time and the associated cost.

4.3 NMT System

As for the NMT system, we employed the OpenNMT toolkit (Klein et al., 2017) to train four MT systems, one for each translation direction. Tokenization was performed by the supplied OpenNMT algorithm. Regarding network parametrization, the following settings were used, while all other parameters were set as default:

- Encoder type: bidirectional recurrent neural network
- Decoder type: Seq2Seq with attention (default)
- Word vector size: 600
- Layers (encoder and decoder): 4
- RNN size: 800
- Batch size: 64
- Vocabulary size: 50000

To train our system, we used the Azure virtual machines with a single NVIDIA Tesla V100 GPU. 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 options for translation were kept as default.

5 Experimental Results

We now detail the results achieved by our SMT and NMT systems on the official test data used in the shared task. Table 4 shows the BLEU scores (Papineni et al., 2002) for both systems and for the submissions made by other teams.

Our submissions achieved the best results for all translation directions we participated, with remarkable BLEU scores for the ES/EN and PT/EN pairs. When compared to the other teams, our results presented similar behavior, with higher scores when English was the target language, which may be explained by the poor English morphosyntactic system. For the English/Spanish pair, the SMT system presented slightly better results than the NMT one, probably due to the dictionary size used in the NMT.

Regarding the superior results achieved, we expect that the large parallel corpora used in our experiments played an essential role. Although we did not use the provided Scielo abstracts corpus (Neves et al., 2016), we used a newer parallel corpus also from Scielo, but comprised of full-text articles (Soares et al., 2018a), which overlaps with the abstracts, but contains more data.

In addition to the biomedical and health corpora, we employed two out-of-domain corpora that we assumed to have a similar structure to scientific texts: the books and the JRC-Acquis (Tiedemann, 2012). We decided not to use the

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

[5] http://www.statmt.org/moses/?n=moses.baseline

Team, Runs	EN/ES	EN/PT	ES/EN	PT/EN
UFRGS run1 (NMT)	39.62	**39.43**	43.31	**42.58**
UFRGS run2 (SMT)	**39.77**	**39.43**	**43.41**	**42.58**
TGF TALP UPC run1	-	-	40.49	39.49
TGF TALP UPC run2	-	-	39.06	38.54
UHH-DS run1	31.32	34.92	36.16	41.84
UHH-DS run2	31.05	34.19	35.17	41.80
UHH-DS run3	31.33	34.49	36.05	41.79

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

large Europarl corpus (Koehn, 2005), since it is comprised of speeches transcripts, which do not follow the usual structure of scientific texts.

6 Conclusions

We presented the UFRGS machine translation systems for the biomedical translation shared task in WMT18. For our submissions, we trained SMT and NMT systems for all four translation directions 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.

Our systems achieved the best results in this shared task for the translation directions we participated, which we attribute to the high quality corpora used and their size.

Regarding future work, we are planning on optimizing our systems by studying the following methods:

- BPE tokenization: as stated by Sennrich et al. (2016b), the use of byte pair encoding tokenization can help to tackle the issue of OOV words by using subword units. We expect that this approach can provide better results for our NMT system on biomedical data, since this domain contains terminologies that are usually based on the use of affixes.

- Backtranslation: the use of synthetic data from back-translation of monolingual proved to be able to increase NMT performance (Sennrich et al., 2016a) by providing additional training data.

- Multilingual training: a study from Google (Johnson et al., 2017) showed that using multilingual data when training NMT systems can improve translation performance, especially when using a many-to-one scheme (i.e. several source languages and one target language). We expect that systems trained using (ES+PT)→EN, for instance, may produce better results due to the similarity between Portuguese and Spanish.

References

José Aires, Gabriel Lopes, and Luís Gomes. 2016. English-portuguese biomedical translation task using a genuine phrase-based statistical machine translation approach. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, volume 2, pages 456–462.

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.

Fabien Cromieres. 2016. Kyoto-nmt: a neural machine translation implementation in chainer. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, pages 307–311.

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.

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 of Computational Linguistics*, 5(1):339–351.

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

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, pages 177–180. Association for Computational Linguistics.

Mariana Neves, Antonio Jimeno Yepes, and Aurlie Nvol. 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.

Rico Sennrich, Alexandra Birch, Anna Currey, Ulrich Germann, Barry Haddow, Kenneth Heafield, Antonio Valerio Miceli Barone, and Philip Williams. 2017. The university of edinburgh's neural mt systems for wmt17. *arXiv preprint arXiv:1708.00726*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. 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.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1715–1725.

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.

Jrg 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).

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.

Antonio Jimeno Yepes, Aurélie Névéol, Mariana Neves, Karin Verspoor, Ondrej Bojar, Arthur Boyer, Cristian Grozea, Barry Haddow, Madeleine Kittner, Yvonne Lichtblau, et al. 2017. Findings of the wmt 2017 biomedical translation shared task. In *Proceedings of the Second Conference on Machine Translation*, pages 234–247.