Terminology-Aware Segmentation and Domain Feature for the WMT19 Biomedical Translation Task

Casimiro Pio Carrino, Bardia Rafieian, Marta R. Costa-jussà, José A. R. Fonollosa
{casimiro.pio.carrino, bardia.rafieian}@upc.edu,
{marta.ruiz, jose.fonollosa}@upc.edu,
TALP Research Center
Universitat Politècnica de Catalunya, Barcelona

Abstract

In this work, we give a description of the TALP-UPC systems submitted for the WMT19 Biomedical Translation Task. Our proposed strategy is NMT model-independent and relies only on one ingredient, a biomedical terminology list. We first extracted such a terminology list by labelling biomedical words in our training dataset using the BabelNet API. Then, we designed a data preparation strategy to insert the terms information at a token level. Finally, we trained the Transformer model (Vaswani et al., 2017) with this terms-informed data. Our best-submitted system ranked 2nd and 3rd for Spanish-English and English-Spanish translation directions, respectively.

1 Introduction

Domain adaptation in Neural Machine Translation (NMT) remains one of the main challenges (Koehn and Knowles, 2017). Domain-specific translations are especially relevant for industrial applications where there is a need for achieving both fluency and terminology in translations. Current state-of-the-art NMT systems achieve high performances when trained with large-scale parallel corpora. However, most of the time, large-scale parallel corpora are not available for specific domains. Consequently, NMT models perform poorly for domain-specific translation when trained in low-resource scenario (Chu and Wang, 2018). Several works have been proposed to overcome the lack of domain parallel data by leveraging on both monolingual domain data (Domhan and Hieber, 2017; Currey et al., 2017) and parallel out-of-domain data (Wang et al., 2017; van der Wees et al., 2017) to improve the performance of domain-specific systems. Furthermore, some attempts have been made to directly insert external knowledge into NMT models through terminology (Chatterjee et al., 2017) and domain information (Kobus et al., 2016). In this work, we designed a data preparation strategy for domain-specific translation systems to enrich data with terminology information without affecting the model architecture. The approach consists on two main steps: 1) Retrieve a biomedical terms list from on our training data 2) use terms to add a domain feature on the source side and define a terminology-aware segmentation. The data preparation is a model-independent process which generates terms-informed token representations that can be used to train any NMT model. For the Biomedical WMT19 task, we decided to train one of the state-of-the-art neural models, the transformer (Vaswani et al., 2017). In our knowledge, this is the first attempt to design a domain-specific text segmentation based on a given terminology list. The rest of the paper is organized as follows. In Sec. 2, we described how terminology is extracted from BabelNet; in Sec. 3 and 4, we defined the terminology-aware segmentation and the domain feature approach, respectively; in Sec. 5, we described the experiments performed, the performance evaluation and the results of the WMT19 competition. Finally, Sec. 6 describes the conclusion and future works.

2 BabelNet

In our work, in order to collect biomedical terms, the domain category of each word was detected with the help of BabelNet (Navigli and Ponzetto, 2012). Specifically, we extracted a list of biomedical terms from our training data using the BabelNet API. To capture biomedical-related domains, we refer to the "biomedical" definition in the BabelNet as stated, "The science of dealing with the maintenance of health and the prevention and treatment of disease". Moreover,

Proceedings of the Fourth Conference on Machine Translation (WMT), Volume 3: Shared Task Papers (Day 2) pages 151–155
Florence, Italy, August 1-2, 2019. ©2019 Association for Computational Linguistics

a biomedical word has BabelNet relations with bio-science, technology, medical practice, medical speciality, neurology and orthopaedics. Consequently, we identified related BabelNet domains to the "biomedical" domain which are: Health and Medicine, Chemistry and Mineralogy, Biology and Engineering and Technology. Based on these domains, we then used the BabelNet API to find the domain of each word in the training dataset by searching through the BabelNet multilingual dictionary. Since a word can have multiple Babel synsets and domains, we collected a domain according to the key concept of a word. For our experiments, we created a list of 10,000 biomedical terms for both English and Spanish.

3 Terminology-aware segmentation

We propose the so-called "bpe-terms segmentation" consisting of both subwords and terms tokens. The idea is to overcome the open-vocabulary problem with subwords and at the same time have the ability to add domain features for terms at the word level. The procedure is rather simple. After learning the bpe codes (Sennrich et al., 2015), they are applied to segment the sentences by explicitly excluding terms belonging to a given domain terminology list. The resulting sentence is a mixture of both subwords and term tokens. In Table 1, we show the differences between standard bpe-segmentation and our bpe-terms segmentation. Unlike general domain words, biomedical terms are not divided into subwords producing a shorter sequence of tokens. It is also important to notice that all the terms that are not present in the terminology list, like "hypertension" and "clot" in the examples, might be split into subwords. These examples show how the effectiveness of bpe-term segmentation depends entirely on the size and quality of the terminology list.

4 Domain features

Following the domain control approach (Kobus et al., 2016), we enrich the data with a word-level binary feature by means of the biomedical terminology. Every word belonging to the terminology list has been labelled as biomedical, while all others as a general domain. The resulting binary feature is then embedded into a dense vector and combined with the word vector. The most common combination strategy consists in concatenating the feature embedding with the word embedding. However, we introduced an additional Multi-Layer perception with one hidden layer after the concatenation. This operation maps the resulting embedding into a new vector that might be more useful for the translation task. More precisely, given the word embedding $\mathbf{x_w} \in R^n$ and the feature embedding $\mathbf{x_f} \in R^m$, the resulting vector $\hat{\mathbf{x}} \in R^d$ is computed as:

$$\hat{\mathbf{x}} = g([\mathbf{x_w}, \mathbf{x_f}]\mathbf{W} + \mathbf{b})$$

where $\mathbf{W} \in R^{n+m,d}$ is the weight matrix, $\mathbf{b} \in R^d$ is the bias term and g is a nonlinear functions for the hidden layer that is applied element-wise. In our experiments, due to the binary nature of the domain feature, we set $m = 3$ as its embedding dimension. The word embedding dimension is set to $n = 512$ instead.

5 Experiments

This section describes the experiments we performed. We first start with the data collection and preprocessing processes. Then, we describe trained systems and their evaluations. Finally, we present the results of the competition in terms of BLEU score. (Papineni et al., 2002).

5.1 Data collection

We gathered data from the resources provided in the official WMT19 web page and from the OPUS collection. For our submissions, all the available biomedical parallel sentences for en/es are chosen both in plain text and Dublin Core format. Then, data have been parsed and merged to create the training and validation sets. Finally, we cleaned the datasets by removing empty sentences and duplicates. In particular, we selected Scielo (Soares et al., 2018), (Neves et al., 2016), UFAL, Pubmed, Medline, IBECS (Villegas et al., 2018) and EMEA (Tiedemann, 2012) sources for the training set and Khresmoi (Dušek et al., 2017) for the validation set.

5.2 Data preprocessing

Data are preprocessed following the standard pipeline by normalizing punctuation, tokenization and true-casing. We also removed sentences longer than 80 tokens and shorter than 2 tokens. For the previous steps, we used the scripts found in the Moses distribution (Koehn et al., 2007). Eventually, we trained shared byte-pairs encoding (BPE) (Sennrich et al., 2015) on both source and

Segmentation	Sentence
Bpe	”the intr@@ ig@@ u@@ ing pro@@ ble@@ m of **cal@@ ci@@ fic@@ ation** and **os@@ s@@ ific@@ ation** ; ne@@ ed to un@@ der@@ st@@ and it for the comp@@ re@@ h@@ ens@@ ion of **b@@ one phys@@ io@@ path@@ ology** .” ”inhibition of **T@@ AF@@ I** activity also resulted in a tw@@ of@@ old increase in clot lysis whereas inhibition of both factor XI and **T@@ AF@@ I** activity had no additional effect . ” ”a 5@@ 7-@@ year-old male with **hepatos@@ plen@@ omegaly** , **p@@ ancy@@ topenia** and hypertension .”
Bpe-terms	”the intr@@ ig@@ u@@ ing pro@@ ble@@ m of **calcification** and **ossification** ; ne@@ ed to un@@ der@@ st@@ and it for the comp@@ re@@ h@@ ens@@ ion of **bone physiopathology** .” inhibition of **TAFI** activity also resulted in a tw@@ of@@ old increase in clot lysis whereas inhibition of both factor XI and **TAFI** activity had no additional effect . ”a 5@@ 7-@@ year-old male with **hepatosplenomegaly** , **pancytopenia** and hypertension .”

Table 1: Different segmentation for some sample sentences extracted from the training data. Biomedical terms are in bold type to highlight the effect of the segmentation on them.

	Training set	**Validation set**
es/en	2812577	500

Table 2: The total number of parallel sentences in the training and validation sets after the preprocessing step.

target data with a number of maximum BPE symbols of 50k. The statistics of the final datasets in terms of the total number of lines are shown in Table 2.

5.3 Training with data enriched with terms information

Our strategy involves a data preparation designed to enrich the sentences with terminology information at the token level before the actual training takes place. There are two important components, the bpe-terms segmentation and the domain feature approach as explained in Sec. 3 and Sec. 4. Both of them are based on the terminology list that was created using the BabelNet API as described in Sec 2. The bpe-terms segmentation is applied to both the source and target side. Instead, the domain feature approach is applied only on the source side. After that, the resulting terms-informed data are used to train the NMT Transformer model. (Vaswani et al., 2017). Thereafter, three different experiments have been performed:

1. The first experiment combined both the terminology-aware segmentation and the domain feature.
2. The second, instead, make just use of the bpe-terms segmentation.
3. The third experiment combined both the terminology-aware segmentation and the domain feature. Additionally, both the vocabularies among source and target and the embedding weights between encoder and decoder are shared during the training.

	en2es	**es2en**
System	WMT18	WMT18
baseline	40.84	43.70
bpe-terms src-tgt + domain feature	**44.26**	43.49
bpe-terms src-tgt + shared vocab & embs	44.04	43.84
bpe-terms src-tgt	44.09	**44.84**

Table 3: The BLEU scores calculated on the WMT18 test set for the three systems compared with the baseline.

	en2es		**es2en**	
System	WMT19 (All)	WMT19 (OK)	WMT18 (All)	WMT19 (OK)
bpe-terms src-tgt	43.40	46.09	37.92	43.55
bpe-terms src-tgt + domain feature	43.01	45.68	37,21	42.70
bpe-terms src-tgt + shared vocab & embs	**43.92**	**46.83**	**39.41**	**45.09**

Table 4: The BLEU scores calculated on the WMT19 test set for the three systems.

Furthermore, we trained a baseline model with standard BPE segmentation to make a comparison with the three proposed experiments. All the models have maximum vocabulary size of 50k tokens. However, the final vocabulary size is affected by both the bpe-terms segmentation and the shared vocabularies between source and target side. It turns out that only the baseline and the third experiment had a vocabulary size of 50k tokens. For the training, we used the Transformer (Vaswani et al., 2017) implementation with its default parameters found in the OpenNMT toolkit (Klein et al.).

5.4 Evaluation and results

We evaluated all the models calculating the BLEU score on the WMT18 test set with the 'multi-bleu-detok.sh' script in the Moses distribution (Koehn et al., 2007). For the WMT19 competition, we first calculated the averages of the training checkpoints that achieved the highest BLEU scores on the validation set. Then, we submitted these averages as our best models. The results for both WMT18 and WMT19 test sets are shown in table 3 and 4. In Table 5, we also calculated how many biomedical terms are found in the validation and WMT18/WMT19 test sets to have an idea of the coverage of the terminology list on the out-of-training data. On the WMT18 test set, our proposed models performed better than the baseline, indicating that the Transformer model (Vaswani et al., 2017) took advantages from the bpe-terms segmentation. On the contrary, the domain feature approach overall hurts the test set performances. The best performing system evaluated on the WMT19 test set is the one with bpe-terms segmentation plus shared vocabulary and embedding layers for both source/target and encoder/decoder layers, respectively, showing consistency across both es/en direction. As a result, we placed 2nd for es2en and 3rd for en2es in the WMT19 competition.

	Validation set	**WMT18**	**WMT19**
es	713	355	399
en	831	363	502

Table 5: The number of biomedical terms from the terminology list found in the validation set and the WMT18 and WMT19 test sets.

6 Conclusions and future works

In this article, we described the TALP-UPC systems submitted to the WMT19 Biomedical Translation Task. Our experiments show an NMT model-independent approach that benefits from terminology to improve translations in the biomedical domain. The future efforts will be devoted to extending our bpe-terms segmentation by taking into account multi-word terms extracted from available biomedical glossaries and collecting a terminology list independent from training data.

Acknowledgments

This work is partially supported by Lucy Software / United Language Group (ULG) and the Catalan Agency for Management of University and Research Grants (AGAUR) through an Industrial PhD Grant. This work is also 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, contract TEC2015-69266-P (MINECO/FEDER, EU) and contract PCIN-2017-079 (AEI/MINECO).

References

Rajen Chatterjee, Matteo Negri, Marco Turchi, Marcello Federico, Lucia Specia, and Frédéric Blain. 2017. Guiding neural machine translation decoding with external knowledge. In *WMT*.

Chenhui Chu and Rui Wang. 2018. A survey of domain adaptation for neural machine translation. *CoRR*, abs/1806.00258.

Anna Currey, Antonio Valerio Miceli Barone, and Kenneth Heafield. 2017. Copied monolingual data improves low-resource neural machine translation. In *WMT*.

Tobias Domhan and Felix Hieber. 2017. Using target-side monolingual data for neural machine translation through multi-task learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1500–1505, Copenhagen, Denmark. Association for Computational Linguistics.

Ondřej Dušek, Jan Hajič, Jaroslava Hlaváčová, Jindřich Libovický, Pavel Pecina, Aleš Tamchyna, and Zdeňka Urešová. 2017. Khresmoi summary translation test data 2.0. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

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

Catherine Kobus, Josep Maria Crego, and Jean Senellart. 2016. Domain control for neural machine translation. *CoRR*, abs/1612.06140.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 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*, ACL '07, pages 177–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Roberto Navigli and Simone Paolo Ponzetto. 2012. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217 – 250.

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*, ACL '02, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *CoRR*, abs/1508.07909.

Felipe Soares, Viviane Moreira, and Karin Becker. 2018. A large parallel corpus of full-text scientific articles. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*. European Language Resource Association.

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).

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 5998–6008. Curran Associates, Inc.

Marta Villegas, Ander Intxaurrondo, Aitor Gonzalez-Agirre, Montserrat Marimon, and Martin Krallinger. 2018. The mespen resource for english-spanish medical machine translation and terminologies: Census of parallel corpora, glossaries and term translations. *Language Resources and Evaluation*.

Rui Wang, Andrew Finch, Masao Utiyama, and Eiichiro Sumita. 2017. Sentence embedding for neural machine translation domain adaptation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 560–566, Vancouver, Canada. Association for Computational Linguistics.

Marlies van der Wees, Arianna Bisazza, and Christof Monz. 2017. Dynamic data selection for neural machine translation. *CoRR*, abs/1708.00712.