A Multilingual Neural Machine Translation Model for Biomedical Data

Alexandre Bérard*, Zae Myung Kim**, Vassilina Nikoulina*, Eunjeong L. Park**, and Matthias Gallé*

*Naver Labs Europe
**Naver Papago

Abstract

We release a multilingual neural machine translation model, which can be used to translate text in the biomedical domain. The model can translate from 5 languages (French, German, Italian, Korean and Spanish) into English. It is trained with large amounts of generic and biomedical data, using domain tags. Our benchmarks show that it performs near state-of-the-art both on news (generic domain) and biomedical test sets, and that it outperforms the existing publicly released models. We believe that this release will help the large-scale multilingual analysis of the digital content of the COVID-19 crisis and of its effects on society, economy, and healthcare policies.

We also release a test set of biomedical text for Korean-English. It consists of 758 sentences from official guidelines and recent papers, all about COVID-19.

1 Motivation

The 2019–2020 coronavirus pandemic has disrupted lives, societies and economies across the globe. Its classification as a *pan*demic highlights its global impact, touching people of all languages. Digital content of all types (social media, news articles, videos) have focused for many weeks predominantly on the sanitary crisis and its effects on infected people, their families, healthcare workers and the society and economy at large. This calls not only for a large set of tools to help during the pandemic (as evidenced by the submissions to this workshop), but also for tools to help digest and analyze this data after it ends. By analyzing the representation and reaction across countries with different guidelines or global trends, it might be possible to inform policies in prevention of and reaction to future epidemics. Several institutions and groups have already started to take snapshots of the digital content shared during these weeks (Croquet, 2020; Banda et al., 2020).

However, because of its global scale, all this digital content is accessible in a variety of different languages, and most existing NLP tools remain English-centric (Anastasopoulos and Neubig, 2019). In this paper we describe the release of a multilingual neural machine translation model (MNMT) that can be used to translate biomedical text. The model is both multi-domain and multilingual, covering translation from French, German, Spanish, Italian and Korean to English.

Our contributions consist in the release of:

- An MNMT model, and benchmark results on standard test sets;

- A new biomedical Korean-English test set.

This paper is structured as follows: in Section 2 we overview previous work upon which we build; Section 3 details the model and data settings, and the released test set; and Section 4 compares our model to other public models and to state-of-the-art results in academic competitions.

The model can be downloaded at `https://github.com/naver/covid19-nmt`: the repository consists in a model checkpoint that is compatible with *Fairseq* (Ott et al., 2019), and a script to preprocess the input text.

2 Related Work

In order to serve its purpose, our model should be able to process multilingual input sentences, and generate tailored translations for COVID-19-related sentences. As far as NMT models are concerned, both multilingual and domain-specific sentences are just sequences of plain tokens that should be distinguished internally and handled in a separate manner depending on the multiple languages or domains. Due to this commonality in both fields

of MNMT and domain adaptation of NMT models, they can be broadly categorized into two groups: 1) data-centric and 2) model-centric (Chu and Wang, 2018).

The former focuses on the preparation of the training data such as handling and selecting from multi-domain (Kobus et al., 2016; Tars and Fishel, 2018) or multilingual parallel corpora (Aharoni et al., 2019; Tan et al., 2019a); and generating synthetic parallel data from monolingual corpora (Sennrich et al., 2015; Edunov et al., 2018).

The model-centric approaches, on the other hand, center on adjusting the training objectives (Wang et al., 2017; Tan et al., 2019b); modifying the model architectures (Vázquez et al., 2018; Dou et al., 2019a); and tweaking the decoding procedure (Hasler et al., 2018; Dou et al., 2019b).

While the two types of approaches are orthogonal and can be utilized in tandem, our released model is trained using data-centric approaches. One of the frequently used data-centric methods for handling sentences of multiple languages and domains is simply prepending a special token that indicates the target language or domain that the sentence should be translated into (Kobus et al., 2016; Aharoni et al., 2019). By feeding the task-specific meta-information via the reserved tags, we signal the model to treat the following input tokens accordingly. Recent works show that this method is also applicable to generating diverse translations (Shu et al., 2019) and translations in specific styles (Madaan et al., 2020).

In addition, back-translation of target monolingual or domain-specific sentences is often conducted in order to augment the low-resource data (Edunov et al., 2018; Hu et al., 2019). Afterward, the back-translated data (and existing parallel data) can be filtered (Xu et al., 2019) or treated with varying amount of importance (Wang et al., 2019) using data selection methods. Back-translated sentences can be tagged to achieve even better results (Caswell et al., 2019).

While myriads of research works on MNMT and domain adaptation exist, the number of publicly available pre-trained NMT models is still low. For example, Fairseq, a popular sequence-to-sequence toolkit maintained by Facebook AI Research, has released ten uni-directional models for translating English, French, German, and Russian sentences.[1] For its widespread usage, we trained our model using this toolkit.

A large number of public MT models are available thanks to OPUS-MT,[2] created by the Helsinki-NLP group. Utilizing the OPUS corpora (Tiedemann, 2012), more than a thousand MT models are trained and released, including several multilingual models which we use to compare with our model.

To the best of our knowledge, we release the first public MNMT model that is capable of producing tailored translations for the biomedical domain.

The COVID-19 pandemic has shown the need for multilingual access to hygiene and safety guidelines and policies (McCulloch, 2020). As an example of crowd-sourced translation, we point out "The COVID Translate Project"[3] which allowed the translation of 75 pages of guidelines for public agents and healthcare workers, from Korean into English in a matter of days. Although our model could assist in furthering such initiatives, we do not recommend relying solely on our model for translating such guidelines, where quality is of uttermost importance. However, the huge amount of digital content created in the last months around the pandemic makes such professional translations of all that content not only infeasible, but sometimes unnecessary depending on the objective. For instance, we believe that the release of this model can unlock the possibility of large-scale translation with the aim of conducting data analysis on the reaction of the media and society on the matter.

3 Model Settings and Training Data

The model uses a variant of the Transformer Big architecture (Vaswani et al., 2017) with a shallower decoder: 16 attention heads, 6 encoder layers, 3 decoder layers, an embedding size of 1024, and a feed-forward dimension of 8192 in the encoder and 4096 in the decoder.

As all language pairs have English as their target language, no special token for target language was used (language detection can be performed internally by the model).

As the model performs many-to-English translation, its encoder should be able to hold most of the complexity. Thus, we increase the capacity of the encoder by doubling the default size of the feed-forward layer as in (Ng et al., 2019).

---

[1] https://github.com/pytorch/fairseq/

blob/master/examples/translation/README.md
[2] https://github.com/Helsinki-NLP/OPUS-MT-train
[3] https://covidtranslate.org/

On the other hand, previous works (Clinchant et al., 2019; Kasai et al., 2020) have shown that it is possible to reduce the number of decoder layers without sacrificing much performance, allowing both faster inference, and smaller network size.

During training, regularization was done with a dropout of 0.1 and label smoothing of 0.1. For optimization, we used Adam (Kingma and Ba, 2014) with warm-up, and maximum learning rate of 0.001. The model was trained for 10 epochs and the best checkpoint was selected based on perplexity on the validation set.

As training data, we used the standard open-accessible datasets, including biomedical data whenever available, for example, the "Corona Crisis Corpora" (TAUS, 2020). Following our past success in domain adaptation (Berard et al., 2019), we used domain tokens (Kobus et al., 2016) to differentiate between domains, allowing multi-domain translation with a single model. We initially experimented with more tags, and combinations of tags (e.g., `medical` → `patent` or `medical` → `political`) to allow for more fine-grained control of the resulting translation. The results however were not very conclusive, and often underperformed. An exception worth noting was the case of transcribed data such as *TED* talks, and *OpenSubtitles*, which are not the main targets of this work. Therefore, for simplicity, we used only two tags: `medical` and `back-translation`. No tag was used with training data that does not belong to one of these two categories.

In addition to biomedical data, we also used back-translated data, although only for Korean, the language with the smallest amount of training data (13.8M sentences). Like Arivazhagan et al. (2019), we used a temperature parameter of 5, to give more chance to Korean. Additionally, the biomedical data was oversampled by a factor of 2. Table 1 details the amount of training sentences used for each language and each domain tag.

As for pre-processing, we cleaned the available data by conducting white-space normalization and NFKC normalization. We filtered noisy sentence pairs based on length (min. 1 token, max. 200), and automatic language identification with `langid.py` (Lui and Baldwin, 2012).

We trained a lower-cased shared BPE model using *SentencePiece* (Kudo and Richardson, 2018) by using 6M random lines for each language (including English). We filtered out single characters

Language	Total	General	BT	Biomed.
French	128.8	125.0	–	3.8
Spanish	92.9	90.8	–	2.1
German	87.3	84.8	–	2.5
Italian	45.6	44.9	–	0.7
Korean	13.8	5.7	8.0	0.1
Total	368.4	351.2	8.0	9.2

Table 1: Amount of training data in millions of sentences. BT refers to back-translation from English to Korean.

with fewer than 20 occurrences from the vocabulary. This results in a shared vocabulary of size 76k.

We reduced the English vocabulary size to speed up training and inference, by setting a BPE frequency threshold of 20, which gives a target vocabulary of size 38k. To get the benefits of a shared vocabulary (i.e., tied source/target embeddings) we sorted the source Fairseq dictionary to put the 38k English tokens at the beginning, which lets us easily share the embedding matrix between the encoder and the decoder.[4]

The BPE segmentation is followed by inline-casing (Berard et al., 2019), where each token is lower-cased and directly followed by a special token specifying its case (`<T>` for title case, `<U>` for all caps, no token for lower-case). Word-pieces whose original case is undefined (e.g., "MacDonalds") are split again into word-pieces with defined case ("mac" and "donalds").

3.1 New Korean-English Test Set

To benchmark the performance on the COVID-19 domain, we built an in-domain test set for Korean-English, as it is the only language pair that is not included in the Corona Crisis Corpora.

The test set contains 758 Korean-English sentence pairs, obtained by having English sentences translated into Korean by four professional Korean translators. We note that any acronym written without its full form in the source sentence is kept the same in the translation unless it is very widely used in general. The English sentences were distributed among the translators with the same guidelines to get consistent tone and manner.

We gathered English sentences from two sources: 1) The official English guidelines and reports from

---

[4]We modified the released checkpoint for it to work out-of-the box with vanilla Fairseq.

Korea Centers for Disease Control and Prevention (KCDC)[5] under Ministry of Health and Welfare of South Korea (258 sentences); and 2) Abstracts of biomedical papers on SARS-CoV-2 and COVID-19 from *arXiv*,[6] *medRxiv*[7] and *bioRxiv*[8] (500 sentences). The sentences were handpicked, focusing on covering diverse aspects of the pandemic, including safety guidelines, government briefings, clinical tests, and biomedical experimentation.

4 Benchmarks

We benchmarked the released multilingual models against: 1) reported numbers in the literature, and 2) other publicly released models. We used OPUS-MT, a large collection (1000+) of pre-trained models released by the NLP group at University of Helsinki. Note that these models were trained with much smaller amounts of training data.

We note that the biomedical test sets (*Medline*) are very small (around 600 lines), and we computed the scores for translations against the untokenized references, which might differ from the official task (from which we reported the numbers). We do not report comparison for Spanish-English *newstest2013*, as the latest reported numbers are outdated (the best WMT entry achieved 30.4).

Our single model obtains competitive results on "generic" test sets (*News* and *IWSLT*), on par with the state of the art. We also obtain strong results on the biomedical test sets. Note the SOTA models were trained to maximize performance in the very specific Medline domain, for which training data is provided. While we included this data in our tagged biomedical data, we did not fine-tune aggressively over it.

Table 2 shows the BLEU scores for the Korean-English COVID-19 test set. The results greatly outperform existing public Korean-English models, even more so than with the IWSLT test sets (Table 3).

5 Conclusion

We describe the release of a multilingual translation model that supports translation in both the general

Model	arXiv	KCDC	Full
Ours	36.5	38.3	37.2
Ours (`medical`)	36.6	38.6	37.4
OPUS-MT	18.7	19.0	18.8

Table 2: Benchmark of the released model on new Korean-English COVID-19 test set.

Language	Model	News	Medline	IWSLT
French	Ours	**41.00**	34.13	**41.09**
	SOTA	40.22*	**35.56**[‡]	–
	OPUS-MT	36.80	32.40	38.90
German	Ours	**41.28**	28.10	31.55
	SOTA	40.98[†]	**28.82**[‡]	**32.01**[†]
	OPUS-MT	39.50	26.50	30.30
Spanish	Ours	**36.63**	**43.94**	**48.79**
	SOTA	–	43.03[‡]	–
	OPUS-MT	30.30	36.30	46.10
Italian	Ours			**42.18**
	SOTA			–
	OPUS-MT			39.70
Korean	Ours			**21.33**
	SOTA			–
	OPUS-MT			17.60

Table 3: Benchmark of the released model against the best reported numbers and the public OPUS-MT models. Our scores on the medical test sets are obtained with the `medical` tag. No tag is used with the other test sets. The SOTA numbers for *News* and *IWSLT* were either obtained by running the corresponding models (Berard et al., 2019; Ng et al., 2019). For *Medline*, we copied the results from the WMT19 Biomedical Task report (Bawden et al., 2019). *News* consists in the *newstest2019* for German (WMT19 News Task test set), *newstest2014* for French, and *newstest2013* for Spanish. For *IWSLT*, the 2017 test set is used for all but Spanish, where the 2016 one is used.

and biomedical domains. Our model is trained on more than 350M sentences, covering French, Spanish, German, Italian and Korean (into English). Benchmarks on public test sets show its strength across domains. In particular, we evaluated the model in the biomedical domain, where it performs near state-of-the-art, but with the advantage of being a single model that operates on several languages. To address the shortage of Korean-English data, we also release a dataset of 758 sentence pairs covering recent biomedical text about COVID-19.

Our aim is to support research studying the international impact that this crisis is causing, at a societal, economical and healthcare level.

---

[5] http://ncov.mohw.go.kr/en
[6] https://arxiv.org
[7] https://www.medrxiv.org
[8] https://www.biorxiv.org
*NLE @ WMT19 Robustness Task (Berard et al., 2019)
[†] FAIR @ WMT19 News Task (Ng et al., 2019)
[‡] Reported results in the WMT19 Biomedical Task (Bawden et al., 2019)

References

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. *arXiv preprint arXiv:1903.00089*.

Antonios Anastasopoulos and Graham Neubig. 2019. Should all cross-lingual embeddings speak english? *arXiv preprint arXiv:1911.03058*.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, et al. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges. *arXiv preprint arXiv:1907.05019*.

Juan M Banda, Ramya Tekumalla, Guanyu Wang, Jingyuan Yu, Tuo Liu, Yuning Ding, and Gerardo Chowell. 2020. A large-scale covid-19 twitter chatter dataset for open scientific research–an international collaboration. *arXiv preprint arXiv:2004.03688*.

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 29–53, Florence, Italy. Association for Computational Linguistics.

Alexandre Berard, Ioan Calapodescu, and Claude Roux. 2019. Naver labs Europe's systems for the WMT19 machine translation robustness task. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 526–532, Florence, Italy. Association for Computational Linguistics.

Isaac Caswell, Ciprian Chelba, and David Grangier. 2019. Tagged back-translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 53–63, Florence, Italy. Association for Computational Linguistics.

Chenhui Chu and Rui Wang. 2018. A survey of domain adaptation for neural machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1304–1319, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Stephane Clinchant, Kweon Woo Jung, and Vassilina Nikoulina. 2019. On the use of BERT for neural machine translation. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 108–117, Hong Kong. Association for Computational Linguistics.

Pauline Croquet. 2020. Comment les archivistes de la BNF sauvegardent la mémoire du confinement sur internet. Accessed June 2020.

Zi-Yi Dou, Junjie Hu, Antonios Anastasopoulos, and Graham Neubig. 2019a. Unsupervised domain adaptation for neural machine translation with domain-aware feature embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1417–1422.

Zi-Yi Dou, Xinyi Wang, Junjie Hu, and Graham Neubig. 2019b. Domain differential adaptation for neural machine translation. *arXiv preprint arXiv:1910.02555*.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.

Eva Hasler, Adrià de Gispert, Gonzalo Iglesias, and Bill Byrne. 2018. Neural machine translation decoding with terminology constraints. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 506–512, New Orleans, Louisiana. Association for Computational Linguistics.

Junjie Hu, Mengzhou Xia, Graham Neubig, and Jaime Carbonell. 2019. Domain adaptation of neural machine translation by lexicon induction. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Jungo Kasai, Nikolaos Pappas, Hao Peng, James Cross, and Noah A. Smith. 2020. Deep encoder, shallow decoder: Reevaluating the speed-quality tradeoff in machine translation.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Catherine Kobus, Josep Crego, and Jean Senellart. 2016. Domain control for neural machine translation. *arXiv preprint arXiv:1612.06140*.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*.

Marco Lui and Timothy Baldwin. 2012. langid.py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 System Demonstrations*, pages 25–30, Jeju Island, Korea. Association for Computational Linguistics.

Aman Madaan, Amrith Setlur, Tanmay Parekh, Barnabas Poczos, Graham Neubig, Yiming Yang, Ruslan Salakhutdinov, Alan W Black, and Shrimai Prabhumoye. 2020. Politeness transfer: A tag and generate approach.

Gretchen McCulloch. 2020. Covid-19 is history's biggest translation challenge.

Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. 2019. Facebook FAIR's WMT19 news translation task submission. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 314–319, Florence, Italy. Association for Computational Linguistics.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. *arXiv preprint arXiv:1904.01038*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Improving neural machine translation models with monolingual data. *CoRR*, abs/1511.06709.

Raphael Shu, Hideki Nakayama, and Kyunghyun Cho. 2019. Generating diverse translations with sentence codes. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1823–1827, Florence, Italy. Association for Computational Linguistics.

Xu Tan, Jiale Chen, Di He, Yingce Xia, Tao Qin, and Tie-Yan Liu. 2019a. Multilingual neural machine translation with language clustering. *arXiv preprint arXiv:1908.09324*.

Xu Tan, Yi Ren, Di He, Tao Qin, and Tie-Yan Liu. 2019b. Multilingual neural machine translation with knowledge distillation. In *International Conference on Learning Representations*.

Sander Tars and Mark Fishel. 2018. Multi-domain neural machine translation. *arXiv preprint arXiv:1805.02282*.

TAUS. 2020. Taus corona crisis corpus.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

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.

Raúl Vázquez, Alessandro Raganato, Jörg Tiedemann, and Mathias Creutz. 2018. Multilingual nmt with a language-independent attention bridge. *arXiv preprint arXiv:1811.00498*.

Rui Wang, Masao Utiyama, Lemao Liu, Kehai Chen, and Eiichiro Sumita. 2017. Instance weighting for neural machine translation domain adaptation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1482–1488, Copenhagen, Denmark. Association for Computational Linguistics.

Shuo Wang, Yang Liu, Chao Wang, Huanbo Luan, and Maosong Sun. 2019. Improving back-translation with uncertainty-based confidence estimation. *arXiv preprint arXiv:1909.00157*.

Guanghao Xu, Youngjoong Ko, and Jungyun Seo. 2019. Improving neural machine translation by filtering synthetic parallel data. *Entropy*, 21(12):1213.