# Machine Translation of Covid-19 Information Resources via Multilingual Transfer

Ivana Kvapilíková and Ondřej Bojar

Charles University, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Malostranské náměstí 25, 118 00 Prague, Czech Republic
<surname>@ufal.mff.cuni.cz

*Abstract:* The Covid-19 pandemic has created a global demand for accurate and up-to-date information which often originates in English and needs to be translated. To train a machine translation system for such a narrow topic, we leverage in-domain training data in other languages both from related and unrelated language families. We experiment with different transfer learning schedules and observe that transferring via more than one auxiliary language brings the most improvement. We compare the performance with joint multilingual training and report superior results of the transfer learning approach.

## 1 Introduction

A global crisis such as the current Covid-19 pandemic requires information to be spread as efficiently as possible. Working with data from different international resources in multiple languages can resolve possible inconsistencies and prevent misinformation. In an emergency situation, new data is released constantly and is communicated to the public not only via national news and authorities, but also foreign media, scientific journals or statements of international agencies. There are extensive data resources written in English which are not accessible for non-English speakers.

In order to quickly access the information in a foreign language, machine translation (MT) can be of great help. However, Covid-related texts use a specific terminology and MT models are known to struggle outside of the general domain.

More than a year after the Covid outbreak, there already is a significant amount of domain-specific multilingual text resources. Furthermore, Covid-related texts are a part of a broader medical domain which can provide additional authentic data for training. Thanks to the MLIA @ Eval[1] initiative who gathered training data for MT and information retrieval related to the pandemic, we can successfully adapt an MT system to the Covid domain or even train it from scratch.

This paper gives an overview of possible methods to automatically translate Covid-related texts. Section 2 outlines different approaches to train a domain-specific MT system using multilingual corpora. Section 3 describes our training data and section 4 gives more details about our MT systems and presents the results. Section 5 concludes the paper.



[1] http://eval.covid19-mlia.eu/

## 2 Methodology

In this section we outline several strategies applicable in the situation where we need to translate from English to multiple languages, we are confined within a specific domain and we have mid-size parallel corpora for every language pair of interest.

Firstly, we can train a standard MT system from scratch for each language separately, possibly resorting to some data augmentation method, e.g. back-translation. Secondly, we can use transfer learning to transfer from a pretrained MT system in one language to another. Finally, we can train a multilingual MT system which learns jointly from all available data.

In our experiments, we do not consider any additional monolingual resources. Although monolingual data are generally easier to obtain, we remain constrained by the datasets provided by the organizers of MLIA @ Eval which are described in Section 3. We also do not evaluate a transfer from a large MT model pretrained on texts from the general domain which would be a promising strategy as well.

### 2.1 Low-resource Neural Machine Translation

When neural machine translation (NMT) became the dominant paradigm in MT [19], it was believed that extremely large parallel resources are required for training. However, Sennrich and Zhang [18] showed that with careful tuning of the hyperparameters, an NMT model can be successfully trained already on 100k sentence pairs, which is less than we have available in the Covid/medical domain for the language pairs of our interest. Furthermore, Conneau and Lample [4] show that translation quality can be further boosted by pretraining a language model and using it to initialize the parameters of both the encoder and the decoder.

An NMT system directly trained to translate in the Covid domain serves as our baseline.

| | de | el | es | fr | it | sv |
|---|---|---|---|---|---|---|
| Train | 925,647 | 834,240 | 1,028,287 | 1,004,215 | 900,472 | 806,425 |
| Dev | 528 | 3,378 | 1,973 | 728 | 3,245 | 723 |
| Dev Test | 500 | 500 | 500 | 500 | 500 | 500 |
| Blind Test | 2,000 | 2,000 | 2,000 | 2,000 | 2,000 | 2,000 |

Table 1: Data summary

### 2.2 Data Augmentation

Back-translation is a crucial method in NMT used to augment training data by translating an existing monolingual corpus [16]. The synthetic text can be either on the source [16] or the target [21] side of the training corpus, or both [15].

When using a bidirectional model (sharing the encoder and decoder for both translation directions), back-translation can be performed *on-the-fly*. During training, the model switches between the training and the inference mode to produce batches of synthetic sentence pairs and learn from both authentic and synthetic samples in each training step. As the system improves, the quality of the generated samples improves as well. This approach was originally proposed for training an unsupervised MT system [2, 14].

In our systems, we translate only the target sentences and generate a synthetic source side *on the fly*. We do not use any additional monolingual data for back-translation.

### 2.3 Transfer Learning

The first strategy to utilize multilingual in-domain training corpora is transfer learning. It can be used to transfer from a different domain [6] or a different language [12, 22]. In this work we focus on the latter.

A trivial transfer learning approach was proposed by Kocmi and Bojar [12] who fine-tune a low-resource child model from a high-resource parent model pretrained for a different language pair. The training procedure consists of first training an NMT model on the parent parallel corpus until it converges and then replacing the training data with the child corpus.

Before training the parent model, it is necessary to designate some vocabulary entries for the new language. Otherwise the model would be forced to completely re-learn its subword embeddings and their connections and would lose its ability to transfer. Kocmi [11] shows that the best strategy is to generate the vocabulary in advance from the concatenation of corpora of both the child and the parent language pair. However, if the child language is not known prior to training the parent, it is enough to leave some "free" slots in the vocabulary and later fill them in with the vocabulary of the child language.

In this work, we experiment with several transfer learning schedules. We repeat the transfer procedure several times with the child becoming the parent for either a completely new language (e.g. German $\rightarrow$ English $\rightarrow$ Spanish $\rightarrow \dots$) or for the original parent (e.g. German $\rightarrow$ English $\rightarrow$ German $\rightarrow \dots$), as illustrated in Figure 1. We always generate the vocabulary from the concatenation of the parent and its "first child". When adding a third (or fourth) language, the joint BPE vocabulary has to be modified by replacing the original parent vocabulary entries with the new child ones. The schedules and their results are described in Section 4.

### 2.4 Multilingual Training

The second strategy to utilize multilingual in-domain training corpora is joint multililngual training.

Multilingual translation systems are either trained with full parameter sharing [1, 7, 9], with language-specific encoders and decoders relying on shared attention [5] or an attention bridge [20]. The results show that multilingual models yield comparable or even superior results to the standard bilingual setup.

In this work, we rely on full parameter sharing and use the same architecture as our bilingual systems, while training it to translate from English into three languages (French, Italian and Spanish) at once. During inference, the target language is determined from indicated language embeddings of the target sentence. We selected these three languages for their similarity which could help the model re-use and share some knowledge. The BPE vocabulary was extracted from the concatenation of all four corpora, using only unique English sentences to reach a comparable corpus size.

## 3 Data

Covid-19 MLIA @ Eval organized a community evaluation effort aimed at accelerating the creation of resources and tools for improved Multilingual Information Access (MLIA). A part of this initiative is a competition to develop the best MT system translating from English to several European languages: German, Modern Greek, French, Italian, Spanish and Swedish. The competition is incremental and so far only the first round was concluded.

The parallel training data provided by the organizers for the first round and used in this paper is summarized in Table 1.[2] It was created based on existing corpora from the medical domain, enriched with sentences directly about

[2]The development test set used for the final model selection was obtained by cutting 500 sentences off of either the train set or the development set, depending on the original development set size.

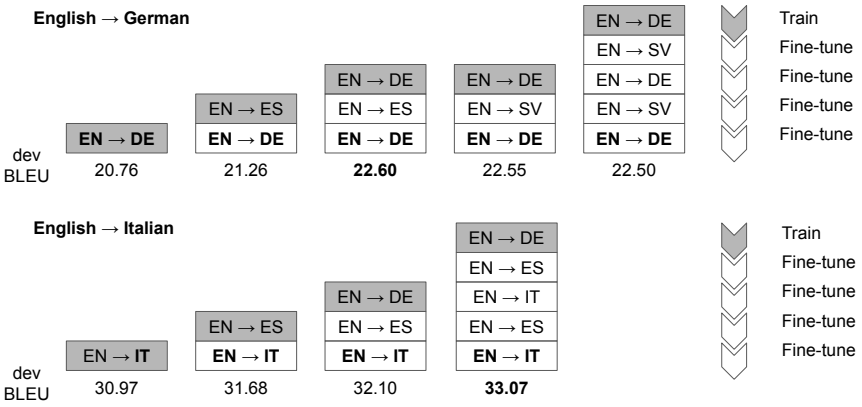


Figure 1: Illustration of the incremental transfer for selected languages: BLEU scores on dev set.

| | de | el | es | fr | it | sv |
|---|---|---|---|---|---|---|
| bidirectional with BT | **21.52** | 22.30 | **40.94** | **38.46** | **33.17** | **20.61** |
| unidirectional without BT | 20.76 | **22.70** | 40.46 | 35.57 | 30.97 | 19.13 |

Table 2: Translating from English using the baseline model and back-translation: BLEU scores on dev set.

Covid, mostly harvested through web crawling and parallel sentence mining [3]. The sentences in different languages might be similar, but the entire corpus collection is not multi-parallel.

All data was segmented into BPE units [17] with a vocabulary of 30k items for the training.

## 4 Experiments & Results

We participated in the MT shared task of the Covid-19 MLIA @ Eval initiative and trained a model for translation into each of the six languages listed in Section 3. The results of our submitted systems are summarized in the preliminary report [13], the overall results are discussed in the shared task findings [3]. Our English → German and English → Swedish systems ranked first (tied with one other system), our other models ranked second.

We experimented with three training strategies compared against one baseline *BASE*:

1. unidirectional training without back-translation (*BASE*);
2. bidirectional training with online back-translation (*BT*);
3. transfer learning (*TRANSFER*);
4. multilingual training (*MULTILING*).

For all our MT models we use a 6-layer Transformer [19] architecture with 8 heads, embedding dimension of 1024 and GELU [8] activations. The training is performed using the XLM[3] toolkit. The translation models were trained on 4 GPUs[4] with 2-step gradient accumulation to reach an effective batch size of $8 \times 3400$ tokens. Effective batch size has a significant impact on the training and we observe that the models converge on lower BLEU scores for smaller batch sizes. We used Adam [10] optimizer with inverse square root decay ($\beta_1 = 0.9$, $\beta_2 = 0.98$, $lr = 0.0001$). Beam search with the beam size of 4 was used during final decoding; greedy decoding was used for back-translation. The vocabulary size was set to 30k. Using larger vocabulary leads to a performance drop. All our model parameters are initialized with a pretrained masked language model as described in Conneau and Lample [4].

### 4.1 Online Back-Translation

For each language pair we first trained a bidirectional back-translation model described in Section 2 and compared it to a standard unidirectional model without back-translation. Online back-translation improved the score by 0.5–2.9 BLEU points, depending on the language, but surprisingly caused a decrease of 0.4 BLEU in the case of the English–Modern Greek model. The reason for this drop is likely in the bidirectionality of the model rather than the data augmentation itself. The results are summarized in Table 2.

We experimented with a dropout of 0.1 and 0.2 and concluded that higher dropout helps in most settings. This observation is in line with Sennrich and Zhang [18] who emphasize the role of higher dropout when working with low- to medium- sized resources.

[3] https://github.com/facebookresearch/XLM

[4] Quadro P5000, 16GB of RAM

| Transfer Combination | de | el | es | fr | it | sv |
|---|---|---|---|---|---|---|
| en-es → en-de | 21.26 | | | | | |
| en-de → en-es | | | 41.28 | | | |
| en-de → en-es → en-de | **22.60** | | | | | |
| en-de → en-es → en-fr | | | | **35.10** | | |
| en-de → en-es → en-it | | | | | 32.10 | |
| en-de → en-es → en-it → en-es | | | **41.34** | | | |
| en-de → en-es → en-it → en-es → en-it | | | | | **33.07** | |
| en-es → en-fr | | | | 32.43 | | |
| en-es → en-it | | | | | 31.68 | |
| en-de → en-el | | **23.29** | | | | |
| en-es → en-el | | 20.91 | | | | |
| en-de → en-sv | | | | | | **21.69** |
| en-de → en-sv → en-de | 22.55 | | | | | |
| en-de → en-sv → en-de → en-sv | | | | | | 20.56 |
| en-de → en-sv → en-de → en-sv → en-de | 22.50 | | | | | |

Table 3: Translating from English using the *TRANSFER* models: BLEU scores on dev set.

| | de | el | es | fr | it | sv |
|---|---|---|---|---|---|---|
| multilingual | - | - | 40.2 | 36.1 | 32.8 | - |
| best transfer | **22.6** | **23.3** | **41.3** | 35.1 | 33.1 | **21.7** |
| best base | 21.5 | 22.7 | 40.9 | **38.5** | **33.2** | 20.6 |

Table 4: Translating from English using the best models from each category: BLEU scores on dev set.

| | de | el | es | fr | it | sv |
|---|---|---|---|---|---|---|
| multilingual | - | - | 47.3 | 48.0 | **28.3** | - |
| best transfer | **31.6** | **24.7** | **47.9** | 47.1 | **28.3** | **30.1** |
| best base | 31.4 | 24.1 | 47.3 | **48.4** | - | 28.5 |

Table 5: Translating from English using the best models from each category: BLEU scores on blind test set.

### 4.2 Transfer Learning

We used the best-performing *BASE* / *BT* models as the parent models and continued with unidirectional training (English → foreign language) for our transfer learning experiments. Since the fine-tuning is unidirectional, we can no longer perform online back-translation.

We observed that it often helped to use the transfer incrementally, having the model converge on one parallel corpus, switch the target language, wait for convergence and switch again. We hypothesize that the model benefits from seeing a larger variety of sentences. For example transferring from German to Spanish to Italian (32.10 BLEU) performs better than transferring directly from Spanish to Italian (31.68 BLEU). The best combination is to even repeat the Spanish-Italian transfer twice (33.07 BLEU).

When translating from English to German, fine-tuning the en-de *BT* model on English→Spanish (or English→Swedish) and switching back to English → German adds around 1 BLEU on top of the original *BT* model. All language combinations used in our transfer learning experiments are described in Table 3 and selected schedules are illustrated in Figure 1.

We observe that transfer learning improves the performance in all cases but French, where the *BASE* model with BT reaches 38.5 BLEU, which is $\sim$ 3 BLEU points more than transfer learning. There is a significant overlap between the training sets in different languages and it is possible that French does not benefit from the transfer because it does not provide enough new sentences. On the other hand, the largest improvement is seen by the language pair with the least amount of training data, English-Swedish, where BLEU increases by 1.1 points on the dev set and 1.6 points on the test set.

### 4.3 Multilingual Training

We train a multilingual model for translation from English to French, Italian and Spanish. The model has the same architecture as our bilingual models, all parameters are shared for all languages. Its encoder and decoder were first pretrained on monolingual data in all three languages and English using the MLM criterion [4].

Table 4 shows the comparison of the *TRANSFER* models with a multilingual model trained jointly. We observe that transfer learning yields superior results and is thus a more effective way to leverage multilingual data than joint multilingual training. However, there is an advantage of a joint model in terms of the training and storage cost. After three days of training, the multilingual model can be used for translation into all three languages. The initial *BASE* models can take between one (without BT) and five (with BT) days to train and fine-tuning on a child language pair adds around 6 hours.

Table 5 lists our task submissions and compares all approaches on the official Covid-19 MLIA @ Eval blind test set.[5]

[5]The BLEU scores in Table 4 and Table 5 cannot be directly com-

## 5 Conclusion

We trained several MT systems specialized in translation of texts related to the topic of Covid-19 and the pandemic from English to six European languages.

We experimented with three training approaches and we conclude that there is not a universal winner that would work the best for all language pairs. However, transfer learning brings promising results across the board, especially when training data is limited. We observed an interesting phenomenon where incremental fine-tuning on multiple languages brings additional gains, as we expose the model to a larger variety of training sentences.

In our setting, transferring knowledge is a more efficient way to leverage multilingual data than joint training. For English→German, we observe that a transfer learning detour via Spanish or Swedish improves the parent model itself. For English→Modern Greek, transfer learning via German works well, despite the unrelatedness of the two languages. For English→French, on the other hand, a bidirectional model with back-translation beats both multilingual and transfer-based models.

## Acknowledgments

This study was supported in parts by the grants CZ.07.1.02/0.0/0.0/16_023/0000108 (Operational Programme – Growth Pole of the Czech Republic), 19-26934X of the Czech Science Foundation, and by the SVV project number 260 575.

## References

[1] Aharoni, R., Johnson, M., Firat, O.: Massively multilingual neural machine translation. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 3874–3884, Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019)

[2] Artetxe, M., Labaka, G., Agirre, E., Cho, K.: Unsupervised neural machine translation. In: Proceedings of the Sixth International Conference on Learning Representations (April 2018)

[3] Casacuberta, F., Ceausu, A., Choukri, K., Deligiannis, M., Domingo, M., Garcia-Martinez, M., Herranz, M., Papavassiliou, V., Piperidis, S., Prokopidis, P., Roussis., D.: The Covid-19 MLIA @ Eval Initiative: Overview of the machine translation task (2021), URL `http://eval.covid19-mlia.eu/meetings/round1/report/20210112-task3-overview.pdf`

pared as the dev scores were calculated by authors and test scores by the organizers.

[4] Conneau, A., Lample, G.: Cross-lingual language model pretraining. In: Wallach, H., Larochelle, H., Beygelzimer, A., d' Alché-Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems 32, pp. 7059–7069, Curran Associates, Inc. (2019)

[5] Firat, O., Cho, K., Bengio, Y.: Multi-way, multilingual neural machine translation with a shared attention mechanism. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 866–875, Association for Computational Linguistics, San Diego, California (Jun 2016)

[6] Freitag, M., Al-Onaizan, Y.: Fast domain adaptation for neural machine translation. CoRR **abs/1612.06897** (2016)

[7] Ha, T.L., Niehues, J., Waibel, A.: Toward multilingual neural machine translation with universal encoder and decoder (2016)

[8] Hendrycks, D., Gimpel, K.: Bridging nonlinearities and stochastic regularizers with gaussian error linear units. CoRR **abs/1606.08415** (2017)

[9] Johnson, M., Schuster, M., Le, Q.V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., Hughes, M., Dean, J.: Google's multilingual neural machine translation system: Enabling zero-shot translation. Transactions of the Association for Computational Linguistic **5**, 339–351 (2017)

[10] Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: Proceedings of the 3rd International Conference for Learning Representations (2015)

[11] Kocmi, T.: Exploring Benefits of Transfer Learning in Neural Machine Translation. Ph.D. thesis, Charles University (2020)

[12] Kocmi, T., Bojar, O.: Trivial transfer learning for low-resource neural machine translation. In: Proceedings of the Third Conference on Machine Translation: Research Papers, pp. 244–252, Association for Computational Linguistics, Brussels (Oct 2018)

[13] Kvapilíková, I.: CUNI machine translation systems for the Covid-19 MLIA initiative (2021), URL `http://eval.covid19-mlia.eu/meetings/round1/report/20210114-cunimt.pdf`

[14] Lample, G., Denoyer, L., Ranzato, M.: Unsupervised machine translation using monolingual corpora only. In: 6th International Conference on Learning Representations (ICLR 2018) (2018)

[15] Niu, X., Denkowski, M., Carpuat, M.: Bi-directional neural machine translation with synthetic parallel data. In: Proceedings of the 2nd Workshop on Neural Machine Translation and Generation, pp. 84–91, Association for Computational Linguistics, Melbourne, Australia (Jul 2018)

[16] Sennrich, R., Haddow, B., Birch, A.: Improving neural machine translation models with monolingual data. In: Proceedings of the 54th Annual Meeting of the ACL (Volume 1: Long Papers), pp. 86–96, Association for Computational Linguistics, Berlin, Germany (Aug 2016)

[17] Sennrich, R., Haddow, B., Birch, A.: Neural machine translation of rare words with subword units. In: Proceedings of the 54th Annual Meeting of the ACL, pp. 1715–1725, Association for Computational Linguistics, Berlin (Aug 2016)

[18] Sennrich, R., Zhang, B.: Revisiting low-resource neural machine translation: A case study. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 211–221, Association for Computational Linguistics, Florence, Italy (Jul 2019)

[19] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems 30, pp. 6000–6010, Curran Associates, Inc. (2017)

[20] Vázquez, R., Raganato, A., Tiedemann, J., Creutz, M.: Multilingual NMT with a language-independent attention bridge. In: Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019), pp. 33–39, Association for Computational Linguistics, Florence, Italy (Aug 2019)

[21] Wu, L., Wang, Y., Xia, Y., Qin, T., Lai, J., Liu, T.Y.: Exploiting monolingual data at scale for neural machine translation. 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), pp. 4207–4216, Association for Computational Linguistics, Hong Kong, China (Nov 2019)

[22] Zoph, B., Yuret, D., May, J., Knight, K.: Transfer learning for low-resource neural machine translation. In: Proceedings of the 2016 Conference on EMNLP, pp. 1568–1575, Association for Computational Linguistics, Austin, Texas (Nov 2016)