CUNI Machine Translation Systems for the Covid-19 MLIA Initiative

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. We participated in the MT task of the Covid-19 MLIA initiative and submitted MT systems translating from English to Arabic, German, Greek, French, Italian, Italian and Swedish. After the first round, we observe that an efficient training strategy is to use transfer learning to leverage in-domain training data in other languages (both from related and unrelated language families).

1 Introduction

A global crisis such as the current Covid-19 pandemic requires information to be spread as efficiently as possible. Working with information from different international resources in multiple languages can resolve possible inconsistencies and prevent misinformation. In an emergency situation, new data are released constantly and are communicated to the general public via national news or government statements, as well as international reports and scientific journals. 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 are a part of a specific domain and MT models are known to struggle outside of the general domain.

The machine translation task of the MLIA @ Eval initiative consists of translating from English to German, Greek, French, Italian, Italian, Swedish and Arabic (added for Round 2). Trained MT systems can be used to make Covid-related texts accessible to speakers of these six languages. Our team participates in all language tracks.

2 Data

In both rounds we only submitted constrained systems trained on the data provided by the task organizers. The data for Rounds 1 and 2 are summarized in

Covid-19 MLIA @ Eval Initiative, http://eval.covid19-mlia.eu/.

Table 1 and Table 2. The development test set used for final model selection was obtained by cutting 500 sentences of either the train set or the development set, depending on the original development set size.

	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 for Round 1

	ar	**de**	**el**	**es**	**fr**	**it**	**sv**
Train	424,434	1,536,411	673,961	2,862.002	2,412,653	1,026,064	374,998
Dev	4,000	4,000	4,000	4,000	4,000	4,000	4,000
Dev Test	–	528	3,878	2,473	728	3,745	723
Blind Test	4,000	4,000	4,000	4,000	4,000	4,000	4,000

Table 2: Data Summary for Round 2. The dev data from Round 1 were used as dev test for Round 2.

3 Methodology

3.1 Round 1

In Round 1 we experimented with three training approaches:

1. standard NMT training with back-translation (*BASE*);
2. transfer learning (*TRANSFER*);
3. multilingual training (*MULTILING*).

The first approach relies on one bidirectional model (sharing the encoder and decoder for both translation directions) which constantly switches between the training and the inference mode to produce batches of synthetic sentence pairs and learn from both authentic and synthetic training samples using online back-translation (BT) [5]. The models are trained on BPE units [6] with a vocabulary of 30k items.

The second transfer learning approach was proposed by Kocmi and Bojar [4] who fine-tune a low-resource child model from a pre-trained high-resource

parent model for a different language pair. The method requires a shared subword vocabulary generated from the concatenation of corpora of both the child and the parent language pair. The training procedure consists of first training an NMT model on the parent parallel corpus until it converges, then replace the training data with the child corpus. We experiment with repeating this procedure several times with the child becoming the parent for either a completely new language (e.g. German $\rightarrow$ English $\rightarrow$ Spanish $\rightarrow \ldots$)) or for the original parent (e.g. German $\rightarrow$ English $\rightarrow$ German $\rightarrow \ldots$). When adding a new language, the joint BPE vocabulary has to be modified by replacing the original parent vocabulary entries with the new child's.

The multilingual approach uses the same architecture as described above and trains one MT model to translate from English into three languages (French, Italian and Spanish). 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 reuse 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.

For all our MT models we use a 6-layer Transformer [8] architecture with 8 heads, embedding dimension of 1024 and GELU [1] activations. The training is performed using the XLM[1] toolkit. The translation models were trained on 4 GPU[2] 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 [3] 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.

3.2 Round 2

In Round 2 we continued experimenting with transfer learning and multilingual training. Due to computational constraints we were facing in Round 2, we changed the methodology and trained unidirectional models instead of bidirectional ones. In Round 1 we showed that training a bidirectional model with online back-translation can add up to $\sim$ 3 BLEU points (highest improvement was observed with French and Italian) but it also requires approx. twice the time to converge as compared to a unidirectional model without back-translation.

We again used a 6-layer Transformer [8] architecture with 8 heads, embedding dimension of 1024 and GELU [1] activations but this time we used the Marian NMT [2] toolkit to trained all bilingual models. The models were trained on 1 GPU[3] with adaptable batch size (–mini-batch-fit). We used Adam [3] optimizer

[1] `https://github.com/facebookresearch/XLM`
[2] Quadro P5000, 16GB of RAM
[3] Quadro P5000, 16GB of RAM

with inverse square root decay ($\beta_1 = 0.9$, $\beta_2 = 0.98$, $lr = 0.0001$). Following the conclusions from Round 1, the dropout was set to 0.2. Beam search with the beam size of 6 was used during decoding.

The multilingual model was trained as described in Section 3.1 but in Round 2 we trained jointly on all languages.

All models for Round 2 were trained on the same vocabulary of 50k BPE units which was generated jointly from all 8 monolingual corpora using fastBPE[4].

4 Results

4.1 Round 1

For each language pair we trained a bidirectional back-translation model described in Section 3.1 and compared it to a standard unidirectional model without back-translation. 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 [7] who emphasize the role of higher dropout when working with low- to medium- sized resources. The results are summarized in Table 3. We submitted the best model for each language pair under the name *BASE*.

	de	**el**	**es**	**fr**	**it**	**sv**
bidirectional w\BT	**21.52**	22.30	**40.94**	**38.46**	**33.17**	**20.61**
unidirectional w\o BT	20.76	**22.70**	40.46	35.57	30.97	19.13

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

We used the best-performing *BASE* models as the parent models and continued with unidirectional training (foreign language → English) for our transfer learning experiments. To our surprise, it often helped to use the transfer several times, having the model converge on one parallel corpus, switch the target language, wait for convergence and switch again. 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 *BASE* model on English→Spanish (or English→Swedish) and switching back to English → German adds around 1 BLEU on top of the original *BASE* model. The language combinations used in out transfer learning experiments are described in Table 4.

[4] `https://github.com/glample/fastBPE`

We observe that transfer learning improves the performance in all cases but French, where the *BASE* model with BT reaches 38.46 BLEU, which is $\sim$ 3 BLEU points more than transfer learning.

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

Table 5 shows the comparison of the *TRANSFER* models with a multilingual model trained jointly for French, Italian and Spanish. We observe that transfer learning is 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 6 lists our task submissions and compares all approaches on the official blind test test.

4.2 Round 2

In Round 2 we again evaluated the difference between joint training and transfer learning and the results correspond to our conclusions from Round 1 – our multilingual model performs worse than transfer learning. However, this time even the *BASE* model often performs better than the multilingual one. Since the winners of this round trained a multilingual model which beat all our models

	de	el	es	fr	it	sv
multilingual	-	-	40.15	36.07	32.76	-
best transfer	**22.60**	**23.29**	**41.34**	35.10	33.07	**21.69**
best base	21.52	22.7	40.94	**38.46**	**33.17**	20.61

Table 5: Translating from English using the best *BASE*, *TRANSFER* and *MULTILING* models: 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 6: Translating from English using the best *BASE*, *TRANSFER* and *MULTILING* models: BLEU scores on blind test set.

by a significant margin, we will have to reassess our multilingual training strategy and model architecture for the next round.

We further investigated the effect of language combinations in the transfer learning schedule and we clearly see that the most significant factor is the training corpus size, see Table 9. Using the English-Spanish corpus (2.9M sentences) or English-French (2.4 sentences) as the parent brings the highest improvement of translation quality. Our results correspond to the results of [4]. For each language, we report the model resulting from the best transfer combination as evaluated on the `dev test` set.

The best performance is achieved by ensembling of different transfer learning models. Unfortunately, we did not manage to submit the ensembled models in time for the competition, but our internal results are summarized in Table 7.

	ar	de	el	es	fr	it	sv
multilingual	15.68	24.47	30.60	40.33	42.26	33.02	13.13
best transfer	18.30	29.05	33.45	46.61	36.69	12.55	
base	15.12	28.75	30.92	39.32	40.58	31.77	12.48
transfer ensemble	20.35	31.08	35.54			38.92	14.52

Table 7: Translating from English using the best *BASE*, *TRANSFER* and *MULTILING* models: BLEU scores on blind test set.

	ar	de	el	es	fr	it	sv
multilingual	21.95	34.78	39.33	51.41	53.81	45.10	16.05
best transfer	**24.82**	**40.21**	**43.86**	**56.58**	**57.87**	**47.38**	**18.75**
base	20.67	38.33	40.80	52.17	52.41	45.15	17.74

Table 8: Translating from English using the best *BASE*, *TRANSFER* and *MULTILING* models: BLEU scores on dev set.

		ar	de	el	es	fr	it	sv
	ar		23.67	41.18	56.09	57.12	45.06	16.90
	de	38.78		43.09	56.31	57.56	46.15	18.74
	el	21.51	39.12	41.24	56.11	57.58	45.8	18.08
transfer from	**es**	23.27	39.89		**56.5**8	**57.87**	**47.38**	**18.75**
	fr	**24.82**	**40.21**	43.58			47.35	18.71
	it	24.49	39.18	**43.86**	56.1	57.45		18.38
	sv	23.87	38.38	43.11	55.77	57.38	45.28	

Table 9: Transfer learning for various language pairs (child - horizontal axis, parent - vertical axis). BLEU scores on dev set.

5 Conclusion

We experimented with three training approaches and conclude that there is not a universal winner that would defeat the other models in all language directions. However, transfer learning brings promising results across the board. In this setting, transferring knowledge is a more effective way to leverage multilingual data than joint training.

In Round 1, we observed that a transfer learning detour via several languages improves the parent model. In Round 2, we experimented with different language combinations for transfer learning. We conclude that when pretraining a model on a different language pair, the most important factor is the corpus size. The transfer works also for completely unrelated languages.

In the following Round we would like to experiment with different architectures and training parameters for the multilingual model. We would also like to train an unconstrained model using a large pretrained model from the general domain.

Acknowledgments

This study was supported in parts by the grants 19-26934X of the Czech Science Foundation, START/SCI/089 and SVV 260 575 of the Charles University. This work has been using language resources and tools stored and distributed by the

LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (LM2018101).

References

[1] Hendrycks, D., Gimpel, K.: Bridging nonlinearities and stochastic regularizers with gaussian error linear units. arXiv [e-Print archive] **abs/1606.08415** (2017), URL https://arxiv.org/abs/1606.08415

[2] Junczys-Dowmunt, M., Grundkiewicz, R., Dwojak, T., Hoang, H., Heafield, K., Neckermann, T., Seide, F., Germann, U., Fikri Aji, A., Bogoychev, N., Martins, A.F.T., Birch, A.: Marian: Fast neural machine translation in C++. In: Proceedings of ACL 2018, System Demonstrations, pp. 116–121, Association for Computational Linguistics, Melbourne, Australia (July 2018), URL http://www.aclweb.org/anthology/P18-4020

[3] Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: Proceedings of the 3rd International Conference for Learning Representations (2015), URL http://arxiv.org/abs/1412.6980

[4] 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), https://doi.org/10.18653/v1/W18-6325, URL https://www.aclweb.org/anthology/W18-6325

[5] Lample, G., Denoyer, L., Ranzato, M.: Unsupervised machine translation using monolingual corpora only. In: 6th International Conference on Learning Representations (ICLR 2018) (2018), URL http://arxiv.org/abs/1711.00043

[6] 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), https://doi.org/10.18653/v1/P16-1162, URL https://www.aclweb.org/anthology/P16-1162

[7] 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), https://doi.org/10.18653/v1/P19-1021, URL https://www.aclweb.org/anthology/P19-1021

[8] 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), URL http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf