The LMU Munich System for the WMT 2021 Large-Scale Multilingual Machine Translation Shared Task

Wen Lai and **Jindřich Libovický** and **Alexander Fraser**
Center for Information and Language Processing, LMU Munich, Germany
{lavine, libovicky, fraser}@cis.lmu.de

Abstract

This paper describes the submission of LMU Munich to the WMT 2021 multilingual machine translation task for small track #1, which studies translation between 6 languages (Croatian, Hungarian, Estonian, Serbian, Macedonian, English) in 30 directions. We investigate the extent to which bilingual translation systems can influence multilingual translation systems. More specifically, we trained 30 bilingual translation systems, covering all language pairs, and used data augmentation techniques such as back-translation and knowledge distillation to improve the multilingual translation systems. Our best translation system scores 5 to 6 BLEU higher than a strong baseline system provided by the organizers (Goyal et al., 2021). As seen in the Dynalab leaderboard, our submission is the only fully constrained submission that uses only the corpus provided by the organizers and does not use any pre-trained models.

1 Introduction

Neural Machine Translation (NMT) (Vaswani et al., 2017) has been shown to be effective with rich and in-domain bilingual parallel corpora. Although the NMT model obtained promising performances for high resource language pairs, it is hardly feasible to train translation models for all directions of the language pairs since the training progress is time- and resource-consuming. Recent work has shown the effectiveness of multilingual neural machine translation (MNMT), which aims to handle the translation from multiple source languages into multiple target languages with a single unified model (Johnson et al., 2017; Aharoni et al., 2019; Arivazhagan et al., 2019; Zhang et al., 2020; Fan et al., 2021; Goyal et al., 2021).

The MNMT model dramatically reduces training and serving costs. It is faster to train a MNMT model than to train bilingual models for all language pairs in both directions, and MNMT significantly simplifies deployment in production systems (Johnson et al., 2017; Arivazhagan et al., 2019). Further, parameter sharing across different languages encourages knowledge transfer, which improves low-resource translation directions and potentially enables zero-shot translation (i.e., direct translation of a language pair not seen during training) (Ha et al., 2017; Gu et al., 2019; Ji et al., 2020; Zhang et al., 2020).

We participate in the WMT 2021 multilingual machine translation task for small track #1. The task aims to train a multilingual model to translate 5 Central/East European languages (Croatian, Hungarian, Estonian, Serbian, Macedonian) and English in 30 directions. The multilingual systems presented in this paper are based on the standard paradigm of MNMT proposed by Johnson et al. (2017), which prefixes the source sentence with a special token to indicate the desired target language and does not change the target sentence at all. Language tags are typically used in MNMT to identify the language to translate to. A language code, in the form of a two- or three-character identification such as `en` for English, is the main constituent of a language tag and is provided by the ISO 639 standard[1] (International Organization for Standardization, nd). Following ISO 639 standard, `en` indicates English, `mk` indicates Macedonian, `sr` indicates Serbian, `et` indicates Estonian, `hr` indicates Croatian and `hu` indicates Hungarian in this paper.

Compared with the other three submissions to the task, our submissions have the following advantages:

- Our submissions are fully constrained, which means we using the data only provided by the organizer, and do not use models pre-trained on extra data.
- Our model only has 313M parameters, which

[1] https://en.wikipedia.org/wiki/ISO_639

Proceedings of the Sixth Conference on Machine Translation (WMT), pages 417–422
Novermber 10–11, 2021.

	Whole	Select
No filter	387M	71M
+ punctuation filter	384M	71M
+ deduplicated filter	304M	44M
+ `langid` filter	302M	43M
+ length filter	274M	42M

Table 1: Number of sentences in bitext datasets (total in 15 directions) for different filtering schemes. **Whole** denotes the use of all data provided by the organizers, **Select** denotes the use of data selection.

is smaller than the other submissions.

2 Data

The training data provided by the organizers come from the public available Opus repository (Tiedemann, 2012), which contains data of mixed quality from a variety of domains (WMT-News, TED, QED, OpenSubtitles, etc.). In addition to the bilingual parallel corpora, in-domain Wikipedia monolingual data for each language is provided. The validation and test sets are obtained from the Flores 101 evaluation benchmark (Goyal et al., 2021), which consists of 3001 sentences extracted from English Wikipedia covering a variety of different topics and domains. See Table 1 for details on data used for training our systems.

2.1 Data Preprocessing

To prepare the data for training, we used the following steps to process all of the corpora:

1. The datasets were truecased and the punctuation was normalized with standard scripts from the Moses toolkit[2](Koehn et al., 2007).

2. Sentences containing 50% punctuation are removed.

3. Duplicate sentences are removed.

4. We used a language detection tool[3] (`langid`) to filter out sentences with mixed language.

5. SentencePiece[4] (Kudo and Richardson, 2018) was used to produce subword units. We trained a model with 0.9995 character coverage to have sufficient coverage of character-based languages.

6. The length filtering removes sentences that are too long (more than 250 subwords after segmentation with Sentencepiece), sentences with a mismatched length ratio (more than 3.0) between source and target language are removed.

2.2 Data Selection

Data selection (Moore and Lewis, 2010; Axelrod et al., 2011; Gascó et al., 2012), aims to select the most relevant sentences from the out-of-domain corpora, which improved the in-domain translation performance. The training data provided by the organizers is large scale and contains multiple domains. Therefore, the data selection becomes a key factor affecting the performance of MNMT. Preliminary experiments (see in Table 1 model #3 and model #4) showed that the performance of using all corpora provided by the organizer was poor. Following the original paper (Goyal et al., 2021), we selected three data sources (CCAligned, MultiCCAligned, WikiMatrix) for further experimentation.

3 Method Description

We first trained bilingual translation models with 30 directions for all language pairs. Next, we trained a single multilingual model that can translate all language pairs. Finally, we use back-translation and knowledge distillation technologies to further improve the performance of the multilingual translation system. The details of these components are outlined next.

3.1 Bilingual NMT Model

We use Transformer (Vaswani et al., 2017) architecture for all bilingual models. To achieve the best BLEU score on the validation dataset, random search was used to select the hyperparameters since the datasets are in different sizes. We segment the data into subword units using SentencePiece jointly learned for all languages. The details of selected hyper-parameters are listed in Section 4.1.

3.2 Multilingual NMT Model

The multilingual model architecture is identical to the bilingual NMT model. To train multilingual models, we used a simple modification to the

[2] https://github.com/moses-smt/mosesdecoder/blob/master/scripts/tokenizer

[3] https://fasttext.cc/docs/en/language-identification.html

[4] https://github.com/google/sentencepiece

source sentence proposed by Johnson et al. (2017) which introduce an artificial token at the beginning of the source sentence indicating the target language (Johnson et al., 2017). For instance, for the English-Macedonian (en→mk) translation direction, we insert a token like <2mk> at the beginning of all English sentences and do not change the Macedonian sentences.

3.3 Back Translation

Back-translation (BT) (Sennrich et al., 2016) is a simple and effective data augmentation technique, which makes use of monolingual corpora and has proven to be effective. Back-translation first trains a target-to-source system that is used to translate monolingual target data into source sentences, resulting in a pseudo-parallel corpus. Then we mix the pseudo-parallel corpus with the authentic parallel data and train the the desired source-to-target translation system. Zhang et al. (2020) has shown how BT can be useful for multilingual MT.

After generating the pseudo parallel corpus, we tag our BT data by adding an artificial token <BT> at the beginning of the source sentence (Caswell et al., 2019), which indicates that the data is generated by back-translation.

3.4 Knowledge Distillation

Knowledge Distillation (KD) is a commonly used technique to improve model performance. The standard KD training (Kim and Rush, 2016) derives a student model from a teacher model by training the student model to mimic the outputs of the teacher. We follow a recent approach to KD proposed by Wang et al. (2021), which uses selection at the batch level and at the global level to choose suitable samples for distillation.

4 Experiments

4.1 Training Details

We use the Transformer architecture (Vaswani et al., 2017) as implemented in fairseq[5] (Ott et al., 2019). For training NMT and MNMT systems, we use the `Transformer-Big` architecture (hidden state 1024, feed-forward layer 4096, 16 attention heads, 6 encoder layers, 6 decoder layers). For optimization, we follow the default settings from the original paper (Vaswani et al., 2017) and used the Adam optimizer with a learning rate of 0.0003. To prevent overfitting, we applied a dropout of 0.3 on all layers. At the time of inference, a beam search of size 5 is used to balance the decoding time and accuracy of the search. The number of warm-up steps was set to 4000 and the vocabulary size is 133k. In addition, we set a length penalty factor of 1.7 to maintain a balance between long and short sentences. The batch size is set to 128 during decoding. We trained our models for approximately 3 weeks on one machine with 8 NVIDIA GTX 2080 Ti 11GB GPUs.

Because of the problems of the international tokenization in the standard BLEU score, the organizers used sentence-piece BLEU (spBLEU)[6] (Goyal et al., 2021) as the official evaluation metric which operates on strings segmented using a SentencePiece model. Recently, the BLEU score was criticized as an unreliable automatic metric (Mathur et al., 2020; Kocmi et al., 2021). Therefore, we also evaluate our models using chrF (Popović, 2015) and BERTScore (Zhang et al., 2019).

4.2 Systems

All of our systems described in Section 3.2 are listed as follows:

Flores. As a baseline system, we use the pretrained models public available by Flores teams. We use `flores101_mm100_615M` tested on the devtest datasets as our baseline.

Bilingual. We trained the bilingual models using standard `Transformer-Big` architecture for 6 languages in 30 directions. The hyperparameters used are discussed in Section 4.1.

Multilingual. We trained the multilingual translation model using standard `Transformer-Big` architecture and a specific language token to indicate the desired translation target language.

Tagged BT. We augment the training data by exploring the monolingual corpus using back-translation proposed by Caswell et al. (2019), with tagged back-translated source sentences with an extra token <BT>.

Selective KD. We focused on selective knowledge distillation proposed by Wang et al. (2021), which uses batch-level and global-level selections to pick suitable samples for distillation.

4.3 Results

The results of our systems on the devtest dataset are presented in Table 2. For models 1–4, we observed

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

[6] https://github.com/ngoyal2707/sacrebleu/tree/adding_spm_tokenized_bleu

#	Systems	spBLEU	chrF	BERTScore	BEST BLEU
0	Flores	28.0	0.528	0.867	sr-mk (36.0)
1	Bilingual$_{whole}$	21.1	0.477	0.831	en-mk (31.3)
2	Bilingual$_{select}$	28.4	0.533	0.863	sr-en (40.6)
3	Multilingual$_{whole}$	16.7	0.431	0.827	sr-en (26.1)
4	Multilingual$_{select}$	30.9	0.555	0.874	sr-en (40.0)
5	Multilingual$_{select}$ + TaggedBT(Multilingual$_{select}$)	30.7	0.548	0.873	sr-en (40.5)
6	Multilingual$_{select}$ + TaggedBT(Bilingual$_{select}$)	**32.3**	**0.562**	**0.879**	sr-en (41.5)
7*	Multilingual$_{select}$ + TaggedBT(Bilingual$_{select}$) + KD$_{batch}$	33.2	0.572	0.883	sr-en (42.0)
8*	Multilingual$_{select}$ + TaggedBT(Bilingual$_{select}$) + KD$_{global}$	**33.9**	**0.576**	**0.887**	sr-en (42.4)

Table 2: The automatic evaluation metrics on devtest data. **spBLEU**, **chrF**, **BERTScore** denotes the average scores of spBLEU, chrF and BERTScore respectively, **BEST BLEU** denotes the language pair with the best BLEU score. Systems with subscript *whole* denote the use of all data provided by the organizers, and systems with subscript *select* denote the use of data selection. Model #6 is our primary system submitted to the Dynalab leaderboard. Systems 7* and 8* were trained after the shared task and were not used for the final submission.

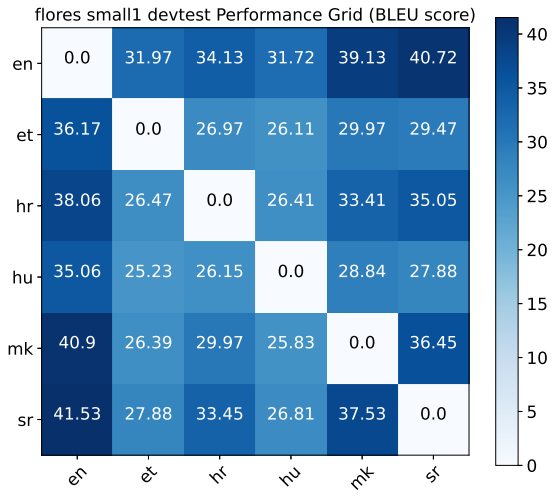

Figure 1: spBLEU scores on devtest data in 30 directions

that the amount of training data is not proportional to the performance of the model for the bilingual or multilingual translation model. The training data provided by the organizers contains multiple domains and does not match the dev/devtext/test data domain. Therefore, we apply the data selection methods to select data-relevant data from the training dataset to do the following experiments. Our multilingual model (#4) performs competitively with the Flores strong baseline (Model #0).

After these initial experiments, we explored how the bilingual models can be used to improve the multilingual model. More specifically, we use the Bilingual$_{select}$ model (#2) and Multilingual$_{select}$ model (#4) to back-translate the relevant monolingual corpora, and then we use the back-translations to train a new multilingual model. Although the overall performance of the Multilingual model (#4) is better than the Bilingual model (#2), back-translation using the Bilingual model (model #6) is better than back-translation using the Multilingual model (model #5). The possible reason is that the multilingual BT is in fact a form of self-training, but bilingual BT uses separate models, which means the knowledge obtained from bilingual BT models is more independent of the knowledge already learned by the baseline multilingual BT model.

Knowledge Distillation further improves performance slightly (Model # 7* and Model # 8*). Based on Model # 6, selective KD (Wang et al., 2021) is added to further improve the performance of the multilingual system.

Our best systems were outperformed by two other shared task submissions, which however used models pre-trained on additional data sources.

The performance grid of our best system (Model # 8*) is presented in Figure 1. We see from the results that the `sr-en` language pair produced the best results in terms of spBLEU score while the `hu-hr` language pair scored the lowest.

5 Conclusions

In this paper, we presented the LMU Munich system for the WMT 2021 Large-scale Multilingual Translation shared task for small track #1. The task evaluates translation between five central/eastern European languages and English, in total 30 translation directions. The system we submitted was fully constrained, using only the data provided by the organizers and not using any pre-trained model. The experiments show that back-translation and knowledge distillation techniques are effective for training multilingual machine translation systems.

6 Acknowledgments

We would like to thank the Flores team for the organization and for the grant of computer credits that we used in our experiments. This work was supported by funding to Wen Lai from LMU-CSC (China Scholarship Council) Scholarship Program (CSC, 202006390016). This work has received funding from the European Research Council under the European Union's Horizon 2020 research and innovation program (grant agreement #640550). This work was also supported by the DFG (grant FR 2829/4-1). We thank the other members of the machine translation group at CIS, LMU Munich, for their ideas and feedback.

References

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. 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)*, pages 3874–3884, Minneapolis, Minnesota. Association for Computational Linguistics.

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*.

Amittai Axelrod, Xiaodong He, and Jianfeng Gao. 2011. Domain adaptation via pseudo in-domain data selection. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 355–362, Edinburgh, Scotland, UK. 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.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Guillem Gascó, Martha-Alicia Rocha, Germán Sanchis-Trilles, Jesús Andrés-Ferrer, and Francisco Casacuberta. 2012. Does more data always yield better translations? In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 152–161, Avignon, France. Association for Computational Linguistics.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzman, and Angela Fan. 2021. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *arXiv preprint arXiv:2106.03193*.

Jiatao Gu, Yong Wang, Kyunghyun Cho, and Victor O.K. Li. 2019. Improved zero-shot neural machine translation via ignoring spurious correlations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1258–1268, Florence, Italy. Association for Computational Linguistics.

Thanh-Le Ha, Jan Niehues, and Alexander Waibel. 2017. Effective strategies in zero-shot neural machine translation. *arXiv preprint arXiv:1711.07893*.

Baijun Ji, Zhirui Zhang, Xiangyu Duan, Min Zhang, Boxing Chen, and Weihua Luo. 2020. Cross-lingual pre-training based transfer for zero-shot neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 115–122.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Yoon Kim and Alexander M. Rush. 2016. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1317–1327, Austin, Texas. Association for Computational Linguistics.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. *arXiv preprint arXiv:2107.10821*.

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 Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

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*.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997, Online. Association for Computational Linguistics.

Robert C. Moore and William Lewis. 2010. Intelligent selection of language model training data. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 220–224, Uppsala, Sweden. 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. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. 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)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

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.

Fusheng Wang, Jianhao Yan, Fandong Meng, and Jie Zhou. 2021. Selective knowledge distillation for neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6456–6466, Online. Association for Computational Linguistics.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639, Online. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.