University of Tartu's Multilingual Multi-domain WMT19 News Translation Shared Task Submission

Andre Tättar **Elizaveta Korotkova** **Mark Fishel**
Institute of Computer Science
University of Tartu, Estonia
{andre.tattar,elizaveta.korotkova,fishel}@ut.ee

Abstract

This paper describes the University of Tartu's submission to the news translation shared task of WMT19, where the core idea was to train a single multilingual system to cover several language pairs of the shared task and submit its results. We only used the constrained data from the shared task. We describe our approach and its results and discuss the technical issues we faced.

1 Introduction

Typically the majority of WMT news translation shared task submissions are based on language pair-specific machine translation (MT) systems (Bojar et al., 2016, 2017, 2018). However, recently several multilingual approaches to MT have been proposed (e.g. Johnson et al., 2017; Vázquez et al., 2018; Aharoni et al., 2019). With them as inspiration, the goal of this paper is to describe our submission to the WMT'2019 news translation shared task, where we trained a single multilingual translation system using the constrained parallel and monolingual data for several language pairs.

In addition to multilinguality we wanted to incorporate the multiple text domains that constitute the constrained set of parallel corpora in the WMT shared task. We approach multi-domain NMT using the method of (Tars and Fishel, 2018): namely, by treating domains as separate languages, therefore creating a "double-multilingual" system.

In addition to multilinguality and multi-domain NMT our submission has more common features, like data filtering, ensembles of several models and fine-tuning on back-translated monolingual data.

Below we describe the architecture of our approach in Section 2, experimental setup in Section 3, results and analysis in Section 4 and conclude the paper in Section 5.

2 Architecture

Our model is a neural MT system based on autoregressive self-attention in the encoder and decoder (Vaswani et al., 2017). We achieve multilinguality in a similar fashion to (Johnson et al., 2017): using an additional input specifying the output language, so that the system would know which language to generate. Differently from Johnson et al. (2017), who include the output language into the input segment itself, we use word factors (Hieber et al., 2017) and specify the output language as a factor of each input token.

In addition to multilinguality, our NMT system also uses the information on which domain the parallel or monolingual corpora come from. The WMT data consist of a variety of text domains (parliamentary speeches, crawled web and news texts, press releases, Wikipedia titles, etc.) and it has been shown (Tars and Fishel, 2018) that multi-domain NMT can get much better results than the default approach of mixing heterogeneous corpora together, as well as yield more efficient solutions than fine-tuning to each domain separately. Our solution is to specify the output text domain as another word factor.

One peculiarity of multilingual NMT is that the model performs back-translation for itself, therefore avoiding the necessity of training more than one translation system.

3 Experiments

3.1 Model Setup

We use the Sockeye (Hieber et al., 2017) machine translation framework for our experiments. The main reason behind this choice is that Sockeye

Proceedings of the Fourth Conference on Machine Translation (WMT), Volume 2: Shared Task Papers (Day 1) pages 382–385
Florence, Italy, August 1-2, 2019.

	CZ-EN	DE-EN	DE-FR	EN-FI	EN-LT	TOTAL
NEWS	2534352	5985498	4372033	2656508	1803323	17351714
OFF	11462432	1797854	1687074	1725792	615219	17288371
SUBS	37251088	-	-	-	-	37251088
OTHER	10932478	34457911	7585341	4012589	1290931	58279250
TOTAL	62180350	42241263	13644448	8394889	3709473	130170423

Table 1: Dataset sizes after filtering. Shown number of parallel sentences.

implements word factors together with the Transformer.

We use traditional transformer NMT architecture with 6 layers for both encoder and decoder, with the transformer model size 1024, transformer attention heads 16, batch size 6000, with a shared byte-pair encoded (BPE) (Sennrich et al., 2015) vocabulary of size 90000. SentencePiece[1] are used to extract BPE vocabulary. The embedding size for source factors is 8. There are 6 different language factors and 4 different domain factors. All other parameters were kept as default.

Models are trained on 4 Tesla V100 GPUs.

3.2 Data

All of the available WMT constrained data for all languages was downloaded and then fed through a data pipeline. The data pipeline consisted of 6 steps:

1. **Filtering** Data filtering included several steps: it filtered out empty/too long sentences, sentences with too many non-alphanumeric characters, sentences where the length difference was too big, and also sentences automatically identified as a different language than the expected one.

2. **Tokenization** The data was tokenized with MosesTokenizer.

3. **Truecasing** A Truecasing model was trained for every language separately, then applied on all the data.

4. **SentencePiece** A SentencePiece model was trained on one big text file which included all data, low-resource language pairs like EN-LT were upscaled and high-resource language pairs like CZ-EN were downscaled. In total 50M lines of text were used for SentencePiece model with vocabulary size 90K.

5. **Factoring** Then the source factors for target domain and target language were generated for all data.

6. **Sharding** Sockeye uses shards to handle massive datasets, which means that a big dataset is divided into more manageable dataset sizes. Each shard is of equal size. A shard size of 1M was used.

Due to time constraints we deviated from the original plan of including all WMT'2019 language pairs and only included languages that use the Latin script in our submissions. The final data set sizes are shown in Table 1.

In order to generate the domain factors we grouped some of the domains by the apparent similarity of texts, additionally grouping smaller corpora together:

- **News** - Rapid2019, Rapid2016, EESC, dev dataset from previous years, EMEA2016, ECB2017, news (from CzEng), News-commentary
- **Subs** - Subtitles from the CzEng corpus
- **Off** - Parts of the CzEng corpus, Europarl
- **Other** - Everything else

Additionally, monolingual data was extracted for back-translation and fine-tuning, mainly News Crawl corpora was used. For every language pair 3M sentences were extracted, with the exception of Lithuanian, where the news crawl size is smaller, and thus other monolingual data like Wiki dumps and Europarl were used.

4 Results and Analysis

Results are presented in Table 2. We separate the results of our **baseline** system, trained on parallel data only, and the **fine-tuned** system that was trained further on monolingual data, back-translated by the baseline system.

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

	Baseline	Fine-tune
EN-CS	22.8	-
DE-EN	29.9	-
EN-DE	39.6	-
DE-FR	32.4	30.7
EN-FI	18.6	-
EN-LT	12.7	-
FI-EN	22.1	24.8
FR-DE	25.9	-
LT-EN	24.5	25.3

Table 2: Results of our multilingual baseline model, trained on parallel data and the fine-tuned model that was further trained on back-translated monolingual data.

	#Sents	#Baseline Wrong	#Ensemble Wrong
DE-EN	33650	214	18596
DE-FR	1698	3	117
EN-CS	9917	256	10137
EN-DE	8853	85	6396
F EN-FI	2606	221	2799
EN-LT	1056	11	1066
FI-EN	4105	8	76
FR-DE	2705	6	843
Total	65684	809	40054
%		1.22	60.24

Table 3: Number of sentences which are classified as having a wrong language after translation using the FastText language classifier.

Our baseline performed reasonably well, however the goal was to achieve state-of-the-art results after doing fine-tuning on back-translated news data. As a result of this second step unexpectedly the model started confusing the output language and generating the output in a different language than requested: for example generating Czech or English instead of Finnish. Automatic language identification with FastText[2] shows the baseline model only produced output in the wrong language in 1.22% of cases, whereas after just a day of fine-tuning on in-domain data, the percentage of translations our model got wrong jumped up to 60.24%. Mostly our ensemble model got English right and other languages wrong. Our ensemble model was done by using 2 snapshots of baseline model and 2 snapshots of fine-tuned model.

For human evaluations published in (Bojar et al., 2019) our model (called TartuNLP-c) performed similarly to other multilingual systems noted as Online-X in the findings paper. Online systems are freely available online systems like Google Translate, Bing Translate etc. Our models performed worse than single language pair NMT systems.

We suspect that the reason for the wrong language output lies in two factors:

- wrong language segments in monolingual crawled data. This mainly occurs in non-English languages like Czech, Finnish and Lithuanian and affects the output side of back-translated data. Before the submission deadline we did not have language-filtering in the data preparation pipeline, which might have caused this effect.
- wrong language output by our model. This affects the input side of the back-translated data. While this does not occur often, filtering out the wrong-language translations should still help learn a more precise translation model.

We are investigating alternative explanations to this behavior further.

5 Conclusions and Future Work

We have described a multilingual multi-domain neural machine translation approach that can be trained on a mixture of different language pairs and text domains.

Our results are modest, mainly due to failing to properly fine-tune the systems on back-translated news texts. Precise reasons for failing the fine-tuning are under investigation.

Other future work includes including more languages and domains, testing online continuous back-translation and experimenting with other ways of providing the output language and domain information to the NMT model.

Acknowledgments

This work was supported in part by funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 825303, Estonian Research Council grant no. 1226 and the AWS Cloud Credits for Research program.

[2] https://github.com/facebookresearch/fastText

References

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. In *Proceedings of NAACL 2019*, page (accepted), Minneapolis, MN, USA.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 conference on machine translation (WMT17). In *Proceedings of the Second Conference on Machine Translation*, pages 169–214, Copenhagen, Denmark.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurelie Neveol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation*, pages 131–198, Berlin, Germany.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of WMT'18: the Third Conference on Machine Translation*, Brussels, Belgium.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, and Christof Monz. 2019. Findings of the 2019 conference on machine translation (wmt19). In *Proceedings of the Fourth Conference on Machine Translation, Volume 2: Shared Task Papers*, Florence, Italy. Association for Computational Linguistics.

Felix Hieber, Tobias Domhan, Michael Denkowski, David Vilar, Artem Sokolov, Ann Clifton, and Matt Post. 2017. Sockeye: A toolkit for neural machine translation. *CoRR*, abs/1712.05690.

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.

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

Sander Tars and Mark Fishel. 2018. Multi-domain neural machine translation. In *Proceedings of EAMT*, pages 259 – 268, Alicante, Spain.

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 *Proceedings of NIPS*, pages 5998–6008, Long Beach, CA, USA.

Raúl Vázquez, Alessandro Raganato, Jörg Tiedemann, and Mathias Creutz. 2018. Multilingual NMT with a language-independent attention bridge. *CoRR*, abs/1811.00498.