This document is part of the Research and Innovation Action "European Live Translator (ELITR)". This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No 825460.

Deliverable D4.2

Intermediate Report on Multi-Lingual MT

Dominik Macháček (CUNI), Vilém Zouhar (CUNI), Rico Sennrich (UEDIN),
Sunit Bhattacharya (CUNI), Thanh-Le Ha (KIT),
Bohdan Ihnatchenko (CUNI), Ondřej Bojar (CUNI)

Dissemination Level: Public

Final (Version 1.0), 31st December, 2020

Grant agreement no.	825460
Project acronym	ELITR
Project full title	European Live Translator
Type of action	Research and Innovation Action
Coordinator	doc. RNDr. Ondřej Bojar, PhD. (CUNI)
Start date, duration	1st January, 2019, 36 months
Dissemination level	Public
Contractual date of delivery	31st December, 2020
Actual date of delivery	31st December, 2020
Deliverable number	D4.2
Deliverable title	Intermediate Report on Multi-Lingual MT
Type	Report
Status and version	Final (Version 1.0)
Number of pages	43
Contributing partners	CUNI, UEDIN, KIT
WP leader	UEDIN
Author(s)	Dominik Macháček (CUNI), Vilém Zouhar (CUNI), Rico Sennrich (UEDIN), Sunit Bhattacharya (CUNI), Thanh-Le Ha (KIT), Bohdan Ihnatchenko (CUNI), Ondřej Bojar (CUNI)
EC project officer	Alexandru Ceausu
The partners in ELITR are:	▪ Univerzita Karlova (CUNI), Czech Republic ▪ University of Edinburgh (UEDIN), United Kingdom ▪ Karlsruher Institut für Technologie (KIT), Germany ▪ PerVoice SPA (PV), Italy ▪ alfatraining Bildungszentrum GmbH (AV), Germany
Partially-participating party	▪ Nejvyšší kontrolní úřad (SAO), Czech Republic

For copies of reports, updates on project activities and other ELITR-related information, contact:

doc. RNDr. Ondřej Bojar, PhD., ÚFAL MFF UK
Malostranské náměstí 25
118 00 Praha, Czech Republic
bojar@ufal.mff.cuni.cz
Phone: +420 951 554 276
Fax: +420 257 223 293

Copies of reports and other material can also be accessed via the project's homepage:

`http://www.elitr.eu/`

Contents

1 Executive Summary **4**

2 Task T4.2 Document-Level Machine Translation (CUNI, UEDIN) **4**

3 Task T4.3 Multi-Target MT (CUNI, UEDIN, KIT) **5**
3.1 Massively Multi-Lingual Models 5
3.2 Large Eastern Europe Multi-Target Models 5
3.3 Unsupervised Transfer Learning for Multilingual Systems 6
3.4 Exploring Mid-Sized Multi-Lingual Models 7

4 Task T4.4 Multi-Source MT (CUNI, UEDIN, KIT) **8**
4.1 Multi-Source Analysis with Multi30k 8
4.2 Initial Analysis of Interpretation Corpus 9

References **9**

Appendices **12**

Appendix A Document-Level Markable Error Exploration **12**

Appendix B Improving Massively Multilingual Neural Machine Translation and Zero-Shot Translation **22**

Appendix C Dynamics of Multilingual Translation **34**

1 Executive Summary

This deliverable summarizes the progress in WP4 Multi-Lingual MT during the first two years of the project. We briefly refer to the work that has been completed and reported in Deliverable D4.1: Initial Report on Multi-Lingual MT, and report all necessary details of new work.

The work package consists of 5 tasks:

T4.1 Baseline MT Models was planned and carried out during the first 6 months of the project. It provided MT systems to the rest of the main processing pipeline, so that integration and technical testing could start. All the details regarding our baseline models are in the previous Deliverable D4.1: Initial Report on Multi-Lingual MT.

We have bilingual or multi-target baseline models for all 43 EUROSAI languages.

T4.2 Document-Level Translation is a research goal somewhat more independent of the remaining tasks. The aim is to substantially improve the practice of handling document-level context across MT processing stages: training, translation and evaluation. In Section 2, we report a new study on document-level MT evaluation.

T4.3 Multi-Target Translation explores the setup most needed for ELITR central event, the EUROSAI congress, where a single speech needs to be translated into up to 43 target languages. In Section 3, we report our new research progress in massively multi-lingual MT.

T4.4 Multi-Source Translation aims to improve translation quality by considering other language versions of the same content. The task is scheduled to start in year 2 and can consider both written or spoken multi-source. As preparatory steps ahead of time, we have begun gathering data from training lessons of interpreters to assess if multi-source could be applied in the ELITR setup of live conference interpretation. More details are in Section 4.

T4.5 Flexible Multi-Lingual MT is planned for year 3 of the project.

2 Task T4.2 Document-Level Machine Translation (CUNI, UEDIN)

Our progress in the first year of the project (Vojtěchová et al., 2019; Rysová et al., 2019; Popel et al., 2019; Voita et al., 2019b,a; Popel et al., 2019) is reported in D4.1.

In 2020, in the second year of the project, Zouhar et al. (2020), CUNI, elaborated a new work on evaluation of document-level translation focused on markables (key terms or expressions to the meaning of the document) and the negative impact of various markable error phenomena on the translation.

The full version of the paper is provided in Appendix A. For the annotation experiment, we chose Czech and English documents (news, audit and lease domains) translated by systems submitted to WMT20 News Translation Task. Short of annotators expert in the auditing domain, we sought for methods where also non-experts could assess the translation quality. We compared three approaches to document translation evaluation, and saw that non-expert annotators rate most MT systems higher than reference with fluency and adequacy, but reference still ranks better than most of them when inspecting markable phenomena and their severity. Inspecting specific markable instances in detail, we found out that MT systems made errors in term translation, which no human translator would do.

Relating the current observation with the impression from the last year (Vojtěchová et al., 2019), we conclude that annotators lacking in-depth domain knowledge are not reliable for annotating on the rather broad scales of fluency and adequacy, but they are capable of spotting term translation errors in the markable style of evaluation. This is important news because expert annotators can not be always secured. The method still however has to be improved

because the inter-annotator agreement was low, possibly due to a rather high number of MT systems compared at once.

3 Task T4.3 Multi-Target MT (CUNI, UEDIN, KIT)

Last year's progress is in D4.1 (Zhang and Sennrich, 2019; Zhang et al., 2019) and in Ihnatchenko (2020).

3.1 Massively Multi-Lingual Models

In D4.1, we reported on our progress towards training deep neural models (Zhang et al., 2019), and reported first results where we applied the technique of a universal multi-lingual NMT model (Johnson et al., 2016), as well as using a mix of shared and language-specific model parameters, in a massively multilingual setting . We have since extended these experiments to consider an English-centric many-to-many setting, using OPUS-100 as our testbed. Our main findings are as follows:

- deeper models and language-specific parameters bring improvements of about 4 BLEU on average in a massively multilingual setting compared to a standard Transformer base.
- in an English-centric setup, we still observe performance degradation (0.5 BLEU) compared to dedicated bilingual systems when translation out of English. When translating into English, we observe large improvements over dedicated bilingual systems (4.8 BLEU), which we attribute to the benefits of sharing the English language generation component across translation directions.
- zero-shot translation, translation quality for translation directions with no parallel training data, is relatively poor, even with our best models, but can be substantially improved with the use of random online backtranslation.

Full details can be found in (Zhang et al., 2020), published at ACL 2020 and reproduced here in Appendix B. For ELITR production systems, we draw the following lessons:

- we will not exclusively rely on massively multilingual models, but will use a mix of bilingual models (for best quality on the most important translation directions) and multilingual models (for wide coverage).
- we are currently investigating the balance of quality and efficiency to decide on whether we can deploy deeper models in production. We also note recent research that shows better speed–quality trade-offs with an asymmetric depth of encoder and decoder (Kasai et al., 2020); we do not yet know how well such configurations scale in multilingual setups, but will consider this for deployment.

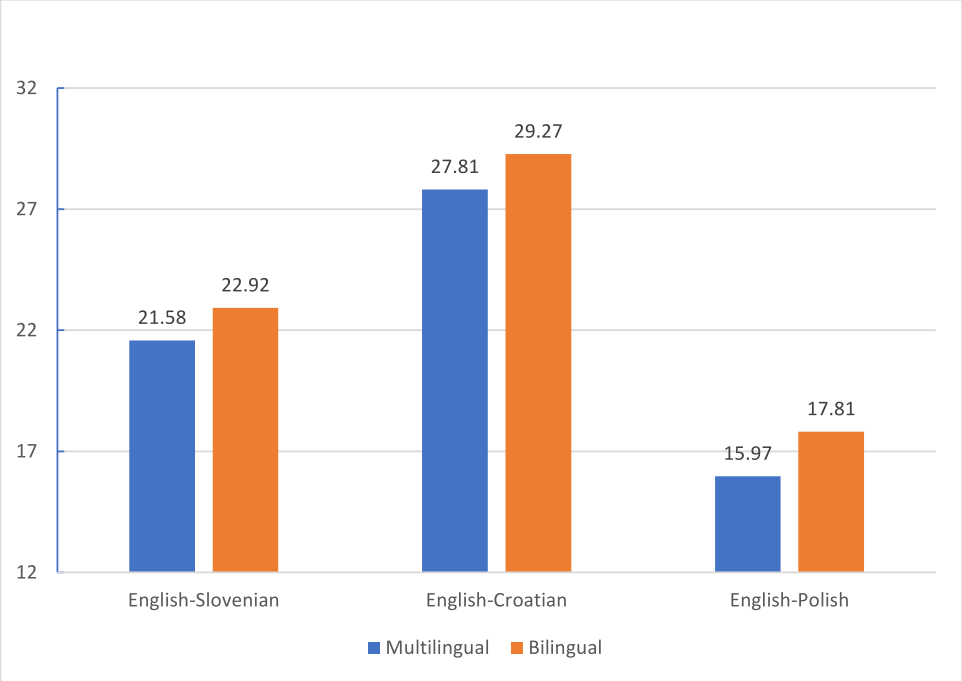
3.2 Large Eastern Europe Multi-Target Models

At KIT, we investigate a large multi-target system which translates from English into three different Eastern European languages: Croatian, Slovenian and Polish. The system has been trained using relatively large corpora harvested from different sources. Table 3.2 shows the corpora we used and their sizes.

The motivations for building such a multilingual system is that Croatian, Slovenian and Polish belong to West and South Slavic language families, sharing similar alphabets and it is convenient to train and deploy a single multilingual translation model instead of building several bilingual ones.

Various studies show that multilingual systems for high-resource languages, i.e. being trained with large corpora, without any special treatment, usually perform worse than the bilingual

Dataset	English-Slovenian	English-Croatian	English-Polish
DGT v2019	3,626,000	669,000	3,665,000
TED	16,800	105,900	166,500
EUconst	6,600	-	6,300
Europarl v8	605,000	-	613,000
GlobalVoices	-	-	48,900
JRC-Acquis	31,000	-	1,055,200
QED	74,900	196,600	482,800
HrWaC	-	96,400	-
OpenSubtitles v2018	16,400,000	28,800,000	34,600,000
ParaCrawl	3,738,000	6,959,000	13,745,000
Tatoeba	3,200	2,400	54,200
SETimes	-	204,100	-
Wikipedia	81,500	2,200	168,600
TOTAL	24,583,000	37,035,600	54,605,500



systems trained on the same amount of data. Our experiments (Figure 3.2) confirm this. However, the reducing cost of training and deployment in our case compromises the degradation of the multilingual system compared to the bilingual systems.

3.3 Unsupervised Transfer Learning for Multilingual Systems

We have been exploring the way to add new languages into already-trained multilingual systems without having parallel or multilingual data of those languages and the existing languages in such systems (i.e. *zero-shot translation* and *continual learning*). Basically, we performed unsupervised transfer learning from the multilingual system with augmented cross-lingual embeddings so that the system can translate unseen inputs from a new language. In particular, the transfer learning consists of the following steps:

1. Train the multilingual system for the language set $\mathcal{L}_{\text{base}}$ using the universal approach (Ha et al., 2016) with the shared embeddings between source and target sides.

2. Create cross-lingual embeddings for all languages including the new language ℓ_{new} we want to add by aligning all the monolingual word embeddings for each language $\ell \in \mathcal{L}_{\text{base}} \cup \ell_{\text{new}}$

from *fastText*'s pre-trained word embeddings (Bojanowski et al., 2017) into a common space following the cross-lingual embedding alignment approach (Joulin et al., 2018)

3. Augmented the multilingual system by replacing its embeddings with the cross-lingual embeddings which have been learned from Step 2.

4. Fine-tune the augmented multilingual system on the synthetic parallel data which is created by taking a noised version of the monolingual corpus of the language ℓ_{new} as the source and its original version as the target. The noised version is derived by slightly shuffling some of the words for every sentence in the monolingual corpus. The purpose is to learn the syntax of that language ℓ_{new} via this kind of denoising auto-encoding.

We experimented with two new languages: Portuguese (Pt) and Russian (Ru), adding to the existing multilingual system of German (De), English (En), Spanish (Es), French (Fr) and Italian (It). Table 3.3 summarizes our experimental results. We can see that our method performs better when adding Portuguese, which has the same alphabet and belongs to the same or similar language families to the existing languages in the multilingual system (Romance in case of Spanish, French and Italian and West Germanic in case of English and German). On the other hand, the performance of Russian system is worse due to the fact that it has a different alphabet (Cyrillic) and belongs to a different language family (East Slavic).

	De-Pt	En-Pt	Es-Pt	Fr-Pt	It-Pt	De-Ru	En-Ru
Unsupervised	17.0	28.1	27.1	21.5	20.8	8.1	8.7
Supervised	21.9	35.8	32.9	27.1	26.2	15.1	17.2

It is expected that our method performs worse than the supervised setting which uses parallel corpora, e.g. German-Portuguese. But there are several important advantages that our method brings:

- We do not need to have parallel data. It is extremely helpful if we work with a low-resource language that does not have any parallel data to the existing languages in the system.

- We can perform continual learning. If we want to translate to the new language, we do not need to train the multilingual system with the new language from scratch.

3.4 Exploring Mid-Sized Multi-Lingual Models

In D4.1 (Section 4.2), we have described a series of experiments we have conducted with training mid-sized multi-lingual models using the ELITR OPUS Corpus v1.0 (described in D1.1). We then extended the experiments with:

- removing the test-train overlapping sentences from the training set;

- training more models.

The results were presented in the master thesis by (Ihnatchenko, 2020).

We arranged the experiments in the same way as in D4.1. We wanted to check how the relatedness of target languages affects the model's translation performance for specific translation directions. To do so, we trained two groups of models: with randomly selected target languages and with the related ones. The "random" group consisted of models where we added random target languages to the mix. The "related" group contained the models, where we have added a group of mutually-related target languages to the mix, i.e., for the En→{De, L1, L2, ...} model, where L1, L2, and others are some of the Germanic languages, such as Dutch, Norwegian, etc.

We later evaluated the selected translation directions of all the trained models on the domain-specific test sets. E.g., for En→De translation direction, in the ELITR OPUS Corpus v1.0, there are such test sets as, for example, News-Commentary/v9.1, MultiUN/v1, OpenSubtitles/v2018,

etc. Each of such domain-specific and translation-direction-specific test sets contains 2000 sentence pairs. The results are later aggregated also by the number of target languages so that for each domain-specific test set there are results for En→{De+1 random language}, En->{De+1 Germanic language}, En→{De+2 random languages}, En→{De+2 Germanic languages}, etc. After that, we compared these mean values for each translation direction, test set, and the number of added target languages separately.

In some cases, we observed an improvement of 1–1.5 BLEU. In most cases, the improvement of mean results was less than 0.5 BLEU for a test set containing 2000 sentence pairs, which may not be enough to draw a strong conclusion (Card et al., 2020).

In other words, selecting the related target languages insignificantly improves the model's performance compared to random languages in the multi-target model, given a fixed test set and a fixed number of target languages in the model. The setting with related languages in the model almost never lead to a decrease compared to random selection.

In most cases, however, multi-target models perform slightly worse than bilingual baselines. The only exceptions are the setups where the bilingual model is under-performing, e.g. due to bad domain match of the training and test data. There the multilingual setup can bring some minimal improvement.

We thus arrive at this recommendation for the rest of the project:

- If affordable in terms of hardware resources, use bilingual models.
- If you have to resort to multi-lingual ones, prefer related target languages.
- In other situations, multi-target models with a random set of target languages are acceptable and not causing too bad loss of performance. This was experimentally tested with up to 5 target languages in the mix.

4 Task T4.4 Multi-Source MT (CUNI, UEDIN, KIT)

This task aims to research ways to create one NMT model, which can translate from multiple parallel language versions of the source into one or more target languages. It may be useful in situations, where translations into many targets must be created in a short time. The existing parallel translations may simplify translating into other languages, it may help to disambiguate the context. On the other hand, it is possible that the ambiguous words that need disambiguation by a combination of languages, are very rare and that the second source confuses the model, so the benefits may not overweight the costs.

We are somewhat behind our plans in multi-source experiments. We have yet to examine large data setting which is more relevant for practical use. So far, we research the behaviour of parallel multi-source models on a small trial data set Multi30k (Section 4.1).

The second usecase of parallel multi-source is the spoken language, the conferences with simultaneous interpretation. The SLT relies on automatic speech recognition, which is very sensitive to the speaker, non-native accents, domain, etc. We assume that a second source may improve the quality of recognition and translation. However, we miss a large data set for the research, and therefore we first create and analyze it. See Section 4.2.

4.1 Multi-Source Analysis with Multi30k

We experimented with Multi-Source MT by experimenting with setups with up to three source languages and one target language (details in Appendix C). The experiments concentrated on trying to study the distribution of attention among different sources and the decoder during prediction of an output token in a forced decoding setting. For the experiments, we used the architectures proposed in Libovický and Helcl (2017) and used the sentinel mechanism proposed in Lu et al. (2017). We used German, French and Czech as source languages and English as the target language for the experiments. The results were presented at the 56th Linguistics Colloquium and our main findings can be summarized as follows:

- In our multi-source models, one single language always ended being relied on by the decoder for prediction of most tokens.
- The inclusion of French as one of the source languages (in models with two and three encoders) improved the BLEU score of the model. However, the French source was never much consulted by the decoder during the prediction of the output tokens.
- In the case of shorter sentences, the model relied less on the information from the source and instead chose to trust its own output.

4.2 Initial Analysis of Interpretation Corpus

We aim to propose a parallel multi-source speech translation system. A machine, contrary to humans, can process multiple parallel input audio streams at once, e.g. an original speaker and one or more simultaneous interpreters, and can use all of them as sources for translation. We suppose that the additional source may help to disambiguate the context and improve the quality, and that it can be used in simultaneous translation.

The first subproblem for the research in this area are the data. There is no corpus of simultaneous interpretation (SI) usable for our work. Therefore, we decided to first create our own corpus from European Parliament (see deliverable D1.5) and use it in the subsequent research on this task. This situation makes us to continue with task T4.4 in the following year.

Although the corpus is not yet completed, we have already started to analyze human SI on the first part. We intend to measure the delay or other features of human SI, which we can use for research of simultaneous machine translation.

We automatically detected the starts and ends of the individual speeches, and processed them by ASR. We filtered out those which were probably not correctly processed, and compiled a set of 4 127 speeches (around 131 hours) given originally in English, with SI into Czech. We processed word alignments, and segmented the speeches to presumably parallel parts, by selecting the aligned words on diagonal, which separated the speech to non-overlapping parts. This approach is very inaccurate, and we plan more work on improving the alignments by validation on manually transcribed validation set, using the timing information properly, etc. However, by this approach we could already measure that the median ear-voice-span delay (the lag of the the interpreter behind the source) is 3.02 seconds, and 99% of words are translated by SI in less than 7.81 seconds. From this, we can conclude the following:

- Machine simultaneous translation should probably target to the delay of 8 seconds, to be comparable to human SI.
- Machine multi-source speech translation may expect the parallel words within 8-second windows.

References

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017. ISSN 2307-387X.

Dallas Card, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald, and Dan Jurafsky. With little power comes great responsibility. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9263–9274, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.745. URL `https://www.aclweb.org/anthology/2020.emnlp-main.745`.

Thanh-Le Ha, Jan Niehues, and Alexander H. Waibel. Toward Multilingual Neural Machine Translation with Universal Encoder and Decoder. In *Proceedings of the 13th International Workshop on Spoken Language Translation*, Seattle, WA, USA, December 2016. URL `https://dblp.org/rec/journals/corr/HaNW16.bib`.

Bohdan Ihnatchenko. Multi-Target Machine Translation. Master's thesis, Univerzita Karlova, Matematicko-fyzikální fakulta, Ústav formální a aplikované lingvistiky, Praha, 2020.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda B. Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's multilingual neural machine translation system: Enabling zero-shot translation. *CoRR*, abs/1611.04558, 2016. URL `http://arxiv.org/abs/1611.04558`.

Armand Joulin, Piotr Bojanowski, Tomas Mikolov, Hervé Jégou, and Edouard Grave. Loss in translation: Learning bilingual word mapping with a retrieval criterion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018.

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

Jindřich Libovický and Jindřich Helcl. Attention strategies for multi-source sequence-to-sequence learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Vancouver, Canada, August 2017. Association for Computational Linguistics.

Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 375–383, 2017.

Martin Popel, Dominik Macháček, Michal Auersperger, Ondřej Bojar, and Pavel Pecina. English-czech systems in wmt19: Document-level transformer. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 342–348, Florence, Italy, August 2019. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W19-5337`.

Kateřina Rysová, Magdaléna Rysová, Tomáš Musil, Lucie Poláková, and Ondřej Bojar. A test suite and manual evaluation of document-level nmt at wmt19. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 455–463, Florence, Italy, August 2019. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W19-5352`.

Elena Voita, Rico Sennrich, and Ivan Titov. Context-aware monolingual repair 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)*, pages 876–885, Hong Kong, China, November 2019a. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/D19-1081`.

Elena Voita, Rico Sennrich, and Ivan Titov. When a Good Translation is Wrong in Context: Context-Aware Machine Translation Improves on Deixis, Ellipsis, and Lexical Cohesion. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, 2019b. Association for Computational Linguistics. URL `https://arxiv.org/pdf/1905.05979.pdf`.

Tereza Vojtěchová, Michal Novák, Miloš Klouček, and Ondřej Bojar. Sao wmt19 test suite: Machine translation of audit reports. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 481–493, Florence, Italy, August 2019. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W19-5355`.

Biao Zhang and Rico Sennrich. Root Mean Square Layer Normalization. In *Advances in Neural Information Processing Systems 32*, Vancouver, Canada, 2019.

Biao Zhang, Ivan Titov, and Rico Sennrich. Improving Deep Transformer with Depth-Scaled Initialization and Merged Attention. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing*, Hong Kong, China, 2019. Association for Computational Linguistics.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 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, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.148. URL `https://www.aclweb.org/anthology/2020.acl-main.148`.

Vilém Zouhar, Tereza Vojtěchová, and Ondřej Bojar. Wmt20 document-level markable error exploration. In *Proceedings of the Fifth Conference on Machine Translation*, pages 369–378, Online, November 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.wmt-1.41.

A Document-Level Markable Error Exploration

WMT20 Document-Level Markable Error Exploration

Vilém Zouhar **Tereza Vojtěchová** **Ondřej Bojar**

Charles University, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Malostranské náměstí 25, 118 00 Prague, Czech Republic
{zouhar, vojtechova, bojar}@ufal.mff.cuni.cz

Abstract

Even though sentence-centric metrics are used widely in machine translation evaluation, document-level performance is at least equally important for professional usage. In this paper, we bring attention to detailed document-level evaluation focused on markables (expressions bearing most of the document meaning) and the negative impact of various markable error phenomena on the translation.

For an annotation experiment of two phases, we chose Czech and English documents translated by systems submitted to WMT20 News Translation Task. These documents are from the News, Audit and Lease domains. We show that the quality and also the kind of errors varies significantly among the domains. This systematic variance is in contrast to the automatic evaluation results.

We inspect which specific markables are problematic for MT systems and conclude with an analysis of the effect of markable error types on the MT performance measured by humans and automatic evaluation tools.

1 Introduction

This paper presents the results of our test suite for WMT20 News Translation Task.[1]

The conclusion of Vojtěchová et al. (2019), a last year's similar effort, states that expert knowledge is vital for correct and comprehensible translation of professional domains, such as Audits or Lease agreements. Furthermore, even MT systems which make fewer mistakes and score above others in both automatic and manual evaluations are prone to making fatal errors related to markable conflicts, which render the whole document translation unusable.

[1] http://www.statmt.org/wmt20/translation-task.html

In this study, we aim to organize and describe a more detailed study with a higher number of annotators. We show three evaluation approaches: (1) automatic evaluation, (2) fluency and adequacy per document line and (3) detailed markable phenomena evaluation. We compare the results of this evaluation across the three domains and try to explain why all of these evaluations do not produce the same ordering of MT systems by performance.

This paper is organized accordingly: Section 1.1 defines the term "Markable", Section 1.2 describes the examined documents and Section 2 introduces the two phases of our annotation experiment and shows the annotator user interface in Section 2.3. In Section 3, we discuss the results from both phases and also automatic evaluation. The main results of this examination are shown in Section 3.5 and specific markable examples are discussed in Section 4. We conclude in Section 5.

1.1 Markable Definition

A markable in this context is an occurrence of any technical or non-technical term or expression that satisfies at least one of the following conditions:

1. The term was translated into two or more different ways *within one document*.
2. The term was translated into two or more different ways *across several translations*.
3. Two or more terms were translated to a specific expression *in one document* but have different meanings.

To be a markable, the term or expression does not have to be a named entity, but it must be vital to the understanding of the document. In the same order we show examples which satisfy the definition conditions.

1. *bytem* – It was translated within one document into *an apartment* and *a residence*.

Document	Sentences	Direction	Markable occurrences	Description
Lease	29	cs→en en→cs	73 70	Housing lease agreement
Cars	18	cs→en	11	Brno Grand Prix competition article + highway accident report
Audit	90	cs→en en→cs	28 18	Supreme Audit Office audit report
Speech	13	en→cs	15	Greta Thunberg's U.N. speech article
Total	**269**	-	**215**	-

Table 1: Summary of examined documents with translation directions, number of lines and number of markable occurrences.

2. *rodné číslo* – It was translated in one translation to *social security number* and in another translation to *identification number*.
3. *nájemce*, *podnájemce* – They have different meanings and in one document they were both translated to tenant.

Markables were proposed first by the annotators in the first phase of annotation in Section 2.1 and then filtered manually by us.

1.2 Test Suite Composition

We selected 4 documents, 2 of which were translated in both directions totalling 6 documents. We chose 2 from the professional domain (Audit and Lease) and 2 from the News domain. The overview of their size is shown in Table 1. The number of markable occurrences is highly dependent on the document domain with the Agreement domain (Lease document) containing the most occurrences.

All of the MT systems are participants of the News Translation Task, and we test their performance even outside of this domain. Most of them were bi-directional, and we join the results from both directions when reporting their performance. The only exceptions are eTranslation (only en→cs) and PROMT_NMT (only cs→en).

1.3 Data and Tools Availability

All of the document translations and measured data are available in the project repository. Furthermore, the used online markable annotation tool written in TypeScript and Python is documented and also open-source.[2]

[2] github.com/ELITR/wmt20-elitr-testsuite

2 Annotation Setup

For both phases of this experiment, we used 10 native Czech annotators with English proficiency. None of them were professional audit or legal translators. Because each annotator annotated only one or two documents, the aggregated results across domains, labelled as *Total*, are of less significance than the results in individual domains.

2.1 Manual Document Evaluation

In this phase of the experiment, we wanted to measure the overall document translation quality and also to collect additional markables for use in the following experiment part. We showed the annotators the source document (in Czech) with a line highlighted and then underneath all its translation variants (in English). The current line was also highlighted. Next to every translation was a set of questions related to the just highlighted lines:

- **Adequacy**: range from 0 (worst) to 1 (best) measuring how much the translated message is content-wise correct regardless of grammatical and fluency errors.
- **Fluency**: range from 0 (worst) to 1 (best) measuring the fluency of the translation, regardless of the relation of the message to the source and the correct meaning.
- **Markables**: A text area for reporting markables for the second phase.
- **Conflicting markables**: checkbox for when there is a markable in conflict (e.g. the terminology change) with a previous occurrence in the document. This corresponds to the first condition in the markable definition in Section 1.1. The default value was *No* (no

conflict) because the distribution was highly imbalanced.

Bojar et al. (2016) summarize several methods for machine translation human evaluation: Fluency-Adequacy, Sentence Ranking, Sentence Comprehension, Direct Assessment, Constituent Rating and Constituent Judgement. For our purposes, we chose a method similar to Fluency-Adequacy as one of the standard sentence-centric methods. The difference to the method described is that we showed all the competing MT systems at once, together with the whole document context. Ultimately, we would like the users to rate Fluency-Adequacy of the whole documents, but we suspected that asking annotators to read the whole document and then rating it on two scales would yield unuseful biased results.

2.2 Manual Markable Evaluation

In the following phase, we focused on markables specifically. For every markable in the source, we asked the annotators to examine 11 phenomena. If the given phenomenon is present in the examined markable occurrence, a checkbox next to it should have been checked (Occurrence). Further on a scale 0–1 (not at all–most) the annotator should mark how negatively it affects the quality of the translation (Severity). We list the 11 phenomena we asked the annotators to work with:

- **Non-translated**: The markable or part of it was not translated.
- **Over-translated**: The markable was translated, but should not have been.
- **Terminology**: The translation terminology choice is terminologically misleading or erroneous.
- **Style**: An inappropriate translation style has been selected, such as too formal, colloquial, general.
- **Sense**: The meaning of the whole markable translation is different from what was intended by the source.
- **Typography**: Typographical errors in translation such as in capitalization, punctuation, special character or other typos.
- **Semantic role**: The markable has a different semantic role in translation than in the source. Without any specific linguistic theory in mind, we provided four basic roles for illustration: agent (story executor), patient (affected by the event), the addressee (recipient of the object in the event), effect (a consequence of the event).
- **Other grammar**: Other grammatical errors such as bad declension or ungrammatical form choice.
- **Inconsistency**: A different lexical translation option than the previous occurrence was used. It is enough to compare only with the previous occurrence and not with all of them.
- **Conflict**: The translation conflicts with another markable or term in the document. This and another markable translates to the same word.
- **Disappearance**: The markable does not appear in translation at all.

The choice to focus on markables was motivated by the aim to find a way to measure document-level performance using human annotators. A good markable translation is not a sufficient condition for document-level performance, but a necessary one. This approach is similar to Constituent Ranking/Judgement described by Bojar et al. (2016) with the difference that we chose to show all the markable occurrences in succession and in all translations in the same screen. We showed the whole translated documents context so that the annotators could refer to previous translations of the markable and the overall context.

2.3 Interface

Figure 1 shows the online interface for the second phase of this experiment. The first text area window contains the source document (e.g. in English). Below it are several translations (e.g. in Czech). Next to each translation is a set of questions. In the source, the current markable occurrence, to which the questions relate, is always displayed in dark blue. The current sentence is highlighted in the translations with light blue. The target words which probably correspond to the current markable (via automatic word alignment) are highlighted in dark blue as well. This alignment is present only for quick navigation as it is not very accurate. In translations, the remaining occurrences of a given markable are highlighted in green to simplify checking for inconsistency.

The FOCUS button is used to scroll to the current line in all text areas in case the user scrolled the view to examine the rest of the document.

In the first phase, the annotators could return to their previous answers and adjust them, but before

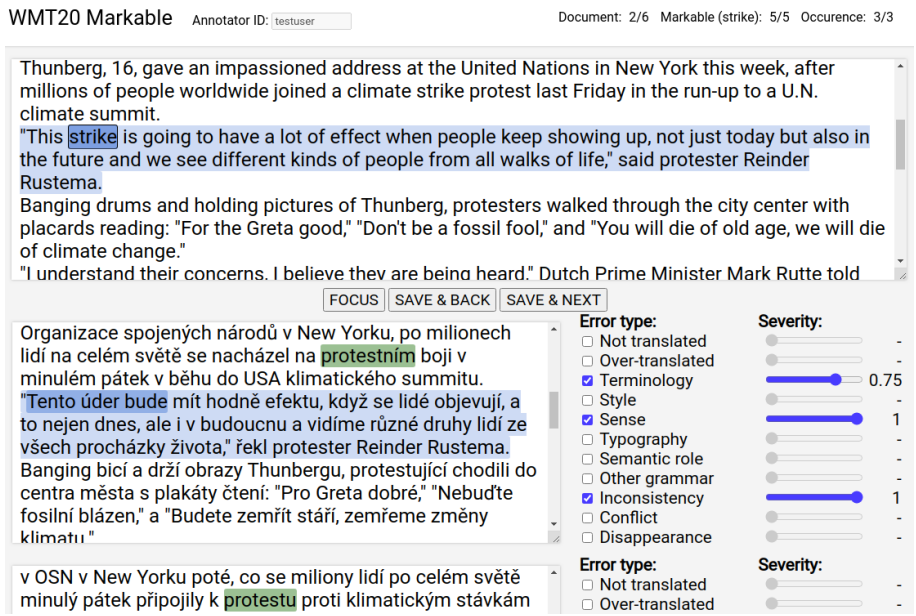


Figure 1: Online interface for markable annotation with highlighted segments. The 12 other translations are in the rest of the page, not fully visible here.

continuing to the next line, they had to fill in the current fluency and adequacy. In the second phase, the annotators could freely return to their previous answers and adjust them. The most straightforward approach for them was to annotate a single markable occurrence across all MT systems and the switch to the next one as opposed to annotating all markable occurrences in the first translation, then all markable occurrences in the second translation, and similarly the rest.

As soon as we aggregate the statistics over multiple documents (or even translation directions), the effects of which particular annotator annotated which document can start playing a role, but we hope they cancel out on average.

3 Results

3.1 Automatic Evaluation

We measured the system quality using BLEU (Papineni et al., 2002) against a single reference. The results sorted by the score across all documents are shown in Table 2. BLEU scores across different test sets are, of course, not comparable directly. Only a very big difference, such as that of eTranslation for News and Audit (39.43% and 23.23%) suggests some statistically sound phenomena. We measured the standard deviation across MT systems within individual domains: News (6.19), Audit (2.34) and News-Lease (2.74). The Audit domain was generally the least successful for most of the submitted systems (see Table 3) and the Lease domain was more stable in terms of variance. The MT system BLEU variance over annotated lines hints that the better the system, the higher variance it has. This may be because most of the best MT systems are focused on News and fail on other domains, while the lower performant MT systems are low performant systematically across all domains.

3.2 Overall Manual Evaluation

From the first phase (Section 2.1) we collected $13 \times 328 = 4264$ line annotations. From the second phase (Section 2.2) we collected $13 \times 499 = 6487$ markable annotations. The average duration for one annotation of one translated line in the first phase was 25s, while one annotation of one system-markable occurrence in the second phase took only 8s.

Fluency and Adequacy correlate per line together

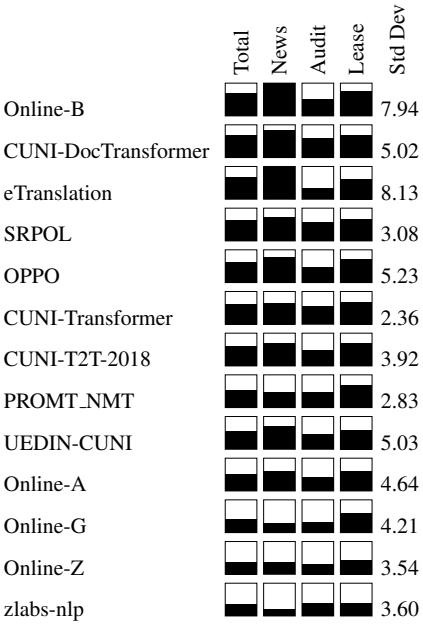


Table 2: MT system results measured by BLEU together with standard deviation measured from all sentences. Sorted by the first column. Full black box indicates 40% BLEU, empty 15% BLEU.

strongly (0.80), and their product correlates negatively (-0.33) with the number of wrong markables. Because of this strong correlation and also the need to describe the result of the first phase by one number, we focus on Fluency×Adequacy. Table 3 shows the average Fluency×Adequacy as well as the average number of reported wrong markables per line.

Document	Mult.	Mkbs.	BLEU
Audit →cs	0.95	0.08	28.61 ±5.13
Audit →en	0.81	1.23	32.68 ±5.07
Lease →cs	0.78	0.33	33.50 ±4.96
Lease →en	0.78	0.30	35.44 ±4.94
News →en	0.74	0.65	30.68 ±5.05
News →cs	0.65	0.83	38.67 ±4.93
Average	0.79	0.73	33.57 ±4.93

Table 3: Document average (across all systems) of Fluency×Adequacy (Mult.), number of reported wrong markables per line (Mkbs.) and BLEU.

3.3 MT System Performance

The performance per MT system and domain can be seen in Table 4. The reference translation received a comparably low rating in especially the Audit domain and fared best in the News domain. We see this as a confirmation of the last year's observation and a consequence of using non-expert annotators, who may have not annotated more complex cases thoroughly and were more content with rather general terms and language than what is correct for the specialized auditing domain.

No system has shown to be risky (high average but also with high variance). The last column in Table 4 shows, that the better the system, the more consistent it is (lower variation across documents). This did not occur with BLEU.

The ordering of systems by annotator assessment is slightly different than by automatic evaluation (Section 3.1). The automatic evaluation correlates with annotator rating (Fluency×Adequacy) with the coefficient of 0.93 (excluding Reference).

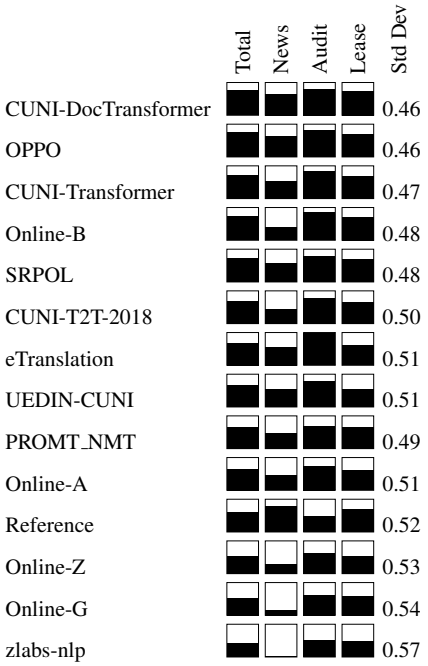


Table 4: MT system results measured by Fluency×Adequacy together with standard deviation measured from Total. Sorted by the first column. Full black box indicates 100%, empty 40%.

Notable is the distinction in the performance of eTranslation in the Audit domain. Its BLEU in this domain (23.23%, Table 2) was below average, however it performed best of all submitted MT systems in terms of Fluency×Adequacy (98.62%, Table 4), above Reference. Closer inspection revealed that the translations were very fluent and adequate but usually used vastly different phrasing than in the Reference, leading to very low BLEU scores.

Source:
In the vast majority of cases, the obligations arising from contracts for financing were properly implemented by the beneficiaries.
Reference:
Ve většině případů byly závazky vyplývající z podmínek podpory příjemci řádně plněny.
eTranslation: (**BLEU**: 9.24%)
Ve velké většině případů příjemci řádně plnili povinnosti vyplývající ze smluv o financování.
CUNI-DocTransformer: (**BLEU**: 41.21%)
V naprosté většině případů byly závazky vyplývající ze smluv o financování příjemci řádně plněny.

Figure 2: Example translations by eTranslation and CUNI-DocTransformer together with Source and Reference. N-grams present in Reference are underlined.

The example in Figure 2 shows activization (opposite of passivization) in the translation by eTranslation (*the beneficiaries fulfilled their obligations*) instead of (*obligations were fulfilled by the beneficiaries*). This resulted in much lower n-gram precision and BLEU score in general, even though the sentence is fluent and more adequate than both the Reference and translation by CUNI-DocTransformer.

3.4 Markable Phenomena and Systems

Table 5 shows an overview of types of markable phenomena with the average number of occurrences and Severity across systems. For all systems, *Terminology* and *Conflicting markables* had the most significant impact on the translation quality. These two categories clearly differ in Severity with markable conflicts being much more severe than terminological mistakes.

Inconsistency, *Typography* and *Disappearance* phenomena also heavily impacted the translation quality, although with varying distribution of Occurrences and Severity.

Reference differs from MT systems by having higher average Occurrence, but lower average Severity (first column in Table 5). Furthermore, the Reference had a higher number of *Inconsistence* occurrences, but with lower Severity. This means that most of these *Inconsitencies* were not actual errors. This is expected, as careful word choice variation improves the style and requires having an overview of previously used terms in the document.

Over-translation occurred rarely and in those cases, mostly in names (example shown in Figure 3). *Other grammar* manifested itself most severely in gender choice when translating sentences with person names without any gender indication from English to Czech. Similarly, *Style* was marked mostly in direct speech translation. The system used informal singular form addressing instead of plural. These two phenomena are shown in Figure 4.

Source & Reference: Karolína Černá
Translation: Caroline Black

Figure 3: Example of overly-translated named entity, it is the name of one of the parties in the Lease agreement.

Source:
“How dare you?” Thunberg’s U.N. speech inspires Dutch climate protesters
Reference:
“Jak se opovažujete?” projev Thunbergové v OSN inspiroval nizozemské protestující proti změnám klimatu
Translation:
“Jak se opovažuješ?” Thunbergův projev OSN inspiruje nizozemské klimatické demonstranty

Figure 4: Example of bad translation style.

Noteworthy is the correlation between phenomena across systems. The highest values were between *Sense* and *Terminology* (0.89), *Terminology* and *Inconsistency* (0.83) and *Sense* and *Other grammar* (0.82). There is no straightforward explanation of this correlation except the obvious that a good system is good across all phenomena. The correlation in the last phenomena pair suggests that the *Other grammar* category is too coarse and contains other subcategories.

3.5 Markable Phenomena and Domains

The results of markable phenomena across different domains is shown in Table 6.

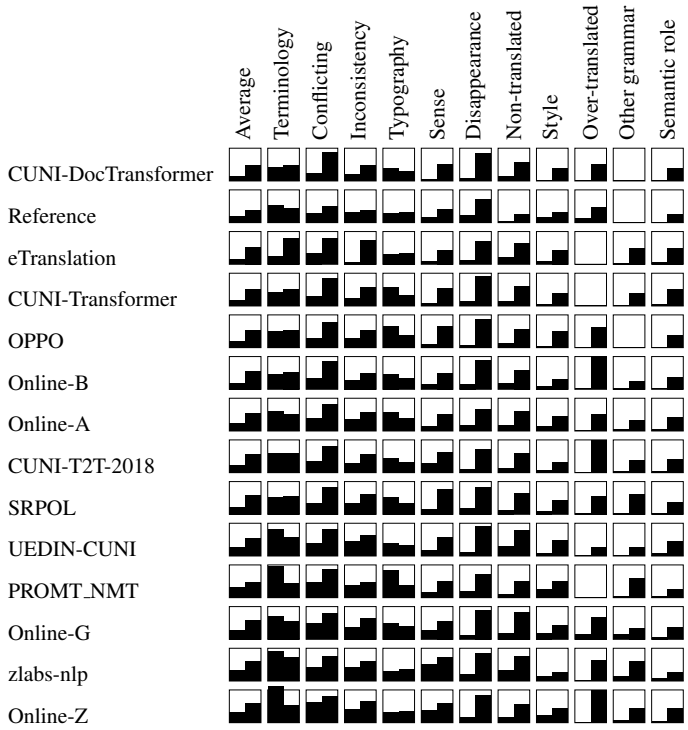


Table 5: Model results across 11 phenomena measured on markables together with their average. Each box is split into two bars: average Occurrence (left) and average Severity (right). Full left and right bars indicate occurrence in 20% of all markable instances and 100% Severity, respectively. Rows are sorted by Occurrence×Severity in the first column and columns, excluding *Average*, by the phenomena average Occurrence×Severity.

The second to last column is the correlation (across systems) between Occurrence×Severity and the BLEU score. The last column in Table 6 shows the correlation (across systems) between the two human scores: Occurrence×Severity and Fluency×Adequacy from the first phase of this experiment.

Since both BLEU and Fluency×Adequacy are positive metrics (the higher the score, the better the performance) and Occurrence×Severity is an error metric (the higher the number, the worse the performance), high negative correlations mean, that the metrics are mutually good performance predictors.

The strongest correlations are: *Conflicting* (-0.58), *Non-translated* (-0.55) and *Semantic role* (-0.41). Except for *Non-translated*, the reason is clear: BLEU is unable to check grammatical relations and never looks across sentences. We find the fact, that BLEU result was in agreement with error marking for these phenomena, to be positive.

Positive correlations (i.e. mismatches) were reached for *Disappearance* (0.28) and *Over-translated* (0.33), which is somewhat surprising because here BLEU *has* a chance to spot these errors from the technical point of view: shorter output could fire brevity penalty and missing terms where the exact wording is clear because they appear already in the source should decrease BLEU score. The overall correlation between Occurrence×Severity and Fluency×Adequacy is more significant than the correlation with BLEU. The most correlating variables are: *Sense* (-0.84), *Other grammar* (-0.84), *Terminology* (-0.81) and *Inconsistency* (-0.59).

Interesting is the markable phenomena *Disappearance* and *Sense* because of their high difference in correlations between BLEU and human score correlations.

	Total	News	Audit	Lease	BLEU corr.	Mult. corr.
Average					-0.45	-0.79
Terminology					-0.38	-0.81
Conflicting					-0.58	-0.45
Inconsistency					-0.36	-0.59
Typography					-0.31	0.25
Sense					-0.29	-0.84
Disappearance					0.28	-0.46
Non-translated					-0.55	-0.50
Style					-0.07	-0.44
Over-translated					0.33	-0.37
Other grammar					-0.37	-0.84
Semantic role					-0.41	-0.24

Table 6: Document domain average (across all systems) of markable phenomena. Sorted by Occurrence×Severity in the first column. Full left and right bars indicate occurrence in 20% of all markable instances and 100% Severity, respecively. The last two columns show correlation between Occurrence×Severity and BLEU and user ratings from Phase 1, respectively.

3.6 Annotator Agreement

We would like to bring attention to inter-annotator agreement for the second annotation phase. Table 7 lists the following metrics, which are computed pairwise and then averaged:

Plain inter-annotator agreement (IAA) reports the percentage of pairs of annotations where the two annotators agree that a given phenomenon was or was not present. IAA shows high numbers in all cases but it is skewed by the heavily imbalanced class distribution: most often, a phenomenon is not present; see the left sides of squares in the leftmost column in Table 6 for distribution reference.

Cohen's Kappa (Kappa), measured also pairwise, isolates the effect of agreeing by chance and reveals that a good agreement is actually reached only in the cases of *Disappearance*, *Terminology* and *Over-translated*, which are less ambiguous to annotate. It is unclear what is the reason behind the low Kappas, but we speculate that it is due to insufficient attention of the annotators: they would perhaps agree much more often that an error occurred but they were overloaded with the complexity of the annotation task and failed to notice on their own.

Phenomenon	IAA	Kappa	Corr.	Corr.+
Disappearance	0.90	0.43	0.52	0.06
Typography	0.95	0.20	0.55	-0.13
Sense	0.91	0.17	0.73	-0.09
Style	0.94	0.24	1.00	0.19
Terminology	0.90	0.41	0.07	-0.03
Inconsistency	0.88	0.13	0.18	-0.08
Non-translated	0.94	0.20	0.64	0.30
Conflicting	0.77	0.02	1.00	0.62
Other grammar	0.96	0.10	1.00	-0.35
Semantic role	0.97	-0.01	-	0.43
Over-translated	0.98	0.37	1.00	1.00

Table 7: Annotator agreement of Occurence marking (Inter Annotator Agreement and Cohen's Kappa) and agreement in Severity (two versions of Pearson Correlation) with respect to every markable phenomenon.

Plain Pearson Correlation (Corr.) was measured on Severities in instances where both annotators marked the phenomenon as present. This, however, disregards the disagreement in cases one annotator did not mark the phenomenon. For this, we also computed Corr.+, which examines all pairs in which at least one annotator reported Severity and replaces the other with zero.

We observe a big difference in the correlations. In cases where both annotators agreed that there was an error they tend to agree on the severity of the mistake, except *Terminology* and *Inconsistency*. If the cases where only one annotator marked the error are included, then the agreement on Severity is non-existent, except *Over-translation* and *Conflicting* translation.

3.7 Translation Direction

We also examined how the language translation directions affect the results. Most notable is CUNI-DocTransformer, which performs worse when translating into Czech. With only 0.01% higher Occurence of markable phenomena, the Severity increased by 20.81%. This is not something which we observed in other systems. The translation into Czech brought on average 0.01% higher Occurrence, but the Severity on average dropped by 3.99% when switching from English→Czech to Czech→English. The explanation supported by

the data is that in translation into English, CUNI-DocTransformer did not make any mistakes (or native Czech annotators did not detect them) and in translating into Czech, more issues were detected. Since the average Severity is measured across all phenomena, then the higher Severity in specific markable cases (Over-translated, Sense, Style and Disappearance) raised the overall average.

4 Annotation Examples

In the following figures (Figure 5, Figure 6 and Figure 7) we show annotated excerpts with BLEU, Fluency, Adequacy and markable phenomena severities. References are here to convey the Czech source segment meanings. They were not shown to the annotators. Examined markables are underlined.

Reference:
This Supplement No. 1 is written and signed in 2 (in words: two) copies, each of which is valid for the original.
Translation:
This Appendix 1 is drawn up and signed in two copies, each of which has the validity of the original.

BLEU: 23.59%, **Fluency:** 1, **Adequacy:** 0.9
Disappearance: 1

Figure 5: Example sentence markable (in words) annotation from Czech Lease document, translated by OPPO.

The example in Figure 5 focuses on intentional, key information duplication (for clarity and security reasons) of the number of signed copies. This duplication was however omitted in the translated output. The output is otherwise fluent and even received higher fluency than the Reference, which has an average fluency of 0.8.

Noteworthy is also another markable visible in the same figure, namely the referred section name: Appendix 1. Even though this word is different from the markable in the Reference: Supplement No. 1, it is used consistently across the whole document. Another variant of the translation is: Amendment No. 1. OPPO, together with Online-Z are the only systems which translated this markable correctly and consistently. Most of the systems (zlabs-nlp, Online-A, Online-B, Online-G, UEDIN-CUNI, CUNI-T2T-2018) switched inconsistently between the lexical choice. Other systems (SRPOL, eTranslation, CUNI-Transformer, CUNI-DocTransformer) were consistent in the main word choice, but not either in capitalization or number (e.g. Appendix No. 1 and Appendix 1).

Word variability (i.e. inconsistency) is often used to make the text more interesting, but in this context, it is vital that the term is translated consistently. Most of the systems, which outperformed even the Reference, made a severe error in this case.

Reference:
The most expensive item to be paid before the Grand Prix is the annual listing fee. This year, the fee was around 115 million Czech crowns. "Masses of people who come to Brno to see the Grand Prix spend money here for their accommodation, food and leisure activities, which should more or less balance out the cost associated with the organization of the event, including the listing fee," economist Petr Pelc evaluated the situation.
Translation:
The most expensive item is a breakdown fee every year before the Grand Prize. This year was about a hundred fifteen million crowns. "Mass of people who will come to Brno at the Grand Prix will spend money on accommodation, food or entertainment, which should more or less balance the costs associated with organizing the event, including the unifying fee," the economist Petr Pelc assessed.

BLEU: 26.59%, **Fluency:** 0.6, **Adequacy:** 0.4
Terminology: 1, **Sense:** 1, **Inconsistency:** 1

Figure 6: Example sentence markable (listing fee) annotation from Czech News document, translated by CUNI-T2T-2018.

Figure 6 shows a listing fee incorrectly translated as breakdown and unifying fee. This markable translation is interesting in the fact that systems were again very inconsistent with the markable translation choice. The wrong lexical choices were: landing, paving, parking, refill, landfill, security, zalistovacího, leasing, drop-in, back-up, reforestration, clearance, referral, padding fee and stamp duty. Good translations were: listing and registration fee.

Online-B and CUNI-DocTransformer made good and consistent lexical choices. SRPOL made good lexical choices but switched between them.

In this instance, this would not be an error, because consistency is not vital for interpreting the text.

The translation by CUNI-T2T-2018 in Figure 6 is not wrong only because of this markable translation choice, but also by poor fluency. The BLEU score, however, does not suggest, that there is anything fundamentally wrong with the translated segment despite the meaning being distorted.

Reference:
In Art. III of the Sublease agreement, entitled "Term of the Lease," the tenant and the lessee agreed that the apartment in question would be rented to the tenant for a fixed period from 13th May 2016 to 31st December 2018.
Translation:
In art. III of the apartment lease agreement, called "sublease period", the tenant and the tenant agreed that the apartment in question will be left to the tenant for use for a fixed period from 13. 5. 2016 to 31. 12. 2018.

BLEU: 31.95%, **Fluency:** 0.7, **Adequacy:** 0.5
Terminology: 0.5, **Sense:** 0.25, **Conflict:** 1,
Other grammar: 0.25

Figure 7: Example sentence markable (lessee) annotation from Czech News document, translated by Online-G.

The last example, in Figure 7, concerns itself with conflicting markables. In this case, two distinct markables (tenant and lessee) were merged into one translation tenant. This is a very fundamental error because, in the Lease agreement, these two markables refer to the two parties, which enter the contract.

Again, the BLEU does not suggest that anything is wrong with the translation. It could be even higher (51.06%) were it not for the localized date format in the Reference.

5 Conclusion

In this article, we compared three approaches to document translation evaluation. We saw that non-expert annotators rate most MT systems higher than Reference with Fluency and Adequacy, but Reference ranks better than most of them when inspecting markable phenomena and their Severity. Inspecting specific instances in detail, we found out that MT systems made errors in terms of markables, which no human translator would do.

Relating the current observation with the impression last year, we conclude that annotators lacking in-depth domain knowledge are not reliable for annotating on the rather broad scales of Fluency and Adequacy but they are capable of spotting term translation errors in the markable style of evaluation. This is important news because expert annotators can not be always secured. Unfortunately, the inter-annotator agreement remains generally low, possibly due to a high cognitive load with many systems annotated.

We further examined these markable phenomena and showed that especially *Sense*, *Other grammar* and *Terminology* kinds of errors negatively influence the Fluency and Adequacy the most. For BLEU the variables of highest importance were *Non-translated* and *Conflicting* errors.

In future work, we would like to examine more of the kinds of markable errors in modern MT systems and their influence on the translation quality. This description could then help researches focus on specific parts of their MT systems.

Furthermore, we would like to explore possible automated metrics, which would help in determining whether the document meaning remained intact with respect to markables.

Annotating markables appears to be easier for human annotators and more reliable for non-expert ones, and the results gave us more insight into the systems' performance than the Fluency-Adequacy method.

Acknowledgement

This study was supported in parts by the grants H2020-ICT-2018-2-825460 (ELITR) and Czech Science Foundation (grant n. 19-26934X, NEUREM3).

References

Ondřej Bojar, Christian Federmann, Barry Haddow, Philipp Koehn, Matt Post, and Lucia Specia. 2016. Ten years of WMT evaluation campaigns: Lessons learnt.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Tereza Vojtěchová, Michal Novák, Miloš Klouček, and Ondřej Bojar. 2019. SAO WMT19 test suite: Machine translation of audit reports. *arXiv preprint arXiv:1909.01701*.

B Improving Massively Multilingual Neural Machine Translation and Zero-Shot Translation

Improving Massively Multilingual Neural Machine Translation and Zero-Shot Translation

Biao Zhang[1] **Philip Williams**[1] **Ivan Titov**[1,2] **Rico Sennrich**[3,1]

[1]School of Informatics, University of Edinburgh
[2]ILLC, University of Amsterdam
[3]Department of Computational Linguistics, University of Zurich
B.Zhang@ed.ac.uk, {pwillia4,ititov}@inf.ed.ac.uk, sennrich@cl.uzh.ch

Abstract

Massively multilingual models for neural machine translation (NMT) are theoretically attractive, but often underperform bilingual models and deliver poor zero-shot translations. In this paper, we explore ways to improve them. We argue that multilingual NMT requires stronger modeling capacity to support language pairs with varying typological characteristics, and overcome this bottleneck via language-specific components and deepening NMT architectures. We identify the off-target translation issue (i.e. translating into a wrong target language) as the major source of the inferior zero-shot performance, and propose random online backtranslation to enforce the translation of unseen training language pairs. Experiments on OPUS-100 (a novel multilingual dataset with 100 languages) show that our approach substantially narrows the performance gap with bilingual models in both one-to-many and many-to-many settings, and improves zero-shot performance by ∼10 BLEU, approaching conventional pivot-based methods.[1]

1 Introduction

With the great success of neural machine translation (NMT) on bilingual datasets (Bahdanau et al., 2015; Vaswani et al., 2017; Barrault et al., 2019), there is renewed interest in multilingual translation where a single NMT model is optimized for the translation of multiple language pairs (Firat et al., 2016a; Johnson et al., 2017; Lu et al., 2018; Aharoni et al., 2019). Multilingual NMT eases model deployment and can encourage knowledge transfer among related language pairs (Lakew et al., 2018; Tan et al., 2019), improve low-resource translation (Ha et al., 2016; Arivazhagan et al., 2019b), and enable zero-shot translation (i.e. direct translation between a language pair never seen in training) (Firat et al., 2016b; Johnson et al., 2017; Al-Shedivat and Parikh, 2019; Gu et al., 2019).

Source	Jusqu'à ce qu'on trouve le moment clé, celui où tu as su que tu l'aimais.
Reference	Bis wir den unverkennbaren Moment gefunden haben, den Moment, wo du wusstest, du liebst ihn.
Zero-Shot	Jusqu'à ce qu'on trouve le moment clé, celui où tu as su que tu l'aimais.
Source	Les États membres ont été consultés et ont approuvé cette proposition.
Reference	Die Mitgliedstaaten wurden konsultiert und sprachen sich für diesen Vorschlag aus.
Zero-Shot	Les Member States have been consultedés and have approved this proposal.

Table 1: Illustration of the off-target translation issue with French→German zero-shot translations with a multilingual NMT model. Our baseline multilingual NMT model often translates into the wrong language for zero-shot language pairs, such as copying the source sentence or translating into English rather than German.

Despite these potential benefits, multilingual NMT tends to underperform its bilingual counterparts (Johnson et al., 2017; Arivazhagan et al., 2019b) and results in considerably worse translation performance when many languages are accommodated (Aharoni et al., 2019). Since multilingual NMT must distribute its modeling capacity between different translation directions, we ascribe this deteriorated performance to the deficient capacity of single NMT models and seek solutions that are capable of overcoming this capacity bottleneck. We propose language-aware layer normalization and linear transformation to relax the representation constraint in multilingual NMT models. The linear transformation is inserted in-between the encoder and the decoder so as to facilitate the induction of language-specific translation correspon-

[1]We release our code at https://github.com/bzhangGo/zero. We release the OPUS-100 dataset at https://github.com/EdinburghNLP/opus-100-corpus.

dences. We also investigate deep NMT architectures (Wang et al., 2019a; Zhang et al., 2019) aiming at further reducing the performance gap with bilingual methods.

Another pitfall of massively multilingual NMT is its poor zero-shot performance, particularly compared to pivot-based models. Without access to parallel training data for zero-shot language pairs, multilingual models easily fall into the trap of *off-target translation* where a model ignores the given target information and translates into a wrong language as shown in Table 1. To avoid such a trap, we propose the random online backtranslation (ROBT) algorithm. ROBT finetunes a pretrained multilingual NMT model for unseen training language pairs with pseudo parallel batches generated by back-translating the target-side training data.[2] We perform backtranslation (Sennrich et al., 2016a) into randomly picked intermediate languages to ensure good coverage of ∼10,000 zero-shot directions. Although backtranslation has been successfully applied to zero-shot translation (Firat et al., 2016b; Gu et al., 2019; Lakew et al., 2019), whether it works in the massively multilingual set-up remained an open question and we investigate it in our work.

For experiments, we collect OPUS-100, a massively multilingual dataset sampled from OPUS (Tiedemann, 2012). OPUS-100 consists of 55M English-centric sentence pairs covering 100 languages. As far as we know, no similar dataset is publicly available.[3] We have released OPUS-100 to facilitate future research.[4] We adopt the Transformer model (Vaswani et al., 2017) and evaluate our approach under one-to-many and many-to-many translation settings. Our main findings are summarized as follows:

- Increasing the capacity of multilingual NMT yields large improvements and narrows the performance gap with bilingual models. Low-resource translation benefits more from the increased capacity.
- Language-specific modeling and deep NMT architectures can slightly improve zero-shot translation, but fail to alleviate the off-target translation issue.
- Finetuning multilingual NMT with ROBT substantially reduces the proportion of off-target translations (by ∼50%) and delivers an improvement of ∼10 BLEU in zero-shot settings, approaching the conventional pivot-based method. We show that finetuning with ROBT converges within a few thousand steps.

[2]Note that backtranslation actually converts the zero-shot problem into a zero-resource problem. We follow previous work and continue referring to *zero-shot* translation, even when using synthetic training data.

[3]Previous studies (Aharoni et al., 2019; Arivazhagan et al., 2019b) adopt in-house data which was not released.

[4]https://github.com/EdinburghNLP/opus-100-corpus

2 Related Work

Pioneering work on multilingual NMT began with multitask learning, which shared the encoder for one-to-many translation (Dong et al., 2015) or the attention mechanism for many-to-many translation (Firat et al., 2016a). These methods required a dedicated encoder or decoder for each language, limiting their scalability. By contrast, Lee et al. (2017) exploited character-level inputs and adopted a shared encoder for many-to-one translation. Ha et al. (2016) and Johnson et al. (2017) further successfully trained a single NMT model for multilingual translation with a target language symbol guiding the translation direction. This approach serves as our baseline. Still, this paradigm forces different languages into one joint representation space, neglecting their linguistic diversity. Several subsequent studies have explored different strategies to mitigate this representation bottleneck, ranging from reorganizing parameter sharing (Blackwood et al., 2018; Sachan and Neubig, 2018; Lu et al., 2018; Wang et al., 2019c; Vázquez et al., 2019), designing language-specific parameter generators (Platanios et al., 2018), decoupling multilingual word encodings (Wang et al., 2019b) to language clustering (Tan et al., 2019). Our language-specific modeling continues in this direction, but with a special focus on broadening normalization layers and encoder outputs.

Multilingual NMT allows us to perform zero-shot translation, although the quality is not guaranteed (Firat et al., 2016b; Johnson et al., 2017). We observe that multilingual NMT often translates into the wrong target language on zero-shot directions (Table 1), resonating with the 'missing ingredient problem' (Arivazhagan et al., 2019a) and the spurious correlation issue (Gu et al., 2019). Approaches to improve zero-shot performance fall into two categories: 1) developing novel cross-lingual regularizers, such as the alignment regularizer (Arivazhagan et al., 2019a) and the consistency regularizer (Al-

Shedivat and Parikh, 2019); and 2) generating artificial parallel data with backtranslation (Firat et al., 2016b; Gu et al., 2019; Lakew et al., 2019) or pivot-based translation (Currey and Heafield, 2019). The proposed ROBT algorithm belongs to the second category. In contrast to Gu et al. (2019) and Lakew et al. (2019), however, we perform online back-translation for each training step with randomly selected intermediate languages. ROBT avoids decoding the whole training set for each zero-shot language pair and can therefore scale to massively multilingual settings.

Our work belongs to a line of research on massively multilingual translation (Aharoni et al., 2019; Arivazhagan et al., 2019b). Aharoni et al. (2019) demonstrated the feasibility of massively multilingual NMT and reported encouraging results. We continue in this direction by developing approaches that improve both multilingual and zero-shot performance. Independently from our work, Arivazhagan et al. (2019b) also find that increasing model capacity with deep architectures (Wang et al., 2019a; Zhang et al., 2019) substantially improves multilingual performance. A concurrent related work is (Bapna and Firat, 2019), which introduces task-specific and lightweight adaptors for fast and scalable model adaptation. Compared to these adaptors, our language-aware layers are jointly trained with the whole NMT model from scratch without relying on any pretraining.

3 Multilingual NMT

We briefly review the multilingual approach (Ha et al., 2016; Johnson et al., 2017) and the Transformer model (Vaswani et al., 2017), which are used as our baseline. Johnson et al. (2017) rely on prepending tokens specifying the target language to each source sentence. In that way a single NMT model can be trained on the modified multilingual dataset and used to perform multilingual translation. Given a source sentence $\mathbf{x}=(x_1, x_2, \ldots, x_{|\mathbf{x}|})$, its target reference $\mathbf{y}=(y_1, y_2, \ldots, y_{|\mathbf{y}|})$ and the target language token t[5], multilingual NMT translates under the encoder-decoder framework (Bahdanau et al., 2015):

$$\mathbf{H} = \text{Encoder}([t, \mathbf{x}]), \quad (1)$$

$$\mathbf{S} = \text{Decoder}(\mathbf{y}, \mathbf{H}), \quad (2)$$

[5] t is in the form of "<2X>" where X is a language name, such as <2EN> meaning *translating into English*.

where $\mathbf{H} \in \mathbb{R}^{|\mathbf{x}|\times d}/\mathbf{S} \in \mathbb{R}^{|\mathbf{y}|\times d}$ denote the encoder/decoder output. d is the model dimension.

We employ the Transformer (Vaswani et al., 2017) as the backbone NMT model due to its superior multilingual performance (Lakew et al., 2018). The encoder is a stack of $L = 6$ identical layers, each containing a self-attention sublayer and a point-wise feedforward sublayer. The decoder follows a similar structure, except for an extra cross-attention sublayer used to condition the decoder on the source sentence. Each sublayer is equipped with a residual connection (He et al., 2015), followed by layer normalization (Ba et al., 2016, LN$(\cdot)$):

$$\bar{\mathbf{a}} = \text{LN}(\mathbf{a} \mid \mathbf{g}, \mathbf{b}) = \frac{\mathbf{a} - \mu}{\sigma} \odot \mathbf{g} + \mathbf{b}, \quad (3)$$

where $\odot$ denotes element-wise multiplication, μ and σ are the mean and standard deviation of the input vector $\mathbf{a} \in \mathbb{R}^d$, respectively. $\mathbf{g} \in \mathbb{R}^d$ and $\mathbf{b} \in \mathbb{R}^d$ are model parameters. They control the sharpness and location of the regularized layer output $\bar{\mathbf{a}}$. Layer normalization has proven effective in accelerating model convergence (Ba et al., 2016).

4 Approach

Despite its success, multilingual NMT still suffers from 1) *insufficient modeling capacity*, where including more languages results in reduction in translation quality (Aharoni et al., 2019); and 2) *off-target translation*, where models translate into a wrong target language on zero-shot directions (Arivazhagan et al., 2019a). These drawbacks become severe in massively multilingual settings and we explore approaches to alleviate them. We hypothesize that the vanilla Transformer has insufficient capacity and search for model-level strategies such as deepening Transformer and devising language-specific components. By contrast, we regard the lack of parallel data as the reason behind the off-target issue. We resort to data-level strategy by creating, in online fashion, artificial parallel training data for each zero-shot language pair in order to encourage its translation.

Deep Transformer One natural way to improve the capacity is to increase model depth. Deeper neural models are often capable of inducing more generalizable ('abstract') representations and capturing more complex dependencies and have shown encouraging performance on bilingual translation (Bapna et al., 2018; Zhang et al., 2019; Wang

et al., 2019a). We adopt the depth-scaled initialization method (Zhang et al., 2019) to train a deep Transformer for multilingual translation.

Language-aware Layer Normalization Regardless of linguistic differences, layer normalization in multilingual NMT simply constrains all languages into one joint Gaussian space, which makes learning more difficult. We propose to relax this restriction by conditioning the normalization on the given target language token t (LALN for short) as follows:

$$\bar{\mathbf{a}} = \text{LN}(\mathbf{a} \mid \mathbf{g}_t, \mathbf{b}_t). \quad (4)$$

We apply this formula to all normalization layers, and leave the study of conditioning on source language information for the future.

Language-aware Linear Transformation Different language pairs have different translation correspondences or word alignments (Koehn, 2010). In addition to LALN, we introduce a target-language-aware linear transformation (LALT for short) between the encoder and the decoder to enhance the freedom of multilingual NMT in expressing flexible translation relationships. We adapt Eq. (2) as follows:

$$\mathbf{S} = \text{Decoder}(\mathbf{y}, \mathbf{H}\mathbf{W}_t), \quad (5)$$

where $\mathbf{W}_t \in \mathbb{R}^{d \times d}$ denotes model parameters. Note that adding one more target language in LALT brings in only one weight matrix.[6] Compared to existing work (Firat et al., 2016b; Sachan and Neubig, 2018), LALT reaches a better trade-off between expressivity and scalability.

Random Online Backtranslation Prior studies on backtranslation for zero-shot translation decode the whole training set for each zero-shot language pair (Gu et al., 2019; Lakew et al., 2019), and scalability to massively multilingual translation is questionable – in our setting, the number of zero-shot translation directions is 9702.

We address scalability by performing online backtranslation paired with randomly sampled intermediate languages. Algorithm 1 shows the detail of ROBT, where for each training instance $(\mathbf{x}_k, \mathbf{y}_k, t_k)$, we uniformly sample an intermediate language t'_k ($t_k \neq t'_k$), back-translate $\mathbf{y}_k$ into t'_k to obtain $\mathbf{x}'_k$, and train on the new instance $(\mathbf{x}'_k, \mathbf{y}_k, t_k)$. Although $\mathbf{x}'_k$ may be poor initially (translations are produced on-line by the model being trained), ROBT still benefits from the translation signal of $t'_k \rightarrow t_k$. To reduce the computational cost, we implement batch-based greedy decoding for line 7.

[6] We also attempted to factorize $\mathbf{W}_t$ into smaller matrices/vectors to reduce the number of parameters. Unfortunately, the final performance was rather disappointing.

Algorithm 1: Algorithm for Random Online Backtranslation

Input : Multilingual training data, D;
Pretrained multilingual model, M;
Maximum finetuning step, N;
Finetuning batch size, B;
Target language set, $\mathcal{T}$;
Output: Zero-shot enabled model, M

1 $i \leftarrow 0$
2 **while** $i \leq N \wedge$ *not converged* **do**
3 $\mathcal{B} \leftarrow$ sample batch from D
4 **for** $k \leftarrow 1$ ***to*** B **do**
5 $(\mathbf{x}_k, \mathbf{y}_k, t_k) \leftarrow \mathcal{B}_k$
6 $t'_k \sim$ Uniform($\mathcal{T}$) such that $t'_k \neq t_k$
7 $\mathbf{x}'_k \leftarrow M([t'_k, \mathbf{y}_k])$
 `// backtrans` $t_k \rightarrow t'_k$ `to produce training example for` $t'_k \rightarrow t_k$
8 $\mathcal{B} \leftarrow \mathcal{B} \cup (\mathbf{x}'_k, \mathbf{y}_k, t_k)$
9 Optimize M using $\mathcal{B}$
10 $i \leftarrow i + 1$
11 **return** M

5 OPUS-100

Recent work has scaled up multilingual NMT from a handful of languages to tens or hundreds, with many-to-many systems being capable of translation in thousands of directions. Following Aharoni et al. (2019), we created an English-centric dataset, meaning that all training pairs include English on either the source or target side. Translation for any language pair that does not include English is zero-shot or must be pivoted through English.

We created OPUS-100 by sampling data from the OPUS collection (Tiedemann, 2012). OPUS-100 is at a similar scale to Aharoni et al. (2019)'s, with 100 languages (including English) on both sides and up to 1M training pairs for each language pair. We selected the languages based on the volume of parallel data available in OPUS.

The OPUS collection is comprised of multiple corpora, ranging from movie subtitles to GNOME

ID	Model Architecture	L	#Param	BLEU_{94}	WR	BLEU_4
1	Transformer, Bilingual	6	106M	-	-	20.90
2	Transformer, Bilingual	12	150M	-	-	**22.75**
3	Transformer	6	106M	24.64	*ref*	18.95
4	3 + MATT	6	99M	23.81	20.2	17.95
5	4 + LALN	6	102M	24.22	28.7	18.50
6	4 + LALT	6	126M	27.11	72.3	20.28
7	4 + LALN + LALT	6	129M	27.18	75.5	20.08
8	4	12	137M	25.69	81.9	19.13
9	7	12	169M	28.04	91.5	19.93
10	7	24	249M	**29.60**	**92.6**	21.23

Table 2: Test BLEU for one-to-many translation on OPUS-100 (100 languages). "*Bilingual*": bilingual NMT, "L": model depth (for both encoder and decoder), "*#Param*": parameter number, "*WR*": win ratio (%) compared to *ref* (③), MATT: the merged attention (Zhang et al., 2019). LALN and LALT denote the proposed language-aware layer normalization and linear transformation, respectively. "$BLEU_{94}$/$BLEU_4$": average BLEU over all 94 translation directions in test set and En→De/Zh/Br/Te, respectively. Higher BLEU and WR indicate better result. Best scores are highlighted in **bold**.

documentation to the Bible. We did not curate the data or attempt to balance the representation of different domains, instead opting for the simplest approach of downloading all corpora for each language pair and concatenating them. We randomly sampled up to 1M sentence pairs per language pair for training, as well as 2000 for validation and 2000 for testing.[7] To ensure that there was no overlap (at the monolingual sentence level) between the training and validation/test data, we applied a filter during sampling to exclude sentences that had already been sampled. Note that this was done cross-lingually, so an English sentence in the Portuguese-English portion of the training data could not occur in the Hindi-English test set, for instance.

OPUS-100 contains approximately 55M sentence pairs. Of the 99 language pairs, 44 have 1M sentence pairs of training data, 73 have at least 100k, and 95 have at least 10k.

To evaluate zero-shot translation, we also sampled 2000 sentence pairs of test data for each of the 15 pairings of Arabic, Chinese, Dutch, French, German, and Russian. Filtering was used to exclude sentences already in OPUS-100.

6 Experiments

6.1 Setup

We perform one-to-many (English-X) and many-to-many (English-X ∪ X-English) translation on OPUS-100 ($|\mathcal{T}|$ is 100). We apply byte pair encoding (BPE) (Sennrich et al., 2016b; Kudo and Richardson, 2018) to handle multilingual words with a joint vocabulary size of 64k. We randomly shuffle the training set to mix instances of different language pairs. We adopt BLEU (Papineni et al., 2002) for translation evaluation with the toolkit SacreBLEU (Post, 2018)[8]. We employ the *langdetect* library[9] to detect the language of translations, and measure the translation-language accuracy for zero-shot cases. Rather than providing numbers for each language pair, we report average BLEU over all 94 language pairs with test sets (BLEU_{94}). We also show the win ratio (WR), counting the proportion where our approach outperforms its baseline.

Apart from multilingual NMT, our baselines also involve bilingual NMT and pivot-based translation (only for zero-shot comparison). We select four typologically different target languages (German/De, Chinese/Zh, Breton/Br, Telugu/Te) with varied training data size for comparison to bilingual models as applying bilingual NMT to each language pair is resource-consuming. We report average BLEU over these four languages as BLEU_4. We reuse the multilingual BPE vocabulary for bilingual NMT.

We train all NMT models with the Transformer base settings (512/2048, 8 heads) (Vaswani et al., 2017). We pair our approaches with the merged attention (MATT) (Zhang et al., 2019) to reduce training time. Other details about model settings are in the Appendix.

6.2 Results on One-to-Many Translation

Table 2 summarizes the results. The inferior performance of multilingual NMT (③) against its

[7] For efficiency, we only use 200 sentences per language pair for validation in our multilingual experiments.

[8] Signature: BLEU+case.mixed+numrefs.1+smooth.exp+tok.13a+version.1.4.1

[9] https://github.com/Mimino666/langdetect

ID	Model Architecture	L	#Param	w/o ROBT			w/ ROBT		
				$BLEU_{94}$	WR	$BLEU_4$	$BLEU_{94}$	WR	$BLEU_4$
1	Transformer, Bilingual	6	110M	-	-	**20.28**	-	-	-
2	Transformer	6	110M	19.50	*ref*	15.35	18.75	4.3	14.73
3	2 + MATT	6	103M	18.49	5.3	14.90	17.85	6.4	14.38
4	3 + LALN + LALT	6	133M	21.39	78.7	18.13	20.81	69.1	17.45
5	3	12	141M	20.77	94.7	16.08	20.24	84.0	15.80
6	4	12	173M	22.86	97.9	19.25	22.39	97.9	18.23
7	4	24	254M	**23.96**	**100.0**	19.83	23.36	97.9	19.45

Table 3: English→X test BLEU for many-to-many translation on OPUS-100 (100 languages). "*WR*": win ratio (%) compared to *ref* (② w/o ROBT). ROBT denotes the proposed random online backtranslation method.

ID	Model Architecture	L	#Param	w/o ROBT			w/ ROBT		
				$BLEU_{94}$	WR	$BLEU_4$	$BLEU_{94}$	WR	$BLEU_4$
1	Transformer, Bilingual	6	110M	-	-	21.23	-	-	-
2	Transformer	6	110M	27.60	*ref*	23.35	27.02	14.9	22.50
3	2 + MATT	6	103M	26.90	2.1	22.78	26.28	4.3	21.53
4	3 + LALN + LALT	6	133M	27.50	37.2	23.05	27.22	23.4	23.30
5	3	12	141M	29.15	**98.9**	24.15	28.80	91.5	24.03
6	4	12	173M	29.49	97.9	24.53	29.54	96.8	25.43
7	4	24	254M	**31.36**	**98.9**	26.03	30.98	95.7	**26.78**

Table 4: X→English test BLEU for many-to-many translation on OPUS-100 (100 languages). "*WR*": win ratio (%) compared to *ref* (② w/o ROBT).

bilingual counterpart (①) reflects the capacity issue (-1.95 $BLEU_4$). Replacing the self-attention with MATT slightly deteriorates performance (-0.83 $BLEU_{94}$ ③→④); we still use MATT for more efficiently training deep models.

Our ablation study (④-⑦) shows that enriching the language awareness in multilingual NMT substantially alleviates this capacity problem. Relaxing the normalization constraints with LALN gains 0.41 $BLEU_{94}$ with 8.5% WR (④→⑤). Decoupling different translation relationships with LALT delivers an improvement of 3.30 $BLEU_{94}$ and 52.1% WR (④→⑥). Combining LALT and LALN demonstrates their complementarity (+3.37 $BLEU_{94}$ and +55.3% WR, ④→⑦), significantly outperforming the multilingual baseline (+2.54 $BLEU_{94}$, ③→⑦), albeit still behind the bilingual models (-0.82 $BLEU_4$, ①→⑦).

Deepening the Transformer also improves the modeling capacity (+1.88 $BLEU_{94}$, ④→⑧). Although deep Transformer performs worse than LALN+LALT under a similar number of model parameters in terms of BLEU (-1.49 $BLEU_{94}$, ⑦→⑧), it shows more consistent improvements across different language pairs (+6.4% WR). We obtain better performance when integrating all approaches (⑨). By increasing the model depth to 24 (⑩), Transformer with our approach yields a score of 29.60 $BLEU_{94}$ and 21.23 $BLEU_4$, beating the baseline (③) on 92.6% tasks and outperforming the base bilingual model (①) by 0.33 $BLEU_4$. Our approach significantly narrows the performance gap between multilingual NMT and bilingual NMT (20.90 $BLEU_4$ → 21.23 $BLEU_4$, ①→⑩), although similarly deepening bilingual models surpasses our approach by 1.52 $BLEU_4$ (⑩→②).

6.3 Results on Many-to-Many Translation

We train many-to-many NMT models on the concatenation of the one-to-many dataset (English→X) and its reversed version (X→English), and evaluate the zero-shot performance on X→X language pairs. Table 3 and Table 4 show the translation results for English→X and X→English, respectively.[10] We focus on the translation performance w/o ROBT in this subsection.

Compared to the one-to-many translation, the many-to-many translation must accommodate twice as many translation directions. We observe that many-to-many NMT models suffer more se-

[10]Note that the one-to-many training and test sets were not yet aggressively filtered for sentence overlap as described in Section 5, so results in Table 2 and Table 3 are not directly comparable.

ID	Model Architecture	L	#Param	English→X			X→English		
				High	Med	Low	High	Med	Low
1	Transformer	6	110M	20.69	20.82	15.18	26.99	28.60	27.49
2	1 + MATT	6	103M	19.70	19.77	14.17	26.32	27.81	26.84
3	2 + LALN + LALT	6	133M	21.07	22.88	19.99	27.03	28.60	26.97
4	2	12	141M	21.67	22.17	16.95	28.39	30.24	29.26
5	3	12	173M	22.48	24.38	21.58	28.66	30.73	29.50
6	3	24	254M	**23.69**	**25.61**	**22.24**	**30.29**	**32.58**	**31.90**

Table 5: Test BLEU for High/Medium/Low (*High/Med/Low*) resource language pairs in many-to-many setting on OPUS-100 (100 languages). We report average BLEU for each category.

ID	Model Architecture	L	#Param	w/o ROBT		w/ ROBT	
				BLEU_{zero}	ACC_{zero}	BLEU_{zero}	ACC_{zero}
1	Transformer, Pivot & Bilingual	6	110M	12.98	84.87	-	-
2	Transformer	6	110M	3.97	36.04	10.11	86.08
3	2 + MATT	6	103M	3.49	31.62	9.67	85.87
4	3 + LALN + LALT	6	133M	4.02	45.43	11.23	87.40
5	3	12	141M	4.71	39.40	11.87	87.44
6	4	12	173M	5.41	51.40	12.62	**87.99**
7	4	24	254M	5.24	47.91	14.08	87.68
8	7 + Pivot	24	254M	14.71	84.81	**14.78**	85.09

Table 6: Test BLEU and translation-language accuracy for zero-shot translation in many-to-many setting on OPUS-100 (100 languages). "$BLEU_{zero}$/ACC_{zero}": average BLEU/accuracy over all zero-shot translation directions in test set, "*Pivot*": the pivot-based translation that first translates one source sentence into English (X→English NMT), and then into the target language (English→X NMT). Lower accuracy indicates severe off-target translation. The average Pearson correlation coefficient between language accuracy and the corresponding BLEU is 0.93 (significant at $p < 0.01$).

rious capacity issues on English→X tasks (-4.93 BLEU_4, ①→② in Table 3 versus -1.95 BLEU_4 in Table 2), where the deep Transformer with LALN + LALT effectively reduces this gap to -0.45 BLEU_4 (①→⑦, Table 3), resonating with our findings from Table 2. By contrast, multilingual NMT benefits X→English tasks considerably from the multitask learning alone, outperforming bilingual NMT by 2.13 BLEU_4 (①→②, Table 4). Enhancing model capacity further enlarges this margin to +4.80 BLEU_4 (①→⑦, Table 4).

We find that the overall quality of English→X translation (19.50/23.96 BLEU_{94}, ②/⑦, Table 3) lags far behind that of its X→English counterpart (27.60/31.36 BLEU_{94}, ②/⑫, Table 4), regardless of the modeling capacity. We ascribe this to the highly skewed training data distribution, where half of the training set uses English as the target. This strengthens the ability of the decoder to translate into English, and also encourages knowledge transfer for X→English language pairs. LALN and LALT show the largest benefit for English→X (+2.9 BLEU_{94}, ③→④, Table 3), and only a small benefit for X→English (+0.6 BLEU_{94}, ③→④, Table 4). This makes sense considering that LALN and LALT are specific to the target language, so capacity is mainly increased for English→X. Deepening the Transformer yields benefits in both directions (+2.57 BLEU_{94} for English→X, +3.86 BLEU_{94} for X→English; ④→⑦, Tables 3 and 4).

6.4 Effect of Training Corpus Size

Our multilingual training data is distributed unevenly across different language pairs, which could affect the knowledge transfer delivered by language-aware modeling and deep Transformer in multilingual translation. We investigate this effect by grouping different language pairs in OPUS-100 into three categories according to their training data size: High ($\geq$ 0.9M, 45), Low ($<$ 0.1M, 18) and Medium (others, 31). Table 5 shows the results.

Language-aware modeling benefits low-resource language pairs the most on English→X translation (+5.82 BLEU, Low versus +1.37/+3.11 BLEU, High/Med, ②→③), but has marginal impact on X→English translation as analyzed in Section 6.3. By contrast, deep Transformers yield similar benefits across different data scales (+2.38 average BLEU, English→X and +2.31 average BLEU, X→English, ②→④). We obtain the best perfor-

mance by integrating both (①→⑥) with a clear positive transfer to low-resource language pairs.

6.5 Results on Zero-Shot Translation

Previous work shows that a well-trained multilingual model can do zero-shot X→Y translation directly (Firat et al., 2016b; Johnson et al., 2017). Our results in Table 6 reveal that the translation quality is rather poor (3.97 BLEU$_{zero}$, ② w/o ROBT) compared to the pivot-based bilingual baseline (12.98 BLEU$_{zero}$, ①) under the massively multilingual setting (Aharoni et al., 2019), although translations into different target languages show varied performance. The marginal gain by the deep Transformer with LALN + LALT (+1.44 BLEU$_{zero}$, ②→⑥, w/o ROBT) suggests that weak model capacity is not the major cause of this inferior performance.

In a manual analysis on the zero-shot NMT outputs, we found many instances of off-target translation (Table 1). We use translation-language accuracy to measure the proportion of translations that are in the correct target language. Results in Table 6 show that there is a huge accuracy gap between the multilingual and the pivot-based method (-48.83% ACC$_{zero}$, ①→②, w/o ROBT), from which we conclude that the off-target translation issue is one source of the poor zero-shot performance.

We apply ROBT to multilingual models by fine-tuning them for an extra 100k steps with the same batch size as for training. Table 6 shows that ROBT substantially improves ACC$_{zero}$ by 35%∼50%, reaching 85%∼87% under different model settings. The multilingual Transformer with ROBT achieves a translation improvement of up to 10.11 BLEU$_{zero}$ (② w/o ROBT→⑦ w/ ROBT), outperforming the bilingual baseline by 1.1 BLEU$_{zero}$ (① w/o ROBT→⑦ w/ ROBT) and approaching the pivot-based multilingual baseline (-0.63 BLEU$_{zero}$, ⑧ w/o ROBT→⑦ w/ ROBT).[11] The strong Pearson correlation between the accuracy and BLEU (0.92 on average, significant at $p < 0.01$) suggests that the improvement on the off-target translation issue explains the increased translation performance to a large extent.

Results in Table 3 and 4 show that ROBT's success on zero-shot translation comes at the cost of sacrificing ∼0.50 BLEU$_{94}$ and ∼4% WR on English→X and X→English translation. We also note that models with more capacity yield higher language accuracy (+7.78%/+13.81% ACC$_{zero}$, ③→⑤/③→④, w/o ROBT) and deliver better zero-shot performance before (+1.22/+0.53 BLEU$_{zero}$, ③→⑤/③→④, w/o ROBT) and after ROBT (+2.20/+1.56 BLEU$_{zero}$, ③→⑤/③→④, w/ ROBT). In other words, increasing the modeling capacity benefits zero-shot translation and improves robustness.

[11] Note that ROBT improves all zero-shot directions due to its randomness in sampling the intermediate languages. We do not bias ROBT to the given zero-shot test set.

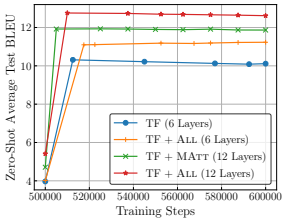


Figure 1: Zero-shot average test BLEU for multilingual NMT models finetuned by ROBT. ALL = MATT + LALN + LALT. Multilingual models with ROBT quickly converge on zero-shot directions.

Setting	BLEU$_{zero}$
6-to-6	11.98
100-to-100	11.23

Table 7: Zero-short translation quality for ROBT under different settings. "*100-to-100*": the setting used in the above experiments; we set $\mathcal{T}$ to all target languages. "*6-to-6*": $\mathcal{T}$ only includes the zero-shot languages in the test set. We employ 6-layer Transformer with LALN and LALT for experiments.

Convergence of ROBT. Unlike prior studies (Gu et al., 2019; Lakew et al., 2019), we resort to an online method for backtranslation. The curve in Figure 1 shows that ROBT is very effective, and takes only a few thousand steps to converge, suggesting that it is unnecessary to decode the whole training set for each zero-shot language pair. We leave it to future work to explore whether different back-translation strategies (other than greedy decoding) will deliver larger and continued benefits with ROBT.

Impact of $\mathcal{T}$ on ROBT. ROBT heavily relies on $\mathcal{T}$, the set of target languages considered, to distribute the modeling capacity on zero-shot directions. To study its impact, we provide a comparison by constraining $\mathcal{T}$ to 6 languages in the zero-shot test set. Results in Table 7 show that the biased ROBT outperforms the baseline by 0.75 BLEU$_{zero}$. By narrowing $\mathcal{T}$, more capacity is scheduled to the focused languages, which results in performance improvements. But the small scale of this improve-

ment suggests that the number of zero-shot directions is not ROBT's biggest bottleneck.

7 Conclusion and Future Work

This paper explores approaches to improve massively multilingual NMT, especially on zero-shot translation. We show that multilingual NMT suffers from weak capacity, and propose to enhance it by deepening the Transformer and devising language-aware neural models. We find that multilingual NMT often generates off-target translations on zero-shot directions, and propose to correct it with a random online backtranslation algorithm. We empirically demonstrate the feasibility of backtranslation in massively multilingual settings to allow for massively zero-shot translation for the first time. We release OPUS-100, a multilingual dataset from OPUS including 100 languages with around 55M sentence pairs for future study. Our experiments on this dataset show that the proposed approaches substantially increase translation performance, narrowing the performance gap with bilingual NMT models and pivot-based methods.

In the future, we will develop lightweight alternatives to LALT to reduce the number of model parameters. We will also exploit novel strategies to break the upper bound of ROBT and obtain larger zero-shot improvements, such as generative modeling (Zhang et al., 2016; Su et al., 2018; García et al., 2020; Zheng et al., 2020).

Acknowledgments

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreements 825460 (ELITR) and 825299 (GoURMET). This project has received support from Samsung Electronics Polska sp. z o.o. - Samsung R&D Institute Poland. Rico Sennrich acknowledges support of the Swiss National Science Foundation (MUTAMUR; no. 176727).

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.

Maruan Al-Shedivat and Ankur Parikh. 2019. Consistency by agreement in zero-shot 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 1184–1197, Minneapolis, Minnesota. Association for Computational Linguistics.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Roee Aharoni, Melvin Johnson, and Wolfgang Macherey. 2019a. The missing ingredient in zero-shot neural machine translation. *CoRR*, abs/1903.07091.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, Wolfgang Macherey, Zhifeng Chen, and Yonghui Wu. 2019b. Massively multilingual neural machine translation in the wild: Findings and challenges. *CoRR*, abs/1907.05019.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Ankur Bapna, Mia Chen, Orhan Firat, Yuan Cao, and Yonghui Wu. 2018. Training deeper neural machine translation models with transparent attention. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3028–3033, Brussels, Belgium. Association for Computational Linguistics.

Ankur Bapna and Orhan Firat. 2019. Simple, scalable adaptation 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)*, pages 1538–1548, Hong Kong, China. Association for Computational Linguistics.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Graeme Blackwood, Miguel Ballesteros, and Todd Ward. 2018. Multilingual neural machine translation with task-specific attention. In *Proceedings of the 27th International Conference on Computational*

Linguistics, pages 3112–3122, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Anna Currey and Kenneth Heafield. 2019. Zero-resource neural machine translation with monolingual pivot data. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 99–107, Hong Kong. Association for Computational Linguistics.

Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-task learning for multiple language translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1723–1732, Beijing, China. Association for Computational Linguistics.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016a. 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*, pages 866–875, San Diego, California. Association for Computational Linguistics.

Orhan Firat, Baskaran Sankaran, Yaser Al-onaizan, Fatos T. Yarman Vural, and Kyunghyun Cho. 2016b. Zero-resource translation with multi-lingual neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 268–277, Austin, Texas. Association for Computational Linguistics.

Xavier García, Pierre Forêt, Thibault Sellam, and Ankur P. Parikh. 2020. A multilingual view of unsupervised machine translation. *ArXiv*, abs/2002.02955.

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. 2016. Toward multilingual neural machine translation with universal encoder and decoder. In *Proceedings of the 13th International Workshop on Spoken Language Translation (IWSLT)*, Seattle, USA.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition. *CoRR*, abs/1512.03385.

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.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*.

Philipp Koehn. 2010. *Statistical Machine Translation*, 1st edition. Cambridge University Press, New York, NY, USA.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Surafel M. Lakew, Marcello Federico, Matteo Negri, and Marco Turchi. 2019. Multilingual Neural Machine Translation for Zero-Resource Languages. *arXiv e-prints*, page arXiv:1909.07342.

Surafel Melaku Lakew, Mauro Cettolo, and Marcello Federico. 2018. A comparison of transformer and recurrent neural networks on multilingual neural machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 641–652, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Jason Lee, Kyunghyun Cho, and Thomas Hofmann. 2017. Fully character-level neural machine translation without explicit segmentation. *Transactions of the Association for Computational Linguistics*, 5:365–378.

Yichao Lu, Phillip Keung, Faisal Ladhak, Vikas Bhardwaj, Shaonan Zhang, and Jason Sun. 2018. A neural interlingua for multilingual machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 84–92, Brussels, Belgium. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Emmanouil Antonios Platanios, Mrinmaya Sachan, Graham Neubig, and Tom Mitchell. 2018. Contextual parameter generation for universal neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 425–435, Brussels, Belgium. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on*

Machine Translation: Research Papers, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Devendra Sachan and Graham Neubig. 2018. Parameter sharing methods for multilingual self-attentional translation models. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 261–271, Brussels, Belgium. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. 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.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Jinsong Su, Shan Wu, Deyi Xiong, Yaojie Lu, Xianpei Han, and Biao Zhang. 2018. Variational recurrent neural machine translation. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Xu Tan, Jiale Chen, Di He, Yingce Xia, Tao QIN, and Tie-Yan Liu. 2019. Multilingual neural machine translation with language clustering. 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 963–973, Hong Kong, China. Association for Computational Linguistics.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Raúl Vázquez, Alessandro Raganato, Jörg Tiedemann, and Mathias Creutz. 2019. Multilingual NMT with a language-independent attention bridge. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 33–39, Florence, Italy. Association for Computational Linguistics.

Qiang Wang, Bei Li, Tong Xiao, Jingbo Zhu, Changliang Li, Derek F. Wong, and Lidia S. Chao. 2019a. Learning deep transformer models for machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1810–1822, Florence, Italy. Association for Computational Linguistics.

Xinyi Wang, Hieu Pham, Philip Arthur, and Graham Neubig. 2019b. Multilingual neural machine translation with soft decoupled encoding. In *International Conference on Learning Representations*.

Yining Wang, Long Zhou, Jiajun Zhang, Feifei Zhai, Jingfang Xu, and Chengqing Zong. 2019c. A compact and language-sensitive multilingual translation method. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1213–1223, Florence, Italy. Association for Computational Linguistics.

Biao Zhang, Ivan Titov, and Rico Sennrich. 2019. Improving deep transformer with depth-scaled initialization and merged attention. 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 898–909, Hong Kong, China. Association for Computational Linguistics.

Biao Zhang, Deyi Xiong, Jinsong Su, Hong Duan, and Min Zhang. 2016. Variational neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 521–530, Austin, Texas. Association for Computational Linguistics.

Zaixiang Zheng, Hao Zhou, Shujian Huang, Lei Li, Xin-Yu Dai, and Jiajun Chen. 2020. Mirror-generative neural machine translation. In *International Conference on Learning Representations*.

A OPUS-100: The OPUS Multilingual Dataset

Table 8 lists the languages (other than English) and numbers of sentence pairs in the English-centric multilingual dataset.

B Model Settings

We optimize model parameters using Adam ($\beta_1 = 0.9, \beta_2 = 0.98$) (Kingma and Ba, 2015) with label smoothing of 0.1 and scheduled learning rate (warmup step 4k). We set the initial learning rate to 1.0 for bilingual models, but use 0.5 for multilingual models in order to stabilize training. We apply dropout to residual layers and attention weights, with a rate of 0.1/0.1 for 6-layer Transformer models and 0.3/0.2 for deeper ones. We group sentence

Table 8: Numbers of training, validation, and test sentence pairs in the English-centric multilingual dataset.

Language		Train	Valid	Test
af	Afrikaans	275512	2000	2000
am	Amharic	89027	2000	2000
an	Aragonese	6961	0	0
ar	Arabic	1000000	2000	2000
as	Assamese	138479	2000	2000
az	Azerbaijani	262089	2000	2000
be	Belarusian	67312	2000	2000
bg	Bulgarian	1000000	2000	2000
bn	Bengali	1000000	2000	2000
br	Breton	153447	2000	2000
bs	Bosnian	1000000	2000	2000
ca	Catalan	1000000	2000	2000
cs	Czech	1000000	2000	2000
cy	Welsh	289521	2000	2000
da	Danish	1000000	2000	2000
de	German	1000000	2000	2000
dz	Dzongkha	624	0	0
el	Greek	1000000	2000	2000
eo	Esperanto	337106	2000	2000
es	Spanish	1000000	2000	2000
et	Estonian	1000000	2000	2000
eu	Basque	1000000	2000	2000
fa	Persian	1000000	2000	2000
fi	Finnish	1000000	2000	2000
fr	French	1000000	2000	2000
fy	Western Frisian	54342	2000	2000
ga	Irish	289524	2000	2000
gd	Gaelic	16316	1605	1606
gl	Galician	515344	2000	2000
gu	Gujarati	318306	2000	2000
ha	Hausa	97983	2000	2000
he	Hebrew	1000000	2000	2000
hi	Hindi	534319	2000	2000
hr	Croatian	1000000	2000	2000
hu	Hungarian	1000000	2000	2000
hy	Armenian	7059	0	0
id	Indonesian	1000000	2000	2000
ig	Igbo	18415	1843	1843
is	Icelandic	1000000	2000	2000
it	Italian	1000000	2000	2000
ja	Japanese	1000000	2000	2000
ka	Georgian	377306	2000	2000
kk	Kazakh	79927	2000	2000
km	Central Khmer	111483	2000	2000
kn	Kannada	14537	917	918
ko	Korean	1000000	2000	2000
ku	Kurdish	144844	2000	2000
ky	Kyrgyz	27215	2000	2000
li	Limburgan	25535	2000	2000
lt	Lithuanian	1000000	2000	2000
lv	Latvian	1000000	2000	2000
mg	Malagasy	590771	2000	2000
mk	Macedonian	1000000	2000	2000
ml	Malayalam	822746	2000	2000
mn	Mongolian	4294	0	0
mr	Marathi	27007	2000	2000
ms	Malay	1000000	2000	2000
mt	Maltese	1000000	2000	2000
my	Burmese	24594	2000	2000
nb	Norwegian Bokmål	142906	2000	2000
ne	Nepali	406381	2000	2000
nl	Dutch	1000000	2000	2000
nn	Norwegian Nynorsk	486055	2000	2000
no	Norwegian	1000000	2000	2000
oc	Occitan	35791	2000	2000
or	Oriya	14273	1317	1318
pa	Panjabi	107296	2000	2000
pl	Polish	1000000	2000	2000
ps	Pashto	79127	2000	2000
pt	Portuguese	1000000	2000	2000
ro	Romanian	1000000	2000	2000
ru	Russian	1000000	2000	2000
rw	Kinyarwanda	173823	2000	2000
se	Northern Sami	35907	2000	2000
sh	Serbo-Croatian	267211	2000	2000
si	Sinhala	979109	2000	2000
sk	Slovak	1000000	2000	2000
sl	Slovenian	1000000	2000	2000
sq	Albanian	1000000	2000	2000
sr	Serbian	1000000	2000	2000
sv	Swedish	1000000	2000	2000
ta	Tamil	227014	2000	2000
te	Telugu	64352	2000	2000
tg	Tajik	193882	2000	2000
th	Thai	1000000	2000	2000
tk	Turkmen	13110	1852	1852
tr	Turkish	1000000	2000	2000
tt	Tatar	100843	2000	2000
ug	Uighur	72170	2000	2000
uk	Ukrainian	1000000	2000	2000
ur	Urdu	753913	2000	2000
uz	Uzbek	173157	2000	2000
vi	Vietnamese	1000000	2000	2000
wa	Walloon	104496	2000	2000
xh	Xhosa	439671	2000	2000
yi	Yiddish	15010	2000	2000
yo	Yoruba	10375	0	0
zh	Chinese	1000000	2000	2000
zu	Zulu	38616	2000	2000

pairs of roughly 50k target tokens into one training/finetuning batch, except for bilingual models where 25k target tokens are used. We train multilingual and bilingual models for 500k and 100k steps, respectively. We average the last 5 checkpoints for evaluation, and employ beam search for decoding with a beam size of 4 and length penalty of 0.6.

C Dynamics of Multilingual Translation

EMNLP 2020 Submission ***. Confidential Review Copy. DO NOT DISTRIBUTE.

Dynamics of Multilingual Translation

Anonymous EMNLP submission

Abstract

Neural networks have been at the helm of landmark progress in several aspects of natural language processing, including machine translation. One of the primary drivers of such progress has been the discovery and the subsequent use of attention mechanisms in neural machine translation systems. Lately attention layers have also been used as a tool for interpretation of model behavior and developing insights about how models work internally. In this paper we attempt to leverage attention weights corresponding to output tokens to understand how sequence-to-sequence models use data from the source and target sides while making predictions. We also show how model behavior is impacted by several linguistic factors. Finally we use our attention-based analysis to gain insights into how multilingual systems behave.

1 Introduction

Neural networks, and more specifically Deep Learning (DL) has lately emerged as one of the most widely used approaches for Machine Translation (MT), a difficult problem of Natural Language Processing (NLP), Brill and Mooney, 1997) in the form of Neural Machine Translation (NMT). NMT is generally modelled as sequence-to-sequence learning (Sutskever et al., 2014) that use two language models (encoder and decoder) to translate a sentence from one language to another. Originally, this required the entire sentence to be crammed into one fixed-length vector by the encoder which causes problems with longer sentences (Cho et al., 2014). The attention mechanism by Bahdanau et al. (2014) and its modifications by others (Hu, 2019; Chaudhari et al., 2019) successfully mitigated the problem to some extent. The attention mechanism simply learns the probability of which encoder state corresponds to a particular decoder state. The information from the attention mechanism along with the final state vector of the encoder is fed to the decoder, which boosts system performance. A series of WMT conferences over the years, established the usage of attention in state-of-the-art systems (Bojar et al., 2016; Ondrej et al., 2017; Barrault et al., 2019). The encoder-decoder approach for MT with the attention mechanism has been further extended to multilingual settings (Firat et al., 2016; Johnson et al., 2017; Schwenk and Douze, 2017; Aharoni et al., 2019) and image captioning (Hossain et al., 2019) to produce good results over different evaluation benchmarks.

The present work attempts to extend the work on interpretability of neural-network-based machine translation systems by using the attention mechanism as a tool for interpreting and understanding neural machine translation models. In this paper, we discuss the results of a set of experiments exploring how attention weights can provide insights into model behavior. All the experiments here follow a basic template of a system having RNN-based encoders and decoders with a hierarchical attention combination mechanism (Libovický and Helcl, 2017) equipped with a sentinel mechanism (Lu et al., 2017). The hierarchical combination has earlier been used for multi-modal tasks by Libovický et al. (2016) where the combination attention was used to selectively focus between the image or text. The purpose of using the attention combination strategy in the present setup was to analyze the dynamics of the encoder(s) and decoder in a pure text-processing setting by observing how attention energies are distributed with progress in training.

Our main contributions in the paper are as follows:

- Show how the Sentinel Attention Activation Ratio (SAAR) can be used to improve model interpretability.
- Use attention weights to understand model

behavior as it is getting trained.

- Use attention-based interpretation methods to study multi-lingual behavior of models.

2 Related Work

A number of researchers have looked at the issue of interpretability of deep neural networks, especially in the context of natural language processing. Li et al. (2015) give an brief review of techniques for interpreting and visualizing neural models in the context of NLP applications. Using the attention mechanism as a tool for understanding model behavior has also been proposed and implemented. Mareček and Rosa (2018) and Pham et al. (2019) use self-attention weights of encoder in a Transformer model to extract syntactic trees in order to analyze networks with respect to learning syntax. Raganato and Tiedemann (2018) use self-attention weights of the encoder of a transformer to extract dependency relations and then do a range of probing tasks with the extracted representation.
There is however a debate pertaining to the usefulness of attention weights as a measure of interpretability. While Serrano and Smith (2019) and Jain and Wallace (2019) argue that attention cannot be used to understand the basis for prediction for models, Vig and Belinkov (2019) show that attention is capable of capturing linguistic notions and giving 'human-interpretable descriptions of model behavior'. Vashishth et al. (2019) have also shown that attention weights are correlated with feature importance measures and conclude that they are interpretable.
Ghader and Monz (2017) and also Koehn and Knowles (2017) show how attention is different from traditional alignment, that is used in statistical neural machine translation systems. Voita et al. (2018) show that an attention layers can be used for learning anaphora resolution. Domhan (2018) analyze different attention based architectures and conclude that multiple source attention mechanisms and residual feedforward blocks bring RNNs closer to Transformers. It is thus evident that the attention mechanism can be used as a tool for interpretability of neural networks and to plug in external knowledge for extending the capabilities of NLP systems (Galassi et al., 2019). Rikters et al. (2017) describe a tool to understand how output translations were produced by models by using attention activation.

3 Experiment

3.1 Data and Tools

All the experiments were done using the Neural-Monkey (Helcl and Libovický, 2017) toolkit and trained on the Multi30k[1] dataset. Multi30k is a Multilingual Multiway Corpus (MMC) containing image captions for 31,014 images in English, German, French and Czech.

3.2 Model Design

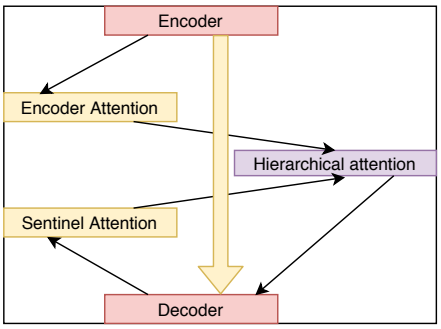


Figure 1: High-level network overview

Each experiment in this study made use of an architecture comprised of one or more GRU-based encoders and a GRU-based decoder along with a hierarchical attention combination mechanism. The key components are schematically captured in Figure 1.

Both the encoder and decoder units consisted of an embedding layer of 300 units and a recurrent layer of 300 units with a dropout rate of 0.5. The decoder unit was also equipped with a conditional GRU mechanism (Calixto et al., 2017). The encoder attention unit was fed into the hierarchical attention combination unit equipped with the sentinel mechanism (decoder attention). The Multi30k dataset was then used to train a range of models with different combinations of three different languages (French, Czech and German) for the encoder and English for the decoder. The models were trained for the period of 100 epochs. At the conclusion of training, the attention weights of each output token in a translated sentence (from the validation set) through different training steps (1 training step = 1 gradient update) were extracted and analyzed. In other words, we analyze the "forced

[1] https://github.com/multi30k/dataset

decoding" model behavior when scoring a given expected output.

For the experiment, no early stopping mechanism was used. However for the sake of analysis a criteria where the BLEU performance of the validation set does not improve in 300 training steps has been chosen as a possible early stopping mechanism. The objective of the inclusion was to see when the model would terminate training ideally and the kind of changes that happen inside the model after that. After the model was trained, the attention weights corresponding to the tokens of sentences in the validation set were extracted. These weights were analyzed to determine how the behavior of the different attention units (decoder,encoder(s)) corresponding to output translation tokens evolved during the course of the training. Experiments with multiple languages were done by adding multiple encoders into the same setup.

3.3 Experiment: Mono-encoder case

The goal of the experiments was to see how the model chose between the encoder and decoder to make a prediction about the output word. Based on linguistic knowledge, one would assume that some words in MT output are more influenced by the source (the encoder) and some more by the target produced so far (the decoder). The experiments were aimed to investigate this dynamics.

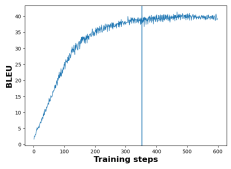


Figure 2: Learning curve of BLEU for DE→EN.

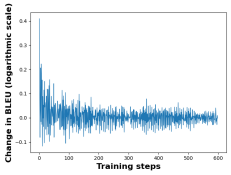


Figure 3: Change in learning curve of BLEU for DE→EN.

German→English: For this experiment, the encoder and decoder were trained with German and English sentences respectively. Figure 2 shows

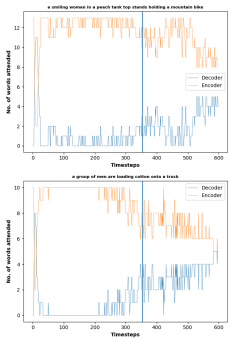


Figure 4: Attention energy distribution for DE→EN.

how the BLEU score improved for the model and Figure 3 shows the rate of change (Δ_τ) of the BLEU scores during the course of the training. For a particular time step τ, Δ_τ was calculated as:

$$\Delta_\tau = log(BLEU_\tau) - log(BLEU_{\tau-1}) \quad (1)$$

Figure 3 and Figure 4 show that while there was no significant "growth" in the BLEU scores after the 300^{th} time step, and the BLEU kept hovering around a set of values. The vertical blue line in the image refers to the point where the early stopping criterion would have kicked in and the training would have terminated. Figure 4 shows how the encoder attention and decoder (sentinel) attention behaved during the course of the training. The plot was obtained by recording the attention energies distributed among the different attention units (encoder, decoder) corresponding to the words of a sentence. We sampled 15 sentences from the validation to study the model dynamics, out of which the results for two sentences are shown here. Then, we identified which of the model components (the encoder or the decoder) was promoted by the hierarchical attention. When then plot the number of words of the target sentence where each of the component 'wins'. Figure 4 shows that a sentence in the validation set, the model starts off by trying to use the decoder attention first and then slowly starts relying more and more on the encoder attention. In most sentences however as the model continues being trained, the number of words being predicted by the encoder attention falls and the decoder attention gradually starts being involved in the prediction of more and more output words. The BLEU score doesn't show any significant increase in this period. It thus seems that even though when

the overall translation performance of the model did not show any significant change, the model slowly started using the decoder attention unit to predict some words. It also seems that the length of the sentence impacted how the model distributed its attention between the encoder and the decoder.

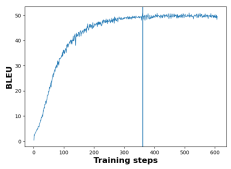


Figure 5: Learning curve of BLEU for FR→EN.

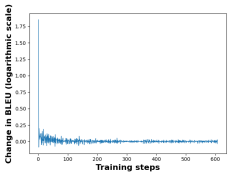


Figure 6: Change in learning curve of BLEU for FR→EN.

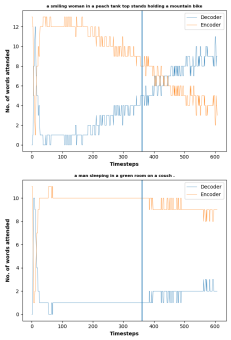


Figure 7: Attention energy distribution for FR→EN.

French→English: The experiment was repeated with a French-English system. And as Figure 5 shows, the system performs reaches higher BLEU scores than the German system. But as Figure 6 shows, the trend of Δ_τ for the model is different from the German model. In the same spirit, although there is an initial tendency of the model to focus on the decoder attention weights for final prediction,the model decides to rely mostly on the encoder for the prediction of most words for the next time steps.

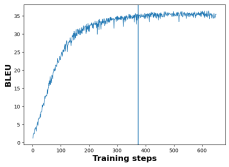


Figure 8: Learning curve of BLEU for CZ→EN.

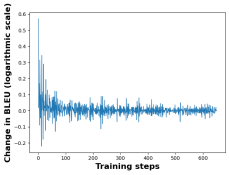


Figure 9: Change in learning curve of BLEU for CZ→EN.

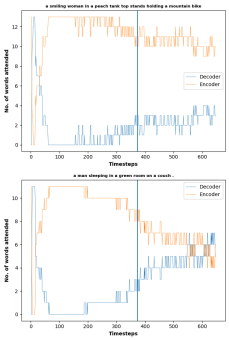


Figure 10: Attention energy distribution for CZ→EN.

Czech→English: The final experiment in this setup was done with a Czech-English system. Figure 8 shows that Δ_τ for this case is lower than the German case but greater than the French case. And in terms of how many output words are predicted by each attention unit, by the end of the training, the Czech and German models tend to use the decoder attention more for final prediction of words. A detailed statistical analysis of this is presented later.

3.4 Bi-encoder experiments

The goal of these experiments was to see how the model decides to use two different encoders and

the decoder to make word predictions.

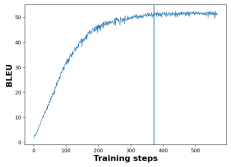


Figure 11: Learning curve of BLEU for FR+DE→EN.

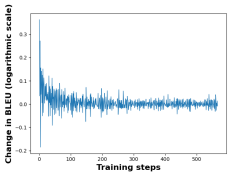


Figure 12: Change in learning curve of BLEU for FR+DE→EN.

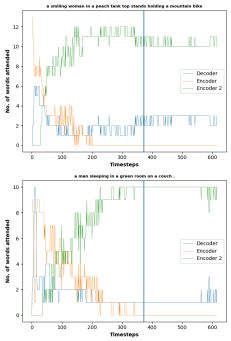


Figure 13: Attention energy distribution for FR+DE→EN.

(German+French)→English: The BLEU performance achieved by the model was similar to the French-English model. The Δ_τ trend was somehow different from the French-English model. It also seems that over time the model decides to use attention from Encoder 2 (German) more than other units for its prediction. It is also interesting to see in Figure 13, the swap in the attentions before the stopping criterion mark. For the monolingual models, this swap was noticed after the stopping criterion mark. But for bilingual models as well as the trilingual models, the swap occurs much before stopping criterion mark.

(French+Czech)→English: In this case, the BLEU performance and Δ_τ trend is alike the French-English model. In terms of attention distribution, the model decides to use attention from Encoder (Czech) more than other units (French encoder/ English decoder) for its prediction over time.

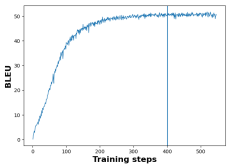


Figure 14: Learning curve of BLEU for CZ+FR→EN.

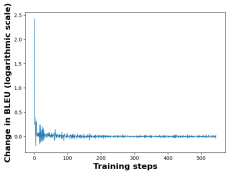


Figure 15: Change in learning curve of BLEU for CZ+FR→EN.

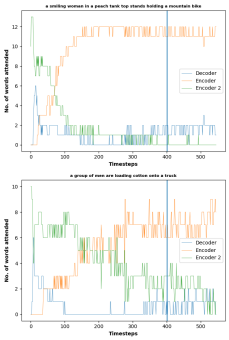


Figure 16: Attention energy distribution for CZ+FR→EN.

(German+Czech)→English: In this case, the BLEU performance is worse than the previous two cases and there is more change in the scores throughout the training period (Figure 18). However, here unlike the two cases above, no particular is a clear primary source and the trends change with sentences.
Thus, when the Czech and German encoders are available, the selection of the encoder becomes dependent on the particular sentence. This is interesting because French has the same word-order as

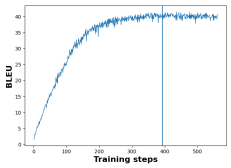


Figure 17: Learning curve of BLEU for CZ+DE→EN.

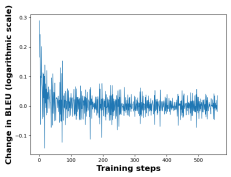


Figure 18: Change in learning curve of BLEU for CZ+DE→EN.

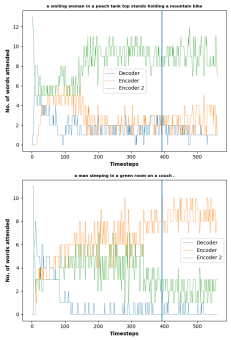


Figure 19: Attention energy distribution for CZ+DE→EN.

English. These observations, may show how language attributes affect model performance (Birch et al., 2008) for neural machine translation systems. Effect of language relatedness on NMT systems in the context of transfer learning has already been demonstrated by Kocmi and Bojar (2018) and for SMT systems by Kolovratník et al. (2009). We speculate that the word order similarity makes French "easier to digest" in the early stages of training. But the source language becomes more informative when the reordering patterns are understood by the model. Why German or Czech would be more informative than French when translating into English is still unclear.

3.5 Tri-encoder experiment

A final experiment was conducted using three encoders as an extension to the previous experiments.

Figure 22 shows the attention distribution for

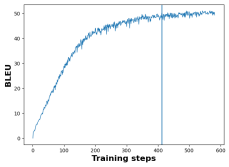


Figure 20: Learning curve of BLEU for FR+CZ+DE→EN.

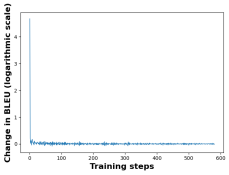


Figure 21: Change in learning curve of BLEU for FR+CZ+DE→EN.

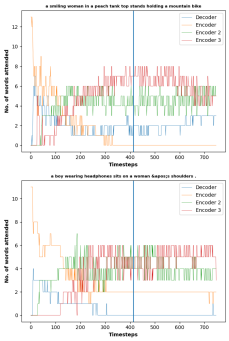


Figure 22: Attention energy distribution for FR+CZ+DE→EN.

two sentences with 13 and 10 words respectively. Each of the words in the sentences are produced with most of the attention coming from one of the attention sources. At each training step of the training, we use the current model and plot the number of target words that were most influenced by the encoder(s) and the decoder. We see that in early stages of the training, the encoder is the most influential element but it later gets virtually ignored as the model learns to rely more on the other two encoders. The BLEU performance of the model

was alike all the models with a French encoder with little change in BLEU performance over time. In terms of attention distribution however all the encoders and the decoder are used by the model to make predictions.

cz_en	fr_en	de_en
30.6	43	33.2

Table 1: Average BLEU scores of monolingual models at the end of training.**cz_en** refers to CZ→EN, **fr_en** refers to FR→EN and **de_en** refers to DE→EN

cz_de_en	cz_fr_en	de_fr_en	3_en
34	43.5	42.6	40.3

Table 2: Average BLEU scores of multilingual models at the end of training.**cz_de_en** refers to CZ+DE→EN, **cz_fr_en** refers to CZ+FR→EN, **de_fr_en** refers to DE+FR→EN and **3_en** refers to FR+CZ+DE→EN.

4 Sentinel attention activation with different lengths

A metric in the form of sentinel attention activation ratio (SAAR) was used to understand how much the decoder was relied upon by the model to make its final predictions. For a particular sentence S_i, SAAR was calculated as:

$$S_i = \frac{A_s}{A_t}$$

where A_s was the number of words whose prediction was based on the decoder during the entire training and A_t represents the total count of attention units activated during training. Now for each model, the corresponding SAAR for all sentences in the validation set was calculated followed by calculating their correlation with sentence length, as shown in Table 3 and Table 4. The correlation values seem to indicate that there is a weak but mostly negative correlation between sentence length and SAAR. Thus, the longer the source sequences were, the greater the model fell back to the decoder attention to make the final prediction. The values also indicate that the maximum negative correlation between sentence length and SAAR was noticed in the German-French-English model. In other words, the greater the source sequence, greater was the tendency of the model to use the decoder attention for making final predictions. However it has already been seen that for this model, the German encoder was preferred over the French encoder. Thus for this model, longer source sequences implied more reliance on the German encoder for predicting the output word. Also, the minimum negative correlation between sentence length and SAAR was noticed in the German-Czech-English model, where there was no clear distinction of which encoder the model chose for final prediction.

cz_en	fr_en	de_en
-0.393	0.010	-0.175

Table 3: Correlation between SAAR and sentence length of monolingual models.

cz_de_en	cz_fr_en	de_fr_en	3_en
-0.1145	-0.242	-0.362	-0.126

Table 4: Correlation between SAAR and sentence length of multilingual models.

5 Extent of Multilinguality

A study of model behavior in the previous section makes is apparent that BLEU scores alone are not a sufficient metric for judging how good a model is, especially for multilingual models. We therefore propose to measure the extent of multilinguality using perplexity.
Mathematically, perplexity is the measure of how well a probability distribution function is capable of predicting a sample. In this case we calculated the perplexity for the model in terms of multilinguality. Given a sentence s_i of the validation set, we used the attention energy matrix for that sentence over all time steps to calculate the entropy of all the components (encoder(s), decoder). For each component. the matrix M_s, was used to obtain the number of words for which it "won". This frequency data was used to calculate the entropy for each component. Thus. the perplexity of each component is calculated as:

$$P_i = 2^{Entropy(x)} \quad (2)$$

Here x represents the frequency counts across training steps obtained from the attention matrix. For each sentence s_i, the average perplexity across all components was calculated as:

$$P_{s_i} = \frac{\sum_{i=1}^{m} P_i}{m} \quad (3)$$

Here m denotes the number of components in the model. Finally, the extent of multilinguality (M) was given as:

$$M = \frac{\sum_{i=1}^{n} P_{s_i} - n}{\sum_{i=1}^{n} P_{s_i}} \quad (4)$$

cz_en	fr_en	de_en
0.987	0.986	0.986
cz_de_en	cz_fr_en	fr_de_en
0.986	0.984	985
cz_de_fr_en		
0.988		

Table 5: Extent of multilinguality in models

6 When Self-Attendance Happens

As suggested in the introduction, we would like to reveal when the decoder is more influenced by the encoder and when it primarily follows itself. We hypothesize that in a given target sentence, the presence of most words is primarily influenced by the source language. In other words, the encoder observes the content words and conveys this information to the decoder. Some words are however dependent only on the target language and the decoder could produce them regardless of the source. These would be any auxiliaries, purely grammatical words, or "non-head words" in the terminology of Fraser and Marcu (2007).

To test this hypothesis, we ran the fully trained model and observed if the majority of attention at each output word came from the encoder or the decoder. For each word in the target language, we gathered how often it was most influenced by the decoder as it was produced in our forced decoding setup of the validation set and gathered the "proportion of activation" given by the ratio between the number of times the decoder "won" in predicting the word and the total number of predictions for the word respectively. Then we listed the words with highest activation proportion.

It is conceivable, that the decision of the decoder is one time step later or earlier than when the auxiliary word was actually produced (in accordance with what Koehn and Knowles (2017) observe for attention vs. word alignment). We thus collected the statistics of the activation unit responsible for the prediction of words at the exact position, one position *before* (Figure 23) and *after* (Figure 24) the decoder was the most influential component of attention. Unfortunately, we could not confirm our hypothesis. English auxiliary words such as the articles or prepositions, or punctuation, cannot be easily recognized in any of the lists.

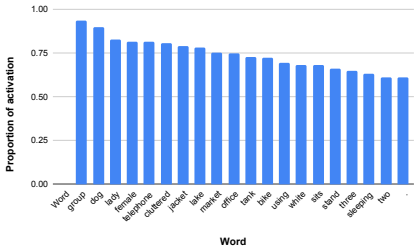


Figure 23: Top 20 words activated by decoder (one step before)

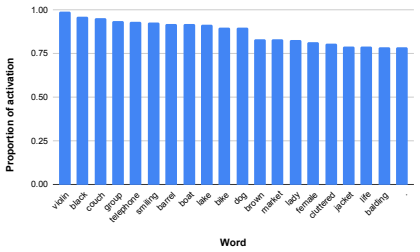


Figure 24: Top 20 words activated by decoder (one step after)

7 Conclusion

The paper has described the observations and results from 7 different experiments performed to understand how analyzing attention weights might give insights into the interpretability of neural machine translation systems. From the experiments, it is clear that using a setup that employs the hierarchical attention combination mechanism can give us an interesting picture of how the model trains and how the model performance is impacted by factors like sentence length or linguistic attributes or even how related the languages being used in the system are. It seems from the results that the model develops some internal criteria to focus on some particular languages during training and then start focusing on other languages after a good enough performance is achieved.

References

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

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Loïc Barrault, Ondřej Bojar, Marta R Costa-Jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, et al. 2019. Findings of the 2019 conference on machine translation (wmt19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61.

Alexandra Birch, Miles Osborne, and Philipp Koehn. 2008. Predicting success in machine translation. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 745–754.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, et al. 2016. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 131–198.

Eric Brill and Raymond J Mooney. 1997. An overview of empirical natural language processing. *AI magazine*, 18(4):13–13.

Iacer Calixto, Qun Liu, and Nick Campbell. 2017. Doubly-attentive decoder for multi-modal neural machine translation. *arXiv preprint arXiv:1702.01287*.

Sneha Chaudhari, Gungor Polatkan, Rohan Ramanath, and Varun Mithal. 2019. An attentive survey of attention models. *arXiv preprint arXiv:1904.02874*.

Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*.

Tobias Domhan. 2018. How much attention do you need? a granular analysis of neural machine translation architectures. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1799–1808, Melbourne, Australia. Association for Computational Linguistics.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism. *arXiv preprint arXiv:1601.01073*.

Alexander Fraser and Daniel Marcu. 2007. Getting the structure right for word alignment: Leaf. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, pages 51–60.

Andrea Galassi, Marco Lippi, and Paolo Torroni. 2019. Attention, please! a critical review of neural attention models in natural language processing. *arXiv preprint arXiv:1902.02181*.

Hamidreza Ghader and Christof Monz. 2017. What does attention in neural machine translation pay attention to? In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 30–39, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Jindřich Helcl and Jindřich Libovický. 2017. Neural monkey: An open-source tool for sequence learning. *The Prague Bulletin of Mathematical Linguistics*, 107(1):5–17.

MD Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. 2019. A comprehensive survey of deep learning for image captioning. *ACM Computing Surveys (CSUR)*, 51(6):1–36.

Dichao Hu. 2019. An introductory survey on attention mechanisms in nlp problems. In *Proceedings of SAI Intelligent Systems Conference*, pages 432–448. Springer.

Sarthak Jain and Byron C Wallace. 2019. Attention is not explanation. *arXiv preprint arXiv:1902.10186*.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Tom Kocmi and Ondřej Bojar. 2018. Trivial transfer learning for low-resource neural machine translation. *arXiv preprint arXiv:1809.00357*.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

David Kolovratník, Natalia Klyueva, and Ondrej Bojar. 2009. Statistical machine translation between related and unrelated languages. In *Proceedings of the Conference on Theory and Practice on Information Technologies*, pages 31–36. Citeseer.

Jiwei Li, Xinlei Chen, Eduard Hovy, and Dan Jurafsky. 2015. Visualizing and understanding neural models in nlp. *arXiv preprint arXiv:1506.01066*.

Jindřich Libovický and Jindřich Helcl. 2017. Attention strategies for multi-source sequence-to-sequence learning. *arXiv preprint arXiv:1704.06567*.

Jindřich Libovický, Jindřich Helcl, Marek Tlustỳ, Pavel Pecina, and Ondřej Bojar. 2016. Cuni system for wmt16 automatic post-editing and multimodal translation tasks. *arXiv preprint arXiv:1606.07481*.

Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2017. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 375–383.

David Mareček and Rudolf Rosa. 2018. Extracting syntactic trees from transformer encoder self-attentions. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 347–349.

Bojar Ondrej, Rajen Chatterjee, Federmann Christian, Graham Yvette, Haddow Barry, Huck Matthias, Koehn Philipp, Liu Qun, Logacheva Varvara, Monz Christof, et al. 2017. Findings of the 2017 conference on machine translation (wmt17). In *Second Conference onMachine Translation*, pages 169–214. The Association for Computational Linguistics.

Thuong-Hai Pham, Dominik Macháček, and Ondřej Bojar. 2019. Promoting the knowledge of source syntax in transformer nmt is not needed. *arXiv preprint arXiv:1910.11218*.

Alessandro Raganato and Jörg Tiedemann. 2018. An analysis of encoder representations in transformer-based machine translation. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 287–297, Brussels, Belgium. Association for Computational Linguistics.

Matīss Rikters, Mark Fishel, and Ondřej Bojar. 2017. Visualizing neural machine translation attention and confidence. *The Prague Bulletin of Mathematical Linguistics*, 109(1):39–50.

Holger Schwenk and Matthijs Douze. 2017. Learning joint multilingual sentence representations with neural machine translation. *arXiv preprint arXiv:1704.04154*.

Sofia Serrano and Noah A Smith. 2019. Is attention interpretable? *arXiv preprint arXiv:1906.03731*.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Shikhar Vashishth, Shyam Upadhyay, Gaurav Singh Tomar, and Manaal Faruqui. 2019. Attention interpretability across nlp tasks. *arXiv preprint arXiv:1909.11218*.

Jesse Vig and Yonatan Belinkov. 2019. Analyzing the structure of attention in a transformer language model. *arXiv preprint arXiv:1906.04284*.

Elena Voita, Pavel Serdyukov, Rico Sennrich, and Ivan Titov. 2018. Context-aware neural machine translation learns anaphora resolution. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1264–1274, Melbourne, Australia. Association for Computational Linguistics.