# Domain Adaptation of Document-Level NMT in IWSLT19

*Martin Popel,*[†‡] *Christian Federmann*[‡]

[†]Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics, Malostranské náměstí 25, 118 00 Prague, Czech Republic

[‡]Microsoft, 1 Microsoft Way, Redmond, WA 98121, USA

popel@ufal.mff.cuni.cz, chrife@microsoft.com

## Abstract

We describe our four NMT systems submitted to the IWSLT19 shared task in English→Czech text-to-text translation of TED talks. The goal of this study is to understand the interactions between document-level NMT and domain adaptation. All our systems are based on the Transformer model implemented in the Tensor2Tensor framework. Two of the systems serve as baselines, which are not adapted to the TED talks domain: SENTBASE is trained on single sentences, DOCBASE on multi-sentence (document-level) sequences. The other two submitted systems are adapted to TED talks: SENTFINE is fine-tuned on single sentences, DOCFINE is fine-tuned on multi-sentence sequences. We present both automatic-metrics evaluation and manual analysis of the translation quality, focusing on the differences between the four systems.

## 1. Introduction

Neural machine translation (NMT) has recently achieved excellent results in the news translation task. Hassan et al. [1] report achieving a "human parity" on Chinese→English news translation. WMT 2018 overview paper [2, p. 291] reports that our English→Czech system "CUNI Transformer" [3] was evaluated as significantly better ($p < 0.05$) than the human reference. However, it has been shown [4, 5] that evaluating the quality of translation of news articles on isolated sentences without the context of the whole document (as done in WMT 2018) is not sufficient. Thus, the research has focused on document-level translation (see e.g. [6, 7, 8]) which is trained simply by training on multi-sentence sequences.[1]

Another line of research focuses on domain adaptation of NMT; see [10] for an overview. One of the most simple and effective techniques is fine-tuning [11], where an NMT model trained on (large) general-domain (or "out-domain") data is further trained on (smaller) in-domain data. The term "domain" in domain adaptation is usually understood very broadly – a domain can be defined by any property of the training data (and expected test data), such as the topic, genre, formality, style, written vs. spoken language etc.

As far as we know, there is no prior work on the interaction of the above-mentioned approaches to NMT: – document-level translation and domain adaptation. Is domain adaptation of document-level systems different from the domain adaptation of sentence-level systems? What are the differences in the translation output? While we have no definite answers to these questions, we hope our present work brings some new insights into the issue.

## 2. Systems overview

We use the following four systems in our experiments:

- SENTBASE is the winning system of the English-Czech WMT 2018 shared task (under name "CUNI Transformer", i.e. Charles University Transformer). It is described in [3]. It is a Transformer model trained with iterated concat backtranslation [3] on single sentences from the WMT (general-domain) training data.

- DOCBASE is one of the winning systems of English-Czech WMT 2019 (under name "DocTransformer T2T"). It is described in [8]. It is trained similarly to SENTBASE, but in a document-level fashion, on sequences of up to 1000 characters (and on slightly larger data than SENTBASE). At inference time, the final translation is produced by merging several overlapping multi-sentence sequences.

- SENTFINE is trained by initializing the parameters with the DOCBASE model[2] and fine-tuning on sentences from the in-domain training data.

[1] Earlier approaches to document-level NMT used more complicated architectures, e.g. adding a special encoder for encoding the context of previous sentences [9].

[2] We trained also a fine-tuned model initialized with SENTBASE, but it achieved slightly worse dev-set BLEU than SENTFINE, so we did not include this system into our submission. Another motivation was to have SENTFINE and DOCFINE as comparable as possible, i.e. trained on the same data and differing only in the fine-tuning.

| data set | sentence pairs (k) | words (k) EN | CS |
|---|---|---|---|
| CzEng 1.7 | 57 065 | 618 424 | 543 184 |
| Europarl v7 | 647 | 15 625 | 13 000 |
| News Commentary v12 | 211 | 4 544 | 4 057 |
| CommonCrawl | 162 | 3 349 | 2 927 |
| WikiTitles | 361 | 896 | 840 |
| EN NewsCrawl 2016–17 | 47 483 | 934 981 | |
| CS NewsCrawl 2007–18 | 78 366 | | 1 108 352 |
| MuST-C train (TED talks) | 128 | 2 414 | 2 001 |
| total | 184 423 | 1 580 233 | 1 674 361 |

Table 1: Training data sizes (in thousands).

- DOCFINE is similar to SENTFINE (also initialized with DOCBASE and fine-tuned), but the fine-tuning was done in a document-level fashion, on multi-sentence sequences of up to 1000 characters from the in-domain training data.

We consider the first two systems as our baselines.

## 3. Experimental Setup

### 3.1. Data sources

Our training data (see Table 1) are constrained to the data allowed in the IWSLT2019 shared task: over half gigaword of parallel out-domain data (mostly CzEng 1.7 [12]), over one gigiword of monolingual out-domain data (Czech NewsCrawl 2007–2018 from WMT)[3] and two megawords of parallel in-domain data (MuST-C v1.1 corpus of TED talks [13]). All the out-domain data were preprocessed, filtered and backtranslated by the same process as in [3].

Our development and test data is reported in Table 2. We used the MuST-C dev set for early stopping of fine-tuning. We also tracked the BLEU performance of our fine-tuning on out-domain development set WMT08-15noncz, which is a concatenation of English-Czech WMT news tests from 2008–2015 excluding originally Czech sentences (i.e. restricting the Czech references to sentences translated from English). After selecting the final four systems for submission, we translated the official IWSLT 2019 test set (tst-IWSLT19) and two additional test sets tst-COMMON and tst-HE included in the MuST-C corpus.

### 3.2. Common training setup

Our four systems are implemented in the Tensor2Tensor (T2T) framework [14], version 1.6.0, following the recommendations of [15]. We used `-batch_size=2900` in all experiments (i.e. a batch size of approximately 2900 tokens per GPU), but we used various numbers of GPUs as indicated in Table 3, resulting in different effective batch size. We use checkpoint averaging of the last eight checkpoints in all experiments. See [3, 8] for the exact hyper-parameter setups.

[3]The English monolingual data was only used for iterated-backtranslation training of our two baseline systems.

| data set | sentence pairs | words EN | CS |
|---|---|---|---|
| WMT08-15noncz dev | 17 841 | 377 712 | 325 480 |
| MuST-C dev | 1 293 | 25 518 | 22 095 |
| MuST-C tst-COMMON | 2 035 | 36 096 | 29 651 |
| MuST-C tst-HE | 600 | 11 899 | 10 020 |
| tst-IWSLT19 | 2 958 | 52 666 | ? |

Table 2: Development and test data sizes.

| system | GPUs | steps | time |
|---|---|---|---|
| SENTBASE | 8x GTX 1080 Ti | 928k | 8 days |
| DOCBASE | 10x GTX 1080 Ti | 661k | 9 days |
| SENTFINE | 4x Titan Xp | 800 | 13 minutes |
| DOCFINE | 4x Titan Xp | 400 | 9 minutes |

Table 3: Hardware used for training/fine-tuning our systems. In case of the two *FINE systems, the number of training steps and time refer only to the fine-tuning phase (excluding the 661k steps of training DOCBASE). Preparation of back-translation data (described in [3]) is not reported here.

### 3.3. Fine-tuning setup

We fine-tuned by simply continuing to train the DOCBASE model on the in-domain parallel data. We have not altered the learning rate schedule, i.e. we continued to decay the learning rate (already quite small after more than 600k steps of training) according to the inverse-square-root schedule. We decreased the checkpoint saving interval to two minutes, so that we can better track the fine-tuning progress and also better use the checkpoint averaging effect. We decreased the effective batch size by training on 4 GPUs (instead of 10 GPUs in the DOCBASE training). Otherwise, we kept all the hyper-parameters the same as in DOCBASE.

We tracked the training progress on the MuST-C dev set and used the checkpoint with the highest BLEU. This happened relatively quickly (400–800 steps), as reported in Table 3.

## 4. Automatic Evaluation

In this section, we evaluate our four systems submitted to IWSLT2019 with three automatic metrics calculated using sacreBLEU 1.3.7 [16]. The metrics' signatures are: BLEU+case.lc+numrefs.1+smooth.exp+tok.intl, BLEU+case.mixed+numrefs.1+smooth.exp+tok.13a and chrF2+case.mixed+numchars.6+numrefs.1+space.False. While the reference translation of the official test set tst-IWSLT19 was not available at the submission time, we report here the evaluation on tst-COMMON (which we have not used before the submission).

| system | BLEU uncased | BLEU cased | chrF2 cased |
|---|---|---|---|
| SENTFINE | 31.39 | 30.50 | 0.5423 |
| DOCFINE | 31.37 | 30.46 | 0.5438 |
| DOCBASE (WMT19, [8]) | 29.56 | 28.36 | 0.5320 |
| SENTBASE (WMT18, [3]) | 29.07 | 27.92 | 0.5255 |

Table 4: Automatic evaluation on tst-COMMON. Significantly different BLEU scores ($p < 0.05$ bootstrap resampling) are separated by a horizontal line.

Table 4 shows that DOCFINE and SENTFINE achieved almost the same BLEU score (31.4). They are significantly ($+1.8$ BLEU) better than DOCBASE, which is significantly ($+0.5$ BLEU) better than SENTBASE. We can confirm our hypothesis that the improvement caused by fine-tuning is smaller for the document-level models than for the sentence-level models.

Naturally a question arises whether the translations of DOCBASE and DOCFINE are substantially different, given the short fine-tuning (400 steps, 9 minutes, cf. Table 3) and only 0.5 BLEU difference. Similarly, we can ask whether the translations of DOCFINE and SENTFINE are substantially different, i.e. whether the document-level translation has any effect on the fine-tuned systems. Table 5 shows that in both cases the outputs are actually more different than could be expected from Table 4. Interestingly, when evaluated on tst-IWSLT19, DOCFINE is more similar to DOCBASE than to SENTFINE, i.e. the document-level aspect seems to be stronger than the fine-tuning aspect. However, it is the other way round when evaluated on tst-COMMON.

Note that for all pairs of our four systems, BLEU(X, Y) is almost the same as BLEU(Y, X).[4] Thus, the n-gram precision (used in BLEU) is approximately the same as n-gram recall and we can interpret it as an overlap (similarity). When focusing on 4-grams only in tst-COMMON, there is only a 63% overlap between DOCFINE and DOCBASE, and only 80% overlap between DOCFINE and SENTFINE.

## 5. Manual analysis

### 5.1. Domain-adaptation effects

In this section, we study different types of differences between our baseline and fine-tuned systems.

#### *5.1.1. Typographic-style adaptation*

We noticed several differences in the typographic style related to the TED talks subtitles. For example, SENTBASE usually translates "*(Laughter)*" as "*smích*", but the other three systems and the reference usually prefer the capitalized version "*Smích*".[5] While this difference has presumably no effect on the translation quality, it affects the cased (case-sensitive) BLEU score. Another similar example is the preference of m-dash (—) vs. hyphen (-), which affects also the uncased BLEU score.

| tst-IWSLT19 | DOCFINE | DOCBASE | SENTFINE | SENTBASE |
|---|---|---|---|---|
| DOCFINE | – | 90.08 | 84.59 | 62.09 |
| DOCBASE | 90.09 | – | 82.16 | 62.23 |
| SENTFINE | 84.56 | 82.11 | – | 62.81 |
| SENTBASE | 62.07 | 62.20 | 62.82 | – |
| **tst-COMMON** | **DOCFINE** | **DOCBASE** | **SENTFINE** | **SENTBASE** |
| DOCFINE | – | 73.66 | 86.10 | 61.07 |
| DOCBASE | 73.69 | – | 66.19 | 78.49 |
| SENTFINE | 86.04 | 66.13 | – | 62.22 |
| SENTBASE | 61.03 | 78.42 | 62.22 | – |

Table 5: BLEU (cased) similarity between different translations of tst-IWSLT19 (top) and tst-COMMON (bottom). For each cell, the system in a given column is taken as the hypothesis and the system in a given row as the reference.

#### *5.1.2. Sentence segmentation*

Yet another example of typographic differences is the rendering of opening double quotation marks. The Czech language rules require the use of lower quotes symbol („), the reference uses straight upper quotes ("), but SENTBASE uses often (25 occurrences in 15 segments in tst-COMMON) two comma symbols (,,). DOCBASE is also affected (20 occurrences in 11 segments), but there are no occurrences of double-commas in the two fine-tuned systems.

When investigating the source of this error, we found out that all the double-commas are in translations of multi-sentence input segments (lines). The IWSLT test and train sets contain usually a single sentence per line, but sometimes more. When translating the test sets, we have forgotten to re-segment the input into sentences. This is unfortunate because our sentence-level[6] models expect sentence-segmented input. Due to some relics of multi-sentence segments in the training data, the models are able to translate also multi-sentence inputs, but with lower quality because the relics are rare and they are usually from noisier data sources.

The fact that the fine-tuned systems did not produce any double-commas suggests that fine-tuning on MuST-C-train (which also contains some multi-sentence lines) helped to prevent this particular translation error resulting from multi-sentence inputs.[7]

[4] The total translation length is about the same in all four systems (36,383–36,817), so the multiplicative brevity penalty used in BLEU is always higher than 0.992.

[5] In tst-COMMON, the capitalized:lower-cased ratio is 38:15 in the reference, 15:37 in SENTBASE, 33:20 in DOCBASE, 52:0 in SENTFINE and 51:1 in DOCFINE.

[6] Our document-level systems are trained on multi-sentence inputs, but the sentences are separated by a special symbol, so even the document-level systems' outputs may be affected if the symbol is missing at inference time.

[7] In some of the segments, we also noticed that the spacing around quotes is wrong in the input segment – whenever the first quote in the segment was

| | |
|---|---|
| SRC | And I'd really love to show **you** *my* week's worth of outfits right now. |
| REF | A opravdu ráda bych **vám** teď ukázala *své* oblečení na týden. |
| SENTBASE, DOCBASE | A moc ráda bych **ti** teď ukázala *moje* oblečení na celý týden. |
| SENTFINE, DOCFINE | A opravdu ráda bych **vám** teď ukázala *své* týdenní oblečení. |

Figure 1: Example of translation differences.

However, double-commas were not the only translation errors in the multi-sentence segments. After manually inspecting all the 15 segments with double-commas, we found that SENTFINE fixed only the quotation symbols but nothing else, relative to SENTBASE (though there were many changes which did not affect the quality). We also found in post-submission experiments that some translations get improved after properly re-segmenting the input. For example: SENTBASE translates the sentence "*I just want to be able to communicate with him and him to be able to communicate with me.*" as "*Jen chci být schopná komunikovat s ním a on se mnou.*", which is an acceptable translation. However, if the source sentence is followed by other text (as in tst-COMMON), SENTBASE produces an incorrect translation "*Jen chci být schopná komunikovat* ***s ním a s ním, aby byli schopni*** *komunikovat se mnou.*" meaning "*I just want to be able to communicate* ***with him and with him, so that they are able*** *to communicate with me.*".

### *5.1.3. Proper TED talks adaptation*

We found also few examples where the domain adaptation actually improved the translation quality. For example, the baseline non-adapted systems translate "*All right, let's go.*" as "*Tak jo, jdeme.*", where *jdeme* means *to go somewhere*. The fine-tuned systems and reference translate the sentence as "*Dobře, jdeme na to.*", where *jdeme na to* means *let's start*, which is the correct translation in a given context.[8]

Another example of an improvement caused by domain adaptation is shown in Figure 1. The fine-tuned systems correctly translated *you* as plural *vám*, instead of singular *ti*. This is an example of a domain adaptation, which would be difficult to achieve with the document-level context only: the document itself does not indicate that there are multiple persons in the audience. We need to know that a given document is a transcription of a TED talk (and a given occurrence of *you* is addressing the audience).

Another difference between the translations in Figure 1 is the word *my*, where the fine-tuned systems use *své*, which is the correct translation in a given context, while the baseline systems use *moje*, which is acceptable only in informal text (or speech). For completeness, we note yet another difference – *na celý týden* vs. *týdenní* – the fine-tuned systems use a contextually worse translation of *week's worth*, although it is questionable whether this difference is related to the fine-tuning (we could not find any similar differences in other sentences).

| | |
|---|---|
| DOCFINE better than SENTFINE | 11 |
| — doc-related | 7 |
| — unrelated | 4 |
| DOCFINE worse than SENTFINE | 4 |
| similar quality | 44 |
| total diffs | 59 |

Table 6: Manual comparison of translation quality of DOCFINE relative to SENTFINEon 100 sentences from tst-COMMON.

## 5.2. Document-level effects

In this section, we study differences between translations of SENTFINE and DOCFINE, i.e. we study the effect of document-level translation on the fine-tuned systems.

In a pilot annotation, we compared the first 100 sentences of tst-COMMON and identified 59 differences.[9] Table 6 shows the results of this annotation: Most of the differences (44) had either none or negligible effect on the translation quality. In 11 cases, DOCFINE was clearly better than SENTFINE and in 7 out of the 11 cases, we were able to prove that the improvements is caused by the document-level context (the improvement disappeared when translating individual sentences with the DOCFINE model). In 4 cases, DOCFINE was clearly worse than SENTFINE.

While the number of sentences annotated in this pilot study is too small for drawing any conclusions about the overall quality of the compared systems (cf. Section 5.3), we use it for selecting example sentences, which we discuss below.

There was a TED talk about rescuing a homeowner with her dog and shoes from a fire. The talk contained four occurrences of the word *homeowner*, which can be translated into

a closing quote (i.e. the segment starts with a continuation of a direct speech from previous segments). For example: *These devices aren't accessible to people. "And I said," Well, how do you actually communicate? "Has everyone seen the movie" The Diving Bell and the Butterfly? "That's how they communicate — so run their finger along.* This could be another reason for the lower-quality translation.

[8] Interestingly, even the DOCBASE actually translated the sentence correctly as "*Dobře, pojd'me na to.*", but the number of sentences in a given translation sequence did not match the number of source sentences, so a backup substitution by SENTBASE translations was used in the post-processing.

[9] Related differences in multiple words (e.g. consistent difference in inflection of a noun phrase) were counted as a single difference. That said, most of the differences were single words.

Czech either with masculine (*majitel*) or feminine (*majitelka*) gender.

The first occurrence was in a sentence that revealed the homeowner's gender (via a coreferring phrase "*her life*"), so both DOCFINE and SENTFINE translated the word correctly.

The second occurrence was in a sentence not revealing the gender; here SENTFINE choose the incorrect gender and DOCFINE the correct one, obviously using the context of the previous two sentences, which revealed the gender via a coreference chain.

The third occurrence of *homeowner* was eight sentences further (and out of the up-to-1000-characters sequence used in DOCFINE inference) and both SENTFINE and DOCFINE translated it with the incorrect gender.

The fourth occurrence was in a sentence immediately following the third occurrence. The sentence was "*A few weeks later, the department received a letter from the homeowner thanking us for the valiant effort displayed in saving her home.*" The pronoun *her* actually refers to the homeowner and SENTFINE used this clue and choose a correct-gender translation. DOCFINE choose a wrong gender, but consistent with the previous sentence. The meaning of the DOCFINE Czech translation was "*... a letter from the homeowner*$_{masc}$*, where he thanked us for the valiant effort which she displayed in saving her home*". So in addition to choosing a wrong gender, DOCFINE resolved the coreference incorrectly (*she* referring to something in previous sentences instead of to the homeowner) and identified incorrectly the agent of *saving*.[10]

The talk ended with a sentence "*Save the shoes*", which DOCFINE correctly translated as "*Zachraňte boty*" (*rescue the shoes*), again using the context of the previous sentences (although this time without any coreference). The translation chosen by SENTFINE – "*Šetřete si boty*" (*spare your shoes*) was incorrect in the context of a given talk.

### 5.3. Manual evaluation

It is well known that BLEU scores do not always correlate with human judgments [2, 17]. Especially, in the human-parity level of MT quality, it is obvious that any metric based on similarity to human references cannot measure the real translation quality.

We thus hired trained evaluators (native Czech speakers with a good knowledge of English) and conducted a manual evaluation using Direct Assessment [18]. We used a source-based variant (src-DA), which means that instead of the (human) reference translation, we showed the source sentence, so the results are not biased by any errors in the reference. It also allowed us to evaluate the quality of the reference as if it was another system. We also added two online systems into the comparison (anonymized as OnlineA and OnlineB, following WMT). We randomly sampled sentences from the tst-COMMON and tst-HE test sets. Each of the compared systems had 1311–1313 assessments.

| | src-DA | |
|---|---|---|
| system | Avg % | Avg z |
| SENTFINE | 88.3 | 0.212 |
| DOCFINE | 87.9 | 0.194 |
| SENTBASE | 87.6 | 0.176 |
| DOCBASE | 87.2 | 0.150 |
| Reference | 84.0 | −0.057 |
| OnlineB | 81.7 | −0.187 |
| OnlineA | 77.0 | −0.497 |

Table 7: Manual evaluation using source-based Direct Assessment on tst-COMMON and tst-HE. Significantly different scores ($p < 0.05$, Wilcoxon signed-rank test) are separated by a horizontal line.

Table 7 summarizes the results using both raw (Avg %) and normalized src-DA scores (Avg z, [18]). We can see that all four our systems were evaluated as significantly better than the reference and the two online systems. The differences in quality among our four systems are not significant (using standard p-value threshold 0.05 and Wilcoxon signed-rank test).

## 6. Conclusion

While the two fine-tuned (domain-adapted) systems scored significantly better than the two baseline systems in the automatic BLEU evaluation (Table 4), the difference was not evaluated as significant in the manual evaluation (Table 7). This could be explained by the observation (Section 5.1) that many of the domain-adaptation BLEU improvements are actually only typographic or other less important style-related differences. Nevertheless, fine-tuning still seems beneficial and for some purposes even the style consistency may be important (e.g. for decreasing the amount of human post-editing).

The results about the effect of document-level decoding are inconclusive. The document-level systems are insignificantly worse than the respective sentence-level systems according to the manual evaluation (Table 7). However, the pilot annotation (Section 5.2) showed several examples where the document-level system (DOCFINE) is better or more consistent than the sentence-level system (SENTFINE). A major weakness of our manual evaluation is that it was based on isolated sentences only, i.e. the evaluators did not see the document context. This setting is likely to bias the comparison of sentence-level and document-level systems. The

[10] In our pilot annotation, we counted this as two translation errors, although it could be considered also a single error or three errors, depending on the exact definition of "related differences" mentioned in the previous footnote. One could wonder why DOCFINE choose a wrong gender for the third and fourth occurrence when both occurrences were translated at once in a single multi-sentence sequence. We hypothesize that DOCFINE was confused by the third-occurrence sentence "*We took our treasures outside to the homeowner, where, not surprisingly, his received much more attention than did mine.*" and mis-interpreting the pronoun *his* as referring to the homeowner, while it was actually referring to one of the rescuers.

evaluators could not appreciate the improved consistency of DOCFINE relative to SENTFINE. It is also possible that the evaluators could judge a correct translation as worse than an incorrect translation in some cases.[11] We plan to conduct a proper document-level manual evaluation in future.

Finally, it is worth noticing that our systems were evaluated as substantially (4%) and significantly better than the human references. However, without further (document-level) manual evaluation, we cannot interpret this as reaching "human parity" or super-human quality.[12]

## 7. Acknowledgements

The work described in this paper was partially done during the first author's research stay in the Microsoft Translator team. The work has been supported by the "NAKI II – Systém pro trvalé uchování dokumentace a prezentaci historických pramenů z období totalitních režimů", project No. DG16P02B048, funded by the Ministry of Culture of the Czech Republic. The resources used are in part available from the LINDAT/CLARIAH-CZ repository, projects No. LM2015071 and LM2018101, which supported this work.

## 8. References

[1] H. Hassan, A. Aue, C. Chen, V. Chowdhary, J. Clark, C. Federmann, X. Huang, M. Junczys-Dowmunt, W. Lewis, M. Li, S. Liu, T. Liu, R. Luo, A. Menezes, T. Qin, F. Seide, X. Tan, F. Tian, L. Wu, S. Wu, Y. Xia, D. Zhang, Z. Zhang, and M. Zhou, "Achieving human parity on automatic chinese to english news translation," *CoRR*, vol. abs/1803.05567, 2018. [Online]. Available: http://arxiv.org/abs/1803.05567

[2] O. Bojar, C. Federmann, M. Fishel, Y. Graham, B. Haddow, M. Huck, P. Koehn, and C. Monz, "Findings of the 2018 conference on machine translation (wmt18)," in *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*. Belgium, Brussels: Association for Computational Linguistics, October 2018, pp. 272–307. [Online]. Available: http://www.aclweb.org/anthology/W18-6401

[3] M. Popel, "CUNI Transformer Neural MT System for WMT18," in *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*. Belgium, Brussels: Association for Computational Linguistics, October 2018, pp. 486–491. [Online]. Available: http://www.aclweb.org/anthology/W18-6424

[4] S. Läubli, R. Sennrich, and M. Volk, "Has machine translation achieved human parity? a case for document-level evaluation," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, Oct.-Nov. 2018, pp. 4791–4796. [Online]. Available: https://www.aclweb.org/anthology/D18-1512

[5] A. Toral, S. Castilho, K. Hu, and A. Way, "Attaining the unattainable? reassessing claims of human parity in neural machine translation," in *Proceedings of the Third Conference on Machine Translation, Volume 1: Research Papers*. Belgium, Brussels: Association for Computational Linguistics, October 2018, pp. 113–123. [Online]. Available: http://www.aclweb.org/anthology/W18-6312

[6] J. Tiedemann and Y. Scherrer, "Neural machine translation with extended context," in *Proceedings of the Third Workshop on Discourse in Machine Translation*. Copenhagen, Denmark: Association for Computational Linguistics, Sept. 2017, pp. 82–92. [Online]. Available: https://www.aclweb.org/anthology/W17-4811

[7] M. Junczys-Dowmunt, "Microsoft translator at wmt 2019: Towards large-scale document-level neural machine translation," in *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*. Florence, Italy: Association for Computational Linguistics, August 2019, pp. 225–233. [Online]. Available: http://www.aclweb.org/anthology/W19-5321

[8] M. Popel, D. Macháček, M. Auersperger, O. Bojar, and P. 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)*. Florence, Italy: Association for Computational Linguistics, August 2019, pp. 342–348. [Online]. Available: http://www.aclweb.org/anthology/W19-5337

[9] J. Zhang, H. Luan, M. Sun, F. Zhai, J. Xu, M. Zhang, and Y. Liu, "Improving the transformer translation model with document-level context," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, Oct.-Nov. 2018, pp. 533–542. [Online]. Available: https://www.aclweb.org/anthology/D18-1049

[10] C. Chu and R. Wang, "A survey of domain adaptation for neural machine translation," in *Proceedings of the 27th International Conference on Computational*

[11] For example, without any context, "*Šetřete si boty*" (*spare your shoes / use your shoes frugally*) seems to be a more plausible translation of "*Save the shoes*" than "*Zachraňte boty*" (*rescue the shoes*).

[12] Another issue is the quality of the references and whether it could be considered as indicative of "human quality". According to `https://www.ted.com/participate/translate/get-started`, the *volunteer translators should be fluently bilingual in both source and target languages* and *well-versed in the topics covered*. The translations *are reviewed by an experienced volunteer* and *approved by a TED Language Coordinator or staff member*.

*Linguistics*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, Aug. 2018, pp. 1304–1319. [Online]. Available: https://www.aclweb.org/anthology/C18-1111

[11] M.-T. Luong and C. D. Manning, "Stanford neural machine translation systems for spoken language domain," in *International Workshop on Spoken Language Translation*, Da Nang, Vietnam, 2015.

[12] O. Bojar, O. Dušek, T. Kocmi, J. Libovický, M. Novák, M. Popel, R. Sudarikov, and D. Variš, "CzEng 1.6: Enlarged Czech-English Parallel Corpus with Processing Tools Dockered," in *Text, Speech, and Dialogue: 19th International Conference, TSD 2016*, ser. Lecture Notes in Artificial Intelligence, P. Sojka, A. Horák, I. Kopeček, and K. Pala, Eds., no. 9924, Masaryk University. Springer International Publishing, 2016, pp. 231–238.

[13] M. A. Di Gangi, R. Cattoni, L. Bentivogli, M. Negri, and M. Turchi, "MuST-C: a Multilingual Speech Translation Corpus," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, Minneapolis, MN, USA, June 2019.

[14] A. Vaswani, S. Bengio, E. Brevdo, F. Chollet, A. N. Gomez, S. Gouws, L. Jones, L. Kaiser, N. Kalchbrenner, N. Parmar, R. Sepassi, N. Shazeer, and J. Uszkoreit, "Tensor2tensor for neural machine translation," *CoRR*, vol. abs/1803.07416, 2018. [Online]. Available: http://arxiv.org/abs/1803.07416

[15] M. Popel and O. Bojar, "Training Tips for the Transformer Model," *The Prague Bulletin of Mathematical Linguistics*, vol. 110, pp. 43–70, April 2018. [Online]. Available: https://ufal.mff.cuni.cz/pbml/110/art-popel-bojar.pdf

[16] M. Post, "A Call for Clarity in Reporting BLEU Scores," *CoRR*, vol. arXiv/1804.08771, Apr. 2018. [Online]. Available: http://arxiv.org/abs/1804.08771

[17] L. Barrault, O. Bojar, M. R. Costa-jussà, C. Federmann, M. Fishel, Y. Graham, B. Haddow, M. Huck, P. Koehn, S. Malmasi, C. Monz, M. Müller, S. Pal, M. Post, and M. Zampieri, "Findings of the 2019 conference on machine translation (wmt19)," in *Proceedings of the Fourth Conference on Machine Translation, Volume 2: Shared Task Papers*. Florence, Italy: Association for Computational Linguistics, August 2019.

[18] Y. Graham, T. Baldwin, A. Moffat, and J. Zobel, "Can machine translation systems be evaluated by the crowd alone," *Natural Language Engineering*, vol. FirstView, pp. 1–28, 1 2016. [Online]. Available: http://journals.cambridge.org/article_S1351324915000339