CUNI English-Czech and English-Polish Systems in WMT20: Robust Document-Level Training

Martin Popel
Charles University, Faculty of Mathematics and Physics,
Institute of Formal and Applied Linguistics,
Malostranské náměstí 25, 118 00 Prague, Czech Republic
popel@ufal.mff.cuni.cz

Abstract

We describe our two NMT systems submitted to the WMT 2020 shared task in English↔Czech and English↔Polish news translation. One system is sentence level, translating each sentence independently. The second system is document level, translating multiple sentences, trained on multi-sentence sequences up to 3000 characters long.

1 Introduction

In this paper, we describe our two NMT systems submitted to the WMT 2020 news translation shared task: "CUNI-Transformer" (Charles University Transformer, sentence-level) and "CUNI-DocTransformer" (document-level). We trained them for English↔Czech and the former one also for English↔Polish (no parallel document-level data was provided for English-Polish, thus we could not train the latter one).

2 Common settings

Both our systems are implemented in the Tensor2Tensor framework (Vaswani et al., 2018) and have the same Transformer (Vaswani et al., 2017) architecture – transformer_big with 12 encoder layers instead of the default 6 (while keeping 6 layers in the decoder). The 32k joint English-Czech subword vocabulary is exactly the same as used by Popel (2018) and Popel et al. (2019), which are the systems we submitted to WMT in the last two years. Also most of the hyperparameters (except for the encoder depth) and the training regime are the same.

The main improvement of our sentence-level system relative to our last-year submission stems from using slightly larger and better-filtered training data – CzEng 2.0 (Kocmi et al., 2020b) with 61M authentic parallel and 127M synthetic (back-translated) sentences (see Table 1), instead of CzEng 1.7 with 57M authentic parallel sentences.

data set	sentence pairs (M)	words (M) EN	CS
authentic	61	617	702
EN-mono (NewsCrawl 2016–2018)	76	1296	1474
CS-mono (NewsCrawl 2013–2018)	51	700	833
total	188	2613	3009

Table 1: Training data sizes (in millions). All the data are taken from CzEng 2.0.

We also enlarged our development-test set: we concatenated WMT newstest 2008–2018, instead of using newstest2016 only. WMT news tests before 2020 did not have paragraph boundaries marked. We thus prepared a version of our dev-set where we joined together several consecutive sentences randomly (except for titles not ended by a punctuation) to simulate WMT2020 paragraph-level setting.

Our document-level system was further improved as described in Sections 3 and 4.

3 Document-level training

Our last-year document-level submission (Popel et al., 2019) introduced a method of training-data context augmentation, where multiple consecutive sentences (within original documents) are merged together into multi-sentence sequences (of parallel source-target data). The sentences within each sequence are separated with a special token, so that we can easily extract the sentence alignment after decoding. The length of the sequences was limited by 1000 characters and 200 subwords (i.e. any sequences longer than any of the limits in either source or target were discarded from training).

An important aspect of this method is that it extracts all possible sequences from the document-level training data. For example, given a document

Proceedings of the 5th Conference on Machine Translation (WMT), pages 269–273
Online, November 19–20, 2020.

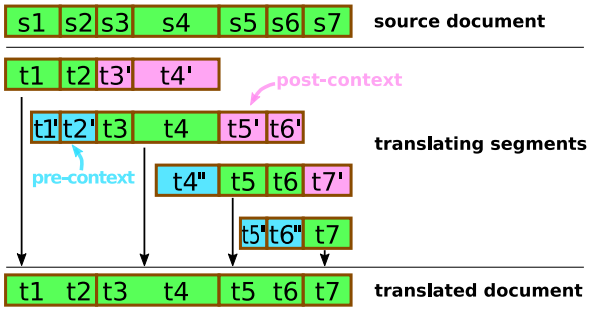

Figure 1: Decoding overlapping multi-sentence sequences with our document-level model. Note that the pre-context part may start not only on sentence boundaries (it improves the results slightly according our initial experiments).

with 5 sentences s1–s5, we extract sequences s1, s1–2, s1–s3, s2, s2–s3, s2–s4, s3, s3–s4, s3–s5, s4, s4–s5 and s5, while ignoring sequences s1–s4, s1–s5 and s2–s5 because these are longer than the limits. Note that this way of context augmentation implicitly upsamples sentences from longer documents relative to sentences from shorter documents. A sentence appearing in N windows of at most 1000 characters is present N times in the augmented training data.

Thus this year, we simply sample non-overlapping sequences of sentences: s1–s3, s4–s5. There are many documents shorter than the limits in CzEng 2.0, including many single-sentence documents (from sources without document-level annotation). Thus, there are naturally occurring trainng sequences which are shorter than the limits and we checked the model is capable of translating also single sentences.

In addition to this change in data sampling, we increased the sequence length limit to 3000 characters and 750 subwords.

4 Document-level decoding

There are many possible ways how to use document-level models (trained as described in the previous section) at decode time.

- We can translate single sentences, thus not using the advantage of document-level training. This may serve as a baseline for comparison with document-level decoding and we used this for our last-year sentence-level submission "Transformer T2T 2019" (Popel et al., 2019).

- We can split each input document into non-overlapping multi-sentence sequences.

- We can split each input document into overlapping multi-sentence sequences (with so-called pre-context and post-context parts, which are ignored in the final translation) as suggested by Popel et al. (2019) and explained in Figure 1.

- We can use the overlapping sequences for some kind of consensus decoding or ensembling. We have not tried this option yet.

Because of the increased limits of training sequences, we increased also the decoding limits two times: pre-context of up to 400 characters, main content of up to 1000 characters and post-context of up to 1800 characters minus the length of the pre-context and main content.

5 Robust training with noising

To make the model more robust to real-world user-generated data, we added a noise to the training data. We followed an approach of Náplava and Straka (2019) and made the source side of the training data more noisy by introducing both grammatical and spelling errors. The basic set of noising operations introducing grammatical errors consisted of the following operations: token replacement with one of its spelling dictionary proposals, token deletion and insertion and swapping of two nearby words. Moreover, we also allowed to replace phrases with one of their most frequent variants, add or delete punctuation and allowed to strip diacritics.

We applied this technique only to our Czech→English sentence-level system by noising the source=Czech side of both the authentic and synthetic parallel data. In preliminary experiments, we observed substantial improvements on artificially noised dev sets, but slight worsenings on WMT dev sets, which contain just a very small amount of typos and other errors (on the source side). We thus decided to mix the noised training data with the original unnoised data 1:1. This resulted in approximately the same BLEU on the original dev sets as without noising, while keeping the improved results on artificially noised dev sets.

For time constraints, we decided to not use any noising in the document-level Czech→English training, as well as in our English→Czech and English↔Polish systems.

6 Results

In Tables 2–5, we report BLEU scores on the newstest2020 for all the systems submitted to WMT.[1] For English→Czech and English→Polish (Tables 2 and 4), we report also gender coreference accuracy scores based on WinoMT testset results (Kocmi et al., 2020a). For English→Czech, we report also manual document-level quality evaluation by Zouhar et al. (2020) of 269 WMT2020 test-suites sentences (i.e. not sentences from newstest2020). The official manual evaluation on newstest2020 is not available yet.

We can see that while our DocTransformer was not the best system according to BLEU, it scored well according to the other two reported metrics, being the best system in English→Czech and the second-best in English→Polish. This could be caused by the low reliability of BLEU (and other metrics based on similarity with reference) for high-quality MT, or by the domain mismatch – the test-suites contain also other domains than news. Unfortunatelly, we cannot answer this question before further analysis using the official WMT manual evaluation, once it is done and published.

Finally, we present several translation examples in the Appendix. The source English documents were taken from the WMT2019 newstest and the same examples were selected already by Popel et al. (2019). Table 6 shows three examples where the document-level model corrects a lexical error of our 2019 sentence-level model. Interestingly, two of these errors were fixed also by our this-year sentence-level model, showing that the cross-sentence context is not *necessary* for correct translation of these examples. Table 7 shows an error of our 2019 document-level model, which is not present in this-year models.

7 Conclusion

We suceeded to improve our baseline system CUNI-T2T-2018 (Popel et al., 2019) by using better training data, doubling the encoder depth (to 12 layers) and by robust training with source-side noising. While all these three techniques are well-known, we show improvements in improving the last-year WMT state of the art in English-Czech translation. We improved also our document-level system (CUNI-DocTransformer) by more careful data sam-

[1]The SacreBLEU signature is BLEU+case.mixed+lang.$src-$trg+numrefs.1+smooth.exp+test.wmt20+tok.13a+version.1.4.13.

system	BLEU cased	g. coref accuracy	TS fluency × adequacy
Online-B	**41.11**	(11) 56.9	(4) 83.3
OPPO	36.78	(3) 78.7	(2) 84.2
SRPOL	36.46	(2) 81.2	(5) 82.2
UEDIN-CUNI	36.27	(6) 72.5	(8) 79.5
CUNI-DocTransformer	35.67	**(1) 83.6**	**(1) 85.1**
eTranslation	35.67	(8) 70.9	(7) 80.5
CUNI-Transformer	35.40	(4) 78.0	(3) 83.4
CUNI-T2T-2018	35.08	(5) 77.6	(6) 81.0
Online-A	30.84	(9) 63.3	(9) 78.9
Online-Z	27.96	(7) 72.2	(10) 72.8
Online-G	25.28	(10) 62.0	(11) 71.7
zlabs-nlp	20.25	(12) 49.9	(12) 64.5

Table 2: Evaluation of English→Czech WMT20 systems. The systems are ordered by BLEU, ordering by the other metrics is provided in parentheses. The gender coreference accuracy scores are based on the WinoMT testset results (Kocmi et al., 2020a). The "TS fluency × adequacy" score is based on manual document-level quality evaluation (Zouhar et al., 2020).

system	BLEU cased
OPPO	**29.91**
CUNI-DocTransformer	29.22
Online-B	28.66
CUNI-Transformer	28.55
SRPOL	28.51
UEDIN-CUNI	27.66
Online-A	26.84
CUNI-T2T-2018	26.08
PROMT_NMT	25.57
Online-G	23.91
Online-Z	23.25
zlabs-nlp	21.76

Table 3: Evaluation of Czech→English WMT20 systems.

system	BLEU cased	g. coref accuracy
SRPOL	**27.56**	**(1) 71.2**
eTranslation	27.20	(3) 68.8
Huoshan_Translate	26.09	(8) 65.7
OPPO	25.49	(4–5) 68.2
SJTU-NICT	25.45	(4–5) 68.2
Online-B	25.17	(12) 57.7
Tilde (1430)	24.93	(9) 64.8
NICT_Kyoto	24.91	(10) 64.2
Tilde (1425)	24.87	(11) 63.3
CUNI-Transformer	24.76	(2) 69.8
Online-G	23.73	(6) 67.3
Online-A	23.71	(13) 53.7
Online-Z	20.75	(7) 65.9
zlabs-nlp	18.64	(14) 46.1

Table 4: Evaluation of English→Polish WMT20 systems.

system	BLEU cased
NICT-Rui	**34.55**
Huoshan_Translate	34.44
SRPOL	34.26
Online-B	33.92
OPPO	32.46
SJTU-NICT	32.16
CUNI-Transformer	31.90
NICT_Kyoto	31.85
Online-A	31.79
PROMT_NMT	31.19
Tilde	30.20
Online-G	29.86
Online-Z	28.64
zlabs-nlp	27.77

Table 5: Evaluation of Polish→English WMT20 systems.

pling which is not biased towards sentences from longer documents.

Acknowledgments

This work has been supported by the grant GX20-16819X (LUSyD) of the Grant Agency of the Czech Republic and by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (LM2015071). The work has been using language resources developed and distributed by the LINDAT/CLARIAH-CZ project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2018101).

References

Tom Kocmi, Tomasz Limisiewicz, and Gabriel Stanovsky. 2020a. Gender Coreference and Bias Evaluation at WMT 2020. Submitted to WMT2020.

Tom Kocmi, Martin Popel, and Ondrej Bojar. 2020b. Announcing CzEng 2.0 Parallel Corpus with over 2 Gigawords. *arXiv preprint arXiv:2007.03006*.

Jakub Náplava and Milan Straka. 2019. Grammatical error correction in low-resource scenarios. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 346–356.

Martin Popel. 2018. CUNI Transformer Neural MT System for WMT18. In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 486–491, Belgium, Brussels. Association for Computational Linguistics.

Martin Popel, Dominik Macháček, Michal Auersperger, Ondřej Bojar, and Pavel Pecina. 2019. 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. Association for Computational Linguistics.

Ashish Vaswani, Samy Bengio, Eugene Brevdo, Francois Chollet, Aidan N. Gomez, Stephan Gouws, Llion Jones, Łukasz Kaiser, Nal Kalchbrenner, Niki Parmar, Ryan Sepassi, Noam Shazeer, and Jakob Uszkoreit. 2018. Tensor2tensor for neural machine translation. *CoRR*, abs/1803.07416.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, pages 6000–6010. Curran Associates, Inc.

Vilem Zouhar, Tereza Vojtechova, and Ondrej Bojar. 2020. WMT20 Document-Level Markable Error Exploration. Submitted to WMT2020.

8 Appendix

source	[...] to meet Craig Halkett's header across goal. The hosts were content to let Rangers play in front of them, knowing they could trouble the visitors at set pieces. And that was the manner in which the crucial **goal** came. Rangers conceded a free-kick [...]
T2T-2019-sent	A to byl způsob, jakým přišel rozhodující **cíl** (*aim*).
others	A to byl způsob, jakým přišel rozhodující **gól** (*goal*).
source	Elizabeth Warren Will Take "Hard Look" At Running For President in 2020, Massachusetts Senator Says Massachusetts Senator Elizabeth Warren said on Saturday she would take a "hard look" at running for president following the midterm elections. During a town hall in Holyoke, Massachusetts, Warren confirmed she'd consider **running**. "It's time for women to go to Washington and fix our broken government and that includes a woman at the top," she said, according to The Hill. [...]
T2T-2019-sent	Na radnici v Holyoke v Massachusetts Warrenová potvrdila, že uvažuje o **útěku** (*escape*).
T2T-2019-doc	Na radnici v Holyoke ve státě Massachusetts Warrenová potvrdila, že o **kandidatuře** (*candidacy*) uvažuje.
T2T-2020-sent	Na radnici v Holyoke v Massachusetts Warrenová potvrdila, že zváží **kandidaturu** (*candidacy*).
T2T-2020-doc	Během jednání na radnici v Holyoke ve státě Massachusetts Warrenová potvrdila, že o **kandidatuře** (*candidacy*) bude uvažovat.
source	At 6am, just as Gegard Mousasi and Rory MacDonald were preparing to face each other, viewers in the UK were left stunned when the coverage changed to Peppa Pig. Some were unimpressed after they had stayed awake until the early hours especially for the **fight**. [...]
T2T-2019-sent	Na některé to neudělalo žádný dojem, když zůstali vzhůru až do časných ranních hodin, zvláště kvůli **rvačce** (*brawl*).
T2T-2019-doc	Na některé to neudělalo žádný dojem, když zůstali vzhůru až do ranních hodin, zejména kvůli **zápasu** (*match*).
T2T-2019-sent	Na některé to neudělalo žádný dojem poté, co zůstali vzhůru až do časných ranních hodin, zejména kvůli **boji** (*combat*).
T2T-2020-doc	Někteří nebyli ohromeni poté, co zůstali vzhůru až do ranních hodin, zejména kvůli **zápasu** (*match*).

Table 6: Three examples of errors by T2T-2019-sent (2019 sentence-level model) corrected by the document-level models and in the first two examples also by T2T-2020-sent (sentence-level CUNI-Transformer from this paper).

source	New cancer vaccine can teach the immune system to 'see' rogue cells New cancer vaccine can teach the immune system to 'see' rogue cells and kill them Vaccine teaches immune system to recognise rogue cells as part of treatment Method involves extracting immune cells from a **patient**, altering them in lab They can then 'see' a protein common to many cancers and then reinjected A trial vaccine is showing promising results in **patients** with a range of cancers. One woman treated with the vaccine, which teaches the immune system to recognise rogue cells, saw her ovarian cancer disappear for more than 18 months. The method involves extracting immune cells from a **patient**, altering them in the laboratory so they can "see" a protein common to many cancers called HER2, and then reinjecting the cells.
T2T-2019-sent	[...] buněk z **pacienta** [...] výsledky u **pacientů** [...] buněk z **pacienta** [...]
T2T-2019-doc	[...] buněk z **pacienta** [...] výsledky u **pacientů** [...] buněk od **pacientky** (*female patient*) [...]
T2T-2020-sent	Nová protinádorová vakcína může naučit imunitní systém „vidět" nepoctivé buňky Nová vakcína proti rakovině může naučit imunitní systém „vidět" buňky darebáků a zabít je Vakcína učí imunitní systém rozpoznat nepoctivé buňky jako součást léčby Metoda zahrnuje extrakci imunitních buněk z **pacienta**, jejich úpravu v laboratoři Mohou pak „vidět" bílkovinu, která je společná mnoha druhům rakoviny a pak ji znovu nasadit Zkušební vakcína vykazuje slibné výsledky u **pacientů** s řadou nádorových onemocnění. Jedna žena léčená vakcínou, která učí imunitní systém rozpoznávat nepoctivé buňky, se postarala o to, že jí na více než 18 měsíců zmizela rakovina vaječníků. Metoda spočívá v extrakci imunitních buněk z **pacienta**, jejich modifikaci v laboratoři, aby mohli „vidět" bílkovinu společnou mnoha druhům rakoviny zvanou HER2, a následném reinjekci buněk.
T2T-2020-doc	Nová protinádorová vakcína může naučit imunitní systém „vidět" darebácké buňky Nová protinádorová vakcína může naučit imunitní systém „vidět" darebácké buňky a zabít je Vakcína učí imunitní systém rozpoznávat darebácké buňky jako součást léčby Metoda spočívá v odebrání imunitních buněk z **pacienta**, jejich pozměnění v laboratoři Mohou pak „vidět" bílkovinu společnou pro mnoho druhů rakoviny a poté znovu použít Zkušební vakcína vykazuje slibné výsledky u **pacientů** s řadou druhů rakoviny. Jedna žena léčená touto vakcínou, která učí imunitní systém rozpoznávat darebácké buňky, viděla, jak její rakovina vaječníků zmizela na více než 18 měsíců. Metoda spočívá v odebrání imunitních buněk z **pacienta**, jejich pozměnění v laboratoři tak, aby mohly „vidět" bílkovinu společnou pro mnoho druhů rakoviny nazývanou HER2, a poté znovu použít buňky.

Table 7: Example of an inconsistency error by T2T-2019-doc. The other three models are consistent.