CUNI Transformer Neural MT System for WMT18

Martin Popel
Charles University, Faculty of Mathematics and Physics,
Institute of Formal and Applied Linguistics,
Prague, Czechia
popel@ufal.mff.cuni.cz

Abstract

We describe our NMT system submitted to the WMT2018 shared task in news translation. Our system is based on the Transformer model (Vaswani et al., 2017). We use an improved technique of backtranslation, where we iterate the process of translating monolingual data in one direction and training an NMT model for the opposite direction using synthetic parallel data. We apply a simple but effective filtering of the synthetic data. We pre-process the input sentences using coreference resolution in order to disambiguate the gender of pro-dropped personal pronouns. Finally, we apply two simple post-processing substitutions on the translated output.

Our system is significantly ($p < 0.05$) better than all other English-Czech and Czech-English systems in WMT2018.

1 Introduction

The quality of Neural Machine Translation (NMT) depends heavily on the amount and quality of the training parallel sentences as well as on various training tricks, which are sometimes surprisingly simple and effective.

In this paper, we describe our NMT system "CUNI Transformer" (Charles University version of Transformer), submitted to the English→Czech and Czech→English news translation shared task within WMT2018. We describe five techniques, which helped to improve our system, so that it outperformed all other systems in these two translation directions: training data filtering (Section 3), improved backtranslation (Section 4), tuning two separate models based on the original language of the text to be translated (Section 5), coreference pre-processing (Section 6) and post-processing using regular expressions (Section 7). Our system significantly outperformed all other systems in WMT2018 evaluation (Section 8).

data set	sentence pairs (k)	words (k) EN	words (k) 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
EN NewsCrawl 2016–17	47 483	934 981	
CS NewsCrawl 2007–17	65 383		927 348
total	170 951	1 576 923	1 490 516

Table 1: Training data sizes (in thousands).

2 Experimental Setup

Our training data is constrained to the data allowed in the WMT2018 shared task. Parallel (authentic) data are: CzEng 1.7, Europarl v7, News Commentary v11 and CommonCrawl. In our backtranslation experiments (Section 4), we used synthetic data translated by backtranslation of monolingual data: Czech and (a subset of) English NewsCrawl articles. We filtered out ca. 3% of sentences from the synthetic data (Section 3). Data sizes are reported in Table 1.

Note that usually the amount of available monolingual data is orders of magnitude larger than the available parallel data, but in our case it is comparable (58M parallel vs. 65M/48M monolingual). We used all the Czech monolingual data allowed in the constrained task.

We used the Transformer self-attentional sequence-to-sequence model (Vaswani et al., 2017) implemented in the Tensor2Tensor framework.[1] We followed the training setup and tips of Popel and Bojar (2018), but we trained our models with the Adafactor optimizer (Shazeer and Stern, 2018) instead of the default Adam: We used T2T version 1.6.0, `transformer_big` and hyper-parameters `learning_rate_schedule=rsqrt_decay`,

[1] https://github.com/tensorflow/tensor2tensor

Proceedings of the Third Conference on Machine Translation (WMT), Volume 2: Shared Task Papers, pages 482–487
Belgium, Brussels, October 31 - Novermber 1, 2018. ©2018 Association for Computational Linguistics
https://doi.org/10.18653/v1/W18-64051

`learning_rate_warmup_steps=8000`, `batch_size=2900`, `max_length=150`, `layer_prepostprocess_dropout=0`, `optimizer=Adafactor`. For decoding, we used `alpha=1`.

We stored model checkpoints each hour and averaged the last eight checkpoints. We used eight GTX 1080 Ti GPUs.

3 Training Data Filtering

We found out that the Czech monolingual data set (NewsCrawl 2007–2017) contains many English sentences. Those sentences were either kept untranslated or paraphrased when preparing the synthetic data with backtranslation. Thus the synthetic data included many English-English sentence pairs. Consequently, the synth-trained models had a higher probability of keeping a sentence untranslated.

In order to filter out the English sentences from the Czech data, we kept only sentences containing at least one accented character.[2] We also filtered out sentences longer than 500 characters from the synthetic data. Most of these sentences would be ignored anyway because we are training our Transformer with `max_length`=150, i.e. filtering out sentences longer than 150 subwords (cf. Popel and Bojar, 2018, § 4.4). Sometimes a Czech sentence was much shorter than its English translation (especially for the translations by Nematus2016) – because of filler words repeated many times, which is a well-known problem of NMT systems (e.g. Sudarikov et al., 2016). We filtered out all sentences with a word (or a pair of words) repeated more than twice using a regular expression `/ (\S+ ?\S+) \1 \1 /`. This way, we filtered out ca. 3% of sentences and re-trained our systems. After this filtering, we did not observe any untranslated sentences in the synth-trained output.

4 Improved Backtranslation

Sennrich et al. (2016b) introduced backtranslation as a simple way how to utilize target-language monolingual data in NMT. The monolingual data sets are translated (by a target-to-source MT system) to the source language, resulting in *synthetic* parallel data, which is used as additional training data (in addition to *authentic* parallel) for the final (source-to-target) NMT system.

Sennrich et al. (2017) compared two regimes of how to incorporate synthetic training data created using backtranslation of monolingual data. In the *fine-tuned* regime, a system is trained first on the authentic parallel data and then after several epochs it is trained on a 1:1 mix of authentic and synthetic data. In the *mixed* regime, the 1:1 mixed data is used from the beginning of training. In both cases, the 1:1 mix means shuffling the data randomly at the sentence level, possibly oversampling the smaller of the two data sources.

We used a third approach, termed *concat* regime, where the authentic and synthetic parallel data are simply concatenated (without shuffling). We observed that this regime leads to improvements in translation quality relative to both *mixed* and *fine-tuned* regimes, especially when checkpoint averaging is used.

For obtaining the final English→Czech system, we iterated the backtranslation process:

1. We downloaded the Nematus2016 models trained by Sennrich et al. (2016a) using *fine-tuned* backtranslation of English NewsCrawl 2015 articles, which were translated "*with an earlier NMT model trained on WMT15 data*" (Sennrich et al., 2016a). We used these Nematus2016 models to translate Czech NewsCrawl 2007–2017 articles to English.

2. We trained an English→Czech Transformer on this data (filtered as described in Section 3) using concat backtranslation with checkpoint averaging. We used this Transformer model to translate English NewsCrawl 2016–2017 articles into Czech.

3. We trained our Czech→English Transformer model (used for our WMT18 submission) on this data using concat backtranslation with averaging. We translated Czech NewsCrawl 2016–2017 articles into English using this system, producing a higher-quality synthetic data than in step 1 (but smaller because of lack of time and resources).

4. We trained our final English→Czech system

[2] `m/[ěščřžýáíéúůd’t’ň]/i` – this simple heuristics is surprisingly effective for Czech. In addition to English sentences, it filters out also *some* short Czech sentences, sentences in other languages (e.g. Chinese) and various "non-linguistic" content, such as lists of football or stock-market results.

on this data, again using concat backtranslation with averaging.

Each training (steps 2, 3 and 4) took eight days on eight GPUs. Translating the monolingual data with Nematus2016 (step 1) took about two weeks and with our Transformer models (steps 2 and 3) took about five days. The final model trained in step 4 is +0.83 BLEU better than the model trained in step 2 without data filtering, as measured on newstest2017 (cf. Table 2).

5 CZ/nonCZ Tuning

In WMT test sets since 2014, half of the sentences for a language pair X-EN originate from English news servers (e.g. bbc.com) and the other half from X-language news servers. All WMT test sets include the server name for each document in metadata, so we were able to split our test set (and dev set newstest2013) into two parts: `CZ` (for Czech-domain articles, i.e. documents with `docid` containing ".cz") and `nonCZ` (for non-Czech-domain articles). We noticed that when training on synthetic data, the model performs much better on the `CZ` test set than on the `nonCZ` test set. When trained on authentic data, it is the other way round. Intuitively, this makes sense: The target side of our synthetic data are original Czech sentences from Czech newspapers, similarly to the `CZ` test set. In our authentic data, over 90% of sentences were originally written in English about "non-Czech topics" and translated into Czech (by human translators), similarly to the `nonCZ` test set. There are two closely related phenomena: a question of domain (topics) in the training data and a question of so-called *translationese* effect, i.e. which side of the parallel training data (and test data) is the original and which is the translation.

Based on these observations, we prepared a `CZ`-tuned model and a `nonCZ`-tuned model. Both models were trained in the same way, they differ only in the number of training steps. For the `CZ`-tuned model, we selected a checkpoint with the best performance on `wmt13-CZ` (Czech-origin portion of newstest2013), which was at 774k steps. Similarly, for the `nonCZ`-tuned model, we selected the checkpoint with the best performance on `wmt13-nonCZ`, which was at 788k steps. Note that both the models were trained jointly in one experiment, just selecting checkpoints at two different moments.

6 Coreference Pre-processing

In Czech, as a pro-drop language, it is common to omit personal pronouns in subject positions. Usually, the information about gender and number of the subject is encoded in the verb inflection, but present-tense verbs have the same form for the feminine and masculine gender. For example, "*Není doma*" can mean either "*She is not home*" or "*He is not home*". When translating such sentences from Czech to English, we must use the context of neighboring sentences in a given document, in order to disambiguate the gender and select the correct translation. However, our Transformer system (similarly to most current NMT systems) translates each sentence independently of other sentences. We observed that in practice it always prefers the masculine gender if the information about gender could not be deduced from the source sentence.

We implemented a simple pre-processing of the Czech sentences, which are then translated with our Czech→English Transformer system – we inserted pronoun *ona* (*she*), where it was "missing". We analyzed the source Czech documents in the Treex NLP framework (Popel and Žabokrtský, 2010), which integrates a coreference resolver (Novák, 2017). We found sentences where a female-gender pronoun subject was dropped and the coreference link was pointing to a different sentence (usually the previous one). We restricted the insertion of *ona* only to the cases in which the antecedent in the coreference chain represents a human (i.e. excluding grammatical-only female gender of inanimate objects and animals). We used a heuristic detection of human entities, which is integrated in Treex.

This preprocessing affected only 1% of sentences in our nestest2017 dev set and for most of them the English translation was improved (according to our judgment), although the overall BLEU score remained the same. We consider this solution as a temporary workaround before document-level NMT (e.g. Kuang et al., 2017) is available in T2T. That said, the advantage of the described preprocessing is that it can be applied to any (N)MT system – without changing its architecture and even without retraining it.

7 RegEx Post-processing

We applied two simple post-processings to the translations, using regular expressions.

English→Czech	BLEU	BLEU	chrF2
system	cased	uncased	cased
Nematus (Sennrich et al., 2016b)	22.80	23.29	0.5059
T2T (Popel and Bojar, 2018)	23.84	24.40	0.5164
our mixed backtranslation	24.85 (+1.01)	25.33	0.5267
our concat backtranslation	25.77 (+0.92)	26.29	0.5352
+ higher quality backtranslation	26.60 (+0.83)	27.10	0.5410
+ CZ/nonCZ tuning	**26.81** (+0.21)	**27.30**	**0.5431**

Table 2: Automatic evaluation on (English→Czech) `newstest2017`. The three scores in parenthesis show BLEU difference relative to the previous line.

We deleted phrases repeated more than twice (immediately following each other); we kept just the first occurrence. We considered phrases of one up to four words. With the training-data filtering described in Section 3, less than 1% sentences needed this post-processing.

For English→Czech, we converted quotation symbols in the translations to the correct-Czech „lower and upper“ quotes using two regexes: ``` s/(^|[({[])("|,,|''|``)/$1„/g ``` and ``` s/("|'')($|[,.?!:;)}\]])/“$2/g ```. In English, the distinction between "straight" and “curly” quotes is considered as a rather typographical (or style-related) issue. However, in Czech, a mismatch between lower (opening) and upper (closing) quotes is considered as an error in formal writing.

8 Evaluation

8.1 WMT2017 Evaluation

Table 2 evaluates the relative improvements described in Sections 4 and 5 on English→Czech newstest2017 and compares the results with the WMT2017 winner – Nematus (Sennrich et al., 2016b), and with the result of Popel and Bojar (2018) – T2T without any backtranslation.

The three reported automatic metrics are: case-sensitive (cased) BLEU, case-insensitive (uncased) BLEU and a character-level metric chrF2 (Popović, 2015). We compute all the three metrics with sacreBLEU (Post, 2018). The reported cased and uncased variants of BLEU differ also in the tokenization. The *cased* variant uses the default (ASCII-only) for better comparability with the results at `http://matrix.statmt.org`. The *uncased* variant uses the international tokenization, which has higher correlation with humans (Macháček and Bojar, 2013). The sacreBLEU signatures of the three metrics are:

- `BLEU+case.mixed+lang.en-cs+numrefs.1+smooth.exp+test.wmt17+tok.13a`,
- `BLEU+case.lc+lang.en-cs+numrefs.1+smooth.exp+test.wmt17+tok.intl` and
- `chrF2+case.mixed+lang.en-cs+numchars.6+numrefs.1+space.False+test.wmt17`.

We performed a small-scale manual evaluation on newstest2017 and noticed that in many cases the human reference translation is actually worse than our Transformer output. Thus the results of BLEU (or any other automatic metric comparing similarity with references) may be misleading.

8.2 WMT2018 Evaluation

Table 3 the reports results of all English↔Czech systems submitted to WMT2018, according to both automatic and manual evaluation. For the automatic evaluation, we use the same three metrics as in the previous section (just with `wmt18` instead of `wmt17`). For the manual evaluation, we report the reference-based direct assessment (refDA) scores, provided by the WMT organizers.

Our Transformer is the best system in English→Czech and Czech→English WMT2018 news task. It is significantly ($p < 0.05$) better than the second-best system – UEdin NMT, in both translation directions and both according to BLEU bootstrap resampling test (Koehn, 2004) and according to refDA Wilcoxon rank-sum test.

9 Conclusion

We have presented five simple but effective techniques for improving (N)MT quality. All five tech-

	English→Czech				Czech→English			
	BLEU	BLEU	chrF2	refDA	BLEU	BLEU	chrF2	refDA
system	uncased	cased	cased	Ave. %	uncased	cased	cased	Ave. %
our Transformer	**26.82**	**26.01**	**0.5372**	**67.2**	**35.64**	**33.91**	**0.5876**	**71.8**
UEdin NMT	24.30	23.42	0.5166	60.6	34.12	33.06	0.5801	67.9
Online-B	20.16	19.45	0.4854	52.1	33.58	31.78	0.5736	66.6
Online-A	16.84	15.74	0.4584	46.0	28.47	26.78	0.5447	62.1
Online-G	16.33	15.11	0.4560	42.0	25.20	22.53	0.5310	57.5

Table 3: WMT2018 automatic (BLEU, chrF2) and manual (refDA = reference-based direct assessment) evaluation on `newstest2018`.

niques can be applied to virtually any NMT system. According to the preliminary results of the manual evaluation, the final translation quality is comparable to or even better than the quality of human references.

As a future work, we would like to assess the relative improvement of each of the five techniques based on manual evaluation (because automatic single-reference evaluation is not reliable when the MT quality is near to the quality of reference translations).

Acknowledgements

The work described in this paper has been supported by the "NAKI II - Systém pro trvalé uchování dokumentace a prezentaci historichých pramenů z období totalitních režimů", project No. DG16P02B048, funded by the Ministry of Culture of the Czech Republic.

References

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of EMNLP*, volume 4, pages 388–395.

Shaohui Kuang, Deyi Xiong, Weihua Luo, and Guodong Zhou. 2017. Cache-based Document-level Neural Machine Translation. *CoRR*, arXiv/1711.11221.

Matouš Macháček and Ondřej Bojar. 2013. Results of the WMT13 Metrics Shared Task. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 45–51, Sofia, Bulgaria. Association for Computational Linguistics.

Michal Novák. 2017. Coreference resolution system not only for czech. In *Proceedings of the 17th conference ITAT 2017: Slovenskočeský NLP workshop (SloNLP 2017)*, pages 193–200, Praha, Czechia. CreateSpace Independent Publishing Platform.

Martin Popel and Ondřej Bojar. 2018. Training Tips for the Transformer Model. *The Prague Bulletin of Mathematical Linguistics*, 110:43–70.

Martin Popel and Zdeněk Žabokrtský. 2010. TectoMT: Modular NLP Framework. *Advances in Natural Language Processing*, pages 293–304.

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

Matt Post. 2018. A Call for Clarity in Reporting BLEU Scores. *CoRR*, arXiv/1804.08771.

Rico Sennrich, Alexandra Birch, Anna Currey, Ulrich Germann, Barry Haddow, Kenneth Heafield, Antonio Valerio Miceli Barone, and Philip Williams. 2017. The University of Edinburgh's Neural MT Systems for WMT17. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 389–399, Copenhagen, Denmark. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Edinburgh Neural Machine Translation Systems for WMT16. In *Proceedings of the First Conference on Machine Translation*, pages 371–376, Berlin, Germany. Association for Computational Linguistics.

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

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive Learning Rates with Sublinear Memory Cost. *CoRR*, arXiv/1804.04235.

Roman Sudarikov, Martin Popel, Ondřej Bojar, Aljoscha Burchardt, and Ondřej Klejch. 2016. Using MT-ComparEval. In *Translation Evaluation:*

From Fragmented Tools and Data Sets to an Integrated Ecosystem, pages 76–82, Portorož, Slovenia. LREC.

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.