CUNI System for the WMT19 Robustness Task

Jindřich Helcl and **Jindřich Libovický** and **Martin Popel**
Charles University, Faculty of Mathematics and Physics,
Institute of Formal and Applied Linguistics,
Malostranské náměstí 25, 118 00 Prague, Czech Republic
{helcl, libovicky, popel}@ufal.mff.cuni.cz

Abstract

We present our submission to the WMT19 Robustness Task. Our baseline system is the Charles University (CUNI) Transformer system trained for the WMT18 shared task on News Translation. Quantitative results show that the CUNI Transformer system is already far more robust to noisy input than the LSTM-based baseline provided by the task organizers. We further improved the performance of our model by fine-tuning on the in-domain noisy data without influencing the translation quality on the news domain.

1 Introduction

Machine translation (MT) is usually evaluated on text coming from news written by a professional journalist. However, in practice, MT should cover more domains, including informal and not carefully spelled text that we encounter in the online world.

Although the MT quality improved dramatically in recent years (Bojar et al., 2018), several studies (Belinkov and Bisk, 2018; Khayrallah and Koehn, 2018) has shown that the current systems are sensitive to the source-side noise. It is also an issue that was not studied intensively in the past because neural systems appear to be more noise-sensitive than the previously used statistical systems (Khayrallah and Koehn, 2018).

Recently, Michel and Neubig (2018) prepared a dataset called Machine Translation of Noisy Text (MTNT) that focuses exclusively on translating texts from the online environment. This dataset is used for the WMT19 Robustness Task (Li et al., 2019).

2 MTNT Dataset and Baselines

The MTNT dataset consists of sentences collected from Reddit[1] posts. Unlike the standard corpora which (in a major part) consist of formal language, often written by professionals, this dataset contains a substantial number of spelling errors, grammatical errors, emoticons, and profanities.

Manual translations are provided along with the crawled source sentences. The translators were asked to keep all the noise-related properties of the source sentence.

There are two language pairs included in the dataset: English-French and English-Japanese in both directions. The dataset comes in three splits, for training, validation, and testing. The English-French part consists of 36k examples in the training split, 852 examples for validation, 1020 examples for testing in the En→Fr direction, and 19k, 886, and 1022 examples for training, validation, and testing respectively in the opposite direction. For English-Japanese, the dataset is substantially smaller, with around 6k training examples in both directions. In our experiments, we focus solely on the translation between French and English.

We noticed that the MTNT dataset as provided for the task has some peculiarities that were probably caused inadvertently during the dataset building. Namely, the training and validation splits seem to come from a single alphabetically sorted file. This means that all validation source sentences start with the letter "Y", and anything that comes after "Y" in the alphabetical order. Because of this, the validation scores are unreliable. Moreover, a system trained on the training split will have a difficult time translating sentences beginning with e.g. the word "You", which is a commonly seen instance in the online discussion domain. This does not affect the test split.

[1] http://www.reddit.com

The baseline system introduced with the dataset is a recurrent sequence-to-sequence model with attention (Bahdanau et al., 2014). The encoder is a bidirectional LSTM with two layers. The decoder is a two-layer LSTM. The hidden state dimension in the LSTMs is 1,024 and the word embedding size is 512.

The model that was used as a baseline for the Robustness Task was trained on the WMT15 parallel data. Additionally, simple fine-tuning using stochastic gradient descent on the MTNT data is shown to improve the translation quality by a large margin. The translation quality of the system is tabulated among our systems in Table 2.

3 Related Work

There have been several attempts to increase the robustness of MT systems in recent years.

Cheng et al. (2018) employ an adversarial training scheme in a multi-task learning setup in order to increase the system robustness. For each training example, its noisy counterpart is randomly generated. The network is trained to yield such input representations such that it is not possible to train a discriminator that decides (based on the input representation) which input is the noisy one. This method improves both the robustness and the translation quality on the clean data.

Liu et al. (2018) attempt to make the translation more robust towards noise from homophones. This type of noise is common in languages with non-phonetic writing systems and concerns words or phrases which are pronounced in the same way, but spelled differently. The authors of the paper train the word embeddings to capture the phonetic information which eventually leads not only to bigger robustness but also to improved translation quality in general.

To our knowledge, the only work that specifically uses the MTNT dataset attempts to improve the system robustness by emulating the noise in the clean data (Vaibhav et al., 2019). They introduce two techniques for noise induction, one employing hand-crafted rules, and one based on back-translation. The techniques offer a similar translation quality gains as fine-tuning on MTNT data.

4 The CUNI Transformer model

Our original plan was to train a system that would be robust by itself and would not require further fine-tuning on the MTNT dataset.

	Corpus	# Sentences
Parallel	10^9 English-French Corpus	22,520k
	Europarl	2,007k
	News Commentary	200k
	UN Corpus	12,886k
	Common Crawl	3,224k
Mono	French News Crawl ('08–'14)	37,320k
	English News Crawl ('11–'17)	127,554k

Table 1: Overview of the data used to train the CUNI Transformer baseline system.

Our baseline is the Transformer "Big" model (Vaswani et al., 2017) as implemented in Tensor2Tensor (Vaswani et al., 2018). We train the model using the procedure described in Popel (2018) and Popel and Bojar (2018), which was the best-performing method for Czech-to-English and English-to-Czech translation in the WMT18 News Translation shared task (Bojar et al., 2018).

We trained our model on all parallel data available for the WMT15 News Translation task (Bojar et al., 2015). We acquired additional synthetic data by back-translation of the WMT News Crawl corpora (from years 2008–2014 for French and 2011–2017 for English). We did not include the News Discussion corpus that we considered too noisy for training the system. Table 1 gives an overview of the training data composition.

5 Fine-Tuning

Similarly to the baseline experiments presented with the MTNT dataset (Michel and Neubig, 2018), we fine-tune our general-domain model on the MTNT dataset.

We continued the training of the models using the training part of the MTNT dataset. Unlike the original model, we used plain stochastic gradient descent with a constant learning rate for updating the weights. We executed several fine-tuning runs with different learning rates and observed that learning rates smaller than 10^{-5} do not change the model outputs at all and learning rates larger than 10^{-4} cause the models to diverge immediately. The models in our final submission were fine-tuned with a learning rate of 10^{-4}.

	English-French				French-English			
	WMT14	WMT15	MTNT	blind	WMT14	WMT15	MTNT	blind
MTNT baseline	33.5	33.0	21.8	22.1	28.9	30.8	23.3	25.6
+ fine-tuning	—	—	29.7	—	—	—	30.3	—
CUNI Transformer	43.6	41.6	34.0	37.0	42.9	39.6	39.9	42.6
+ fine-tuning	43.5	41.6	36.6	38.5	41.5	40.9	42.1	44.8

Table 2: BLEU scores of the baseline and CUNI models measured on several datasets.

	en-fr	fr-en
Naver Labs Europe	41.4	47.9
this work	38.5	44.8
Baidu & Oregon State Uni.	36.4	43.6
Johns Hopkins Uni.	—	40.2
Fraunhofer FOKUS – VISCOM	24.2	29.9
MTNT Baseline	22.1	25.6

Table 3: Quantiative comparison of the CUNI Transformer system + fine-tuning (this work) with other submitted systems.

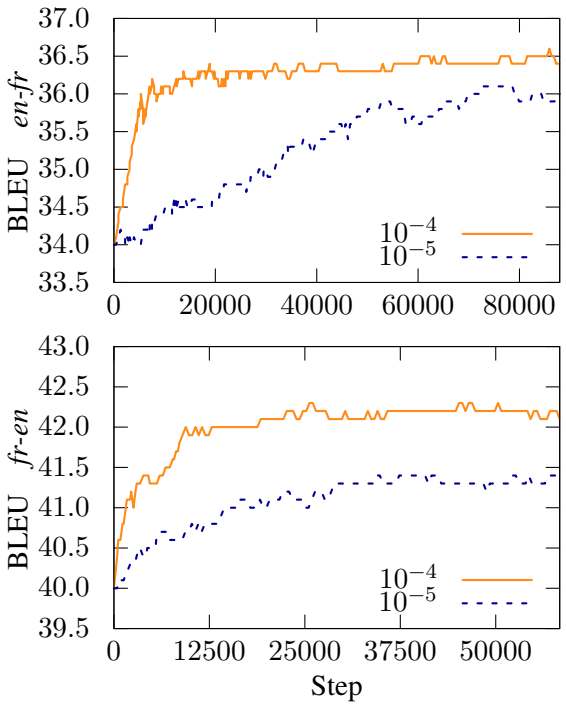

Figure 1: Learning curves showing the progress of fine-tuning on the MTNT test split for English-to-French (top) and French-to-English (bottom) systems with two different learning rates.

6 Results

We evaluate the results on four datasets. The first one is *neswtest2014* (Bojar et al., 2014), a standard WMT test set consisting of manually translated newspaper texts where one half is originally in English and the other half originally in French.

Because of the large amount of training data available, even the statistical MT systems achieved high translation quality on the news domain. Because of that, a slightly different test set (*newsdiscusstest2015*) was used as the evaluation test set for the WMT15 competition (Bojar et al., 2015). The test set consists of sentences from discussions under news stories from The Guardian and Le Monde. Even though the topics are the same as the news stories, the language used in the discussions is less formal and contains grammatical and spelling errors, which makes them somewhat closer to the MTNT dataset.

Finally, we evaluate the models on the test part of the MTNT dataset (described in Section 2) and the blind test set for the WMT19 Robustness Task, which was collected in the same way as the original MTNT dataset.

The quantitative results are shown in Table 2. The Transformer-based baseline outperforms the RNN-based MTNT baseline by a large margin on both WMT and MTNT test datasets.

The fine-tuning of the RNN-based models brings a substantial translation quality boost of 8 and 7 BLEU points in each direction respectively. This effect is much smaller with our stronger baseline and only improves the performance by around 2 BLEU points in either direction. This may indicate that sufficiently strong models are robust enough and do not need further fine-tuning for the type of noise present in the MTNT dataset. Especially in French-to-English translation, the fine-tuning improvement is reached at the expense of decreased translation quality in the news domain.

We observe that the fine-tuning has only a small

negative impact on the translation quality of our models on the general-domain data. It would be interesting to see how big impact made the fine-tuning of the MTNT baseline model, which gained such a large improvement on the domain-specific data. However, the authors of the baseline (Michel and Neubig, 2018) do not report these results.

We plot the learning curves from the progress of the system fine-tuning in Figure 1. Even though the fine-tuning improved the model performance on both language pairs by approximately the same margin, the courses of the fine-tuning differ fundamentally. For English-to-French translation, we see that the translation quality slowly increases until convergence. For the opposite direction, it improves immediately and keeps oscillating during the remaining training steps. We found that this effect was similar regardless of the learning rate.

Although we observed a strong effect of checkpoint averaging during the baseline model training, it has almost no effect on the fine-tuned models. Therefore, we report only the performance for parameter checkpoints with the highest validation BLEU scores.

Table 3 compares the automatic scores with other WMT19 Robustness Task participants. Our submission was outperformed by submissions by Naver Labs Europe in both translation directions. Their submission used the same architecture as our submission, but in addition, it employed corpus tags and synthetic noise generation. Details about other systems were not known at the time of our submission.

7 Conclusions

In our submission to the WMT19 Robustness Task, we experiment with fine-tuning of strong Transformer-based baselines for translation between English and French.

Our results show that when using a strong baseline, the effect of fine-tuning on a domain-specific dataset is much smaller than for weaker models introduced as a baseline with the MTNT dataset.

Acknowledgements

This research has been supported by the from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 825303 (Bergamot), Czech Science Foundtion grant No. 19-26934X (NEUREM3), and Charles University grant No. 976518, and has been using language resources distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (LM2015071).

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and natural noise both break neural machine translation. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. 2015. Findings of the 2015 workshop on statistical machine translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 1–46, Lisbon, Portugal. Association for Computational Linguistics.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 272–303, Belgium, Brussels. Association for Computational Linguistics.

Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA. Association for Computational Linguistics.

Yong Cheng, Zhaopeng Tu, Fandong Meng, Junjie Zhai, and Yang Liu. 2018. Towards robust neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1756–1766, Melbourne, Australia. Association for Computational Linguistics.

Huda Khayrallah and Philipp Koehn. 2018. On the impact of various types of noise on neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 74–83, Melbourne, Australia. Association for Computational Linguistics.

Xian Li, Paul Michel, Antonios Anastasopoulos, Yonatan Belinkov, Nadir K. Durrani, Orhan Firat, Philipp Koehn, Graham Neubig, Juan M. Pino, and Hassan Sajjad. 2019. Findings of the first shared task on machine translation robustness. In *Proceedings of the Fourth Conference on Machine Translation, Volume 2: Shared Task Papers*, Florence, Italy. Association for Computational Linguistics.

Hairong Liu, Mingbo Ma, Liang Huang, Hao Xiong, and Zhongjun He. 2018. Robust neural machine translation with joint textual and phonetic embedding. *CoRR*, abs/1810.06729.

Paul Michel and Graham Neubig. 2018. MTNT: A testbed for machine translation of noisy text. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 543–553, Brussels, Belgium. Association for Computational Linguistics.

Martin Popel. 2018. CUNI transformer neural MT system for WMT18. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 482–487, Belgium, Brussels. Association for Computational Linguistics.

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

Vaibhav, Sumeet Singh, Craig Stewart, and Graham Neubig. 2019. Improving robustness of machine translation with synthetic noise. *CoRR*, abs/1902.09508.

Ashish Vaswani, Samy Bengio, Eugene Brevdo, Francois Chollet, Aidan Gomez, Stephan Gouws, Llion Jones, Łukasz Kaiser, Nal Kalchbrenner, Niki Parmar, Ryan Sepassi, Noam Shazeer, and Jakob Uszkoreit. 2018. Tensor2Tensor for neural machine translation. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Papers)*, pages 193–199, Boston, MA. Association for Machine Translation in the Americas.

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, Long Beach, CA, USA. Curran Associates, Inc.