System Description
– the submission of FOKUS to the WMT 19 robustness task –

Cristian Grozea
Fraunhofer FOKUS
cristian.grozea@fokus.fraunhofer.de

Abstract

This paper describes the systems of Fraunhofer FOKUS for the WMT 2019 machine translation robustness task. We have made submissions to the EN-FR, FR-EN, and JA-EN language pairs. The first two were made with a baseline translator, trained on clean data for the WMT 2019 biomedical translation task. These baselines improved over the baselines from the MTNT paper by 2 to 4 BLEU points, but where not trained on the same data. The last one used the same model class and training procedure, with induced typos in the training data to increase the model robustness.

1 Introduction

Our submissions to the robustness task (Li et al., 2019) aimed to investigate two questions: a) how robust are well-performing models that are trained on clean text and b) does making small intentional "typos" in the training data lead to robust models?

2 Methods

FR-EN, EN-FR

We reproduce here for the sake of self-containment the description of the baseline model we have developed for the biomedical translation task. In order to create a baseline for that task, we have tried to emulate a non-expert who uses a slightly modified NMT tutorial on the data listed in the competition page to compete (minimal effort). The tutorial our submissions are based on was written for the MT Marathon 2018 Labs and is available online [1]. It uses the Marian NMT system(Junczys-Dowmunt et al., 2018).

As training data we have used the UFAL medical corpus(UFA), from which we have removed the "Subtitles" pairs, as they are lower quality than the rest, less medically oriented – if at all, and with the wrong addressing (dialogue, as opposed to narration). As validation dataset we have used Khresmoi(Pecina et al., 2013), which we did not find to be included in UFAL, despite being mentioned as one of the sources.

The training was set to stop when either the cross-entropy or the the BLEU on the validation stalled for 5 training and evaluation cycles. One such cycle processed 10000 sentence pairs.

The model implemented by Marian NMT in the tutorial used here is Sequence2Sequence with shallow networks. The text data has been preprocessed with BPE. Here we deviated for efficiency reasons from the MOSES BPE(Koehn et al., 2007) and used FastBPE[2].

The vocabulary size for BPE was set to 85000, the workspace memory to be reserved on the GPU was reduced to 6 GB to avoid out of memory errors on GTX 1080 Ti. The tests were run on machines with 8 GPUs, the training process of a single language pair took in general a couple of days.

JA-EN

For the Japanese to English submission, we have employed the same models and training as above, but with a preprocessing intended to increase the robustness to typos of two types: missing letters, duplicated letters.

3 Results

The results are presented in Table 1

4 Discussion and Conclusion

The models trained on the UFAL medical corpus are fairly robust and generic, not excessively specialized for the biomedical domain. Despite being trained for the biomedical translation task,

[1] https://marian-nmt.github.io/examples/mtm2018-labs

[2] https://github.com/glample/fastBPE

Proceedings of the Fourth Conference on Machine Translation (WMT), Volume 2: Shared Task Papers (Day 1) pages 537–538
Florence, Italy, August 1-2, 2019. ©2019 Association for Computational Linguistics

Source	Target	BLEU uncased	BLEU cased	WMT19 Biomed.
EN	FR	24.8	24.2	32.5
FR	EN	30.8	29.9	29.9
JA	EN	7.3	6.4	ZH2EN 16.7

Table 1: BLEU scores of our submissions, contrasted with the results of the same models on the biomedical translation task, except for JA-EN, where the result on the closest language pair is given, Chinese to English

the EN2FR and FR2EN models trained by us behaved reasonably well in the WMT ROBUSTNESS task, surpassing the NTMT paper baseline by 2.5 (EN2FR) and 4 (FR2EN) BLEU points, with the caveat of not being a constrained system, in the sense that the training has not been done on the data listed and intended for that task. Still, as Reddit is not among the sources of UFAL, this should not affect the validation results.

One choice that we made, and we think it is right for the biomedical task, to avoid dialogues and direct speech (the subtitles part of UFAL medical corpus) has probably influenced negatively the performance in the robustness task - the Reddit text used for evaluation contains often the first person and second person addressing modes.

In comparison with the performance on the biomedical text, the performance of FR-EN was apparently not affected by the noisy text, whereas for EN-FR there was a strong decrease of the BLEU score, 8.3 points from 32.5 down to 24.2. We did apply the postprocessing of the French text to fix the punctuation marks, thus there should be another explanation for the decrease of performance.

The performance of the JA-EN was very low. Visual inspection of the results shows typical early stage training RNN issues like this translation: Our model's translation: "It's very, very, ..." repeated 17 times. The reference translation was "Minpaku has such cool content and it was fun". In general, numbers are changed to other numbers or ignored completely by our JA-EN translation model. One can assume the training data was not sufficient in quantity to train a reliable Japanese to English translation model. In addition to that, due to an error, we have introduced the intentional typos not only in the source text but also in the target text.

The quality of the FR-EN and EN-FR is on the surface better, but they miss fairly easy translations by translating too literally ("I'm **on the train**" translated as "Je suis **sur le train**") or by missing the correct sense of the word, probably because we didn't use the context at all ("I don't think we're are making any **trades** til the off season." translated as "Je ne pense pas que nous ne faisons aucun **métier** en dehors de la saison."). Meaning got changed ("tu crois vraiment qu'il n'y a vraiment **aucune solution que** la ségrégation ?" went to "Do you really believe that there is really **no solution to** segregation?"), coreference is not properly processed ("**Comme** Nelson Mandela ne voulait pas le pouvoir aux noirs(...), il voulait la fin du racisme." was translated "**As** Nelson Mandela(...)he wanted to see the end of racism.").

References

UFAL medical corpus 1.0. https://ufal.mff.cuni.cz/ufal_medical_corpus. Accessed: 2018-07-24.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, pages 177–180. 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.

Pavel Pecina, Ondřej Dušek, Jan Hajič, and Zdeňka Urešová. 2013. Khresmoi query translation test data 1.0. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.