Allegro.eu submission to WMT21 News Translation Task

Mikołaj Koszowski, Karol Grzegorczyk, Tsimur Hadeliya
ML Research Lab at Allegro.eu
{mikolaj.koszowski, karol.grzegorczyk, tsimur.hadeliya}@allegro.pl

Abstract

This paper describes Allegro.eu submission for the WMT21 news translation shared task. We focus on exploring data filtering and data augmenting methods. We submitted two single-directional models, one for English→Icelandic direction and other for Icelandic→English direction. Our news translation system is based on the transformer-big architecture, it makes use of corpora filtering, back-translation and forward translation applied to parallel and monolingual data alike.

1 Introduction

We participated in the WMT21 news translation shared task for English↔Icelandic language pair. It is a medium-resource regime with under 10M parallel sentences. In our experiments we focused on two approaches for improving translation system: data filtering methods inspired by work of (Jónsson et al., 2020) and data augmentation methods like back-translation or self-training (Edunov et al., 2018; Sennrich et al., 2016; He et al., 2019). We tried to use bi-directional translation models but single-directional proved to be better. We also tried to make use of pretraining on monolingual corpora, but it also was unsuccessful. Krubiński et al. (2020) showed in their ablation study that pretraining is the most successful for low-resource regimes under 1M parallel sentences.

2 Data

2.1 Data Preprocessing

We removed malformed utf-8 encodings, normalized text with NFKC Unicode normalization form, unescaped HTML, removed control characters and converted different whitespaces to a basic space character.

2.2 Data Filtering

We took part in a constrained track for the English↔Icelandic language pair for the news translation task. We used similar heuristic for filtering monolingual and parallel data. A proper sentence pair should fulfil these criteria:
For each sentence separately:

- length in chars $\in (10, 500)$
- length in words $\in (2, 100)$
- average word length in chars < 12
- max word length in chars < 28
- digit ratio < 0.15
- outside alphabet ratio < 0.015
- language detection probability > 0.9

Criteria calculated on a sentence pair:

- no digit sequence mismatch
- Levenshtein distance > 5
- Poisson based length logprob $>$ -10

For language identification we used the CLD2 library. We arrived at these threshold values by analyzing outliers of clean corpora: newsdev2021 development dataset and Jónsson's cleaned ParIce corpus (Jónsson et al., 2020). Our filtering procedure is inspired by Jónsson's and extracts 72% of the same sentences they extracted from the raw ParIce corpus (Barkarson and Steingrímsson, 2019). Each heuristic removes up to 5% of lines from those clean corpora, when all thresholds would be applied they would remove around 9% from the cleaned ParIce corpus. For all available raw parallel corpora this procedure would remove 35% of sentences. Table 1 shows sizes of raw and filtered corpora available in the constrained track.

2.3 Poisson based length filtering

This section describes an improved method of filtering sentences based on their lengths. A simple ratio of sentence lengths is a common method, but it is often too strict for short sentences and too loose

Proceedings of the Sixth Conference on Machine Translation (WMT), pages 140–143
November 10–11, 2021.

Parallel corpora	raw	filtered	left
ParIce.1_1	3.56M	1.98M	0.56
ParaCrawl.7_1	2.39M	1.95M	0.81
WikiMatrix.1	313k	177k	0.57
wikititles.3	50k	2k	0.04
Total	6.31M	4.1M	0.65

Table 1: Sizes of parallel corpora.

for longer ones. We are using a simple assumption, that the distribution of lengths of expected translation is given by the Poisson distribution with a mean equal to a length of the source sentence. This type of length filtering is used by bicleaner framework (Sánchez-Cartagena et al., 2018). We use a correction factor $scl = 1.04$, which is a ratio of chars in the English side to the Icelandic side for the whole parallel corpus. We multiply source length by it or by its reciprocal before calculating probabilities, depending on the context. Figure 1 compares this method with a ratio-based heuristic where the allowable ratio range is $(0.5, 2)$. For this language pair the correction factor is close to 1.0, but for other language pairs it can deviate more, which can lead to bias when using a simple ratio-based heuristic.

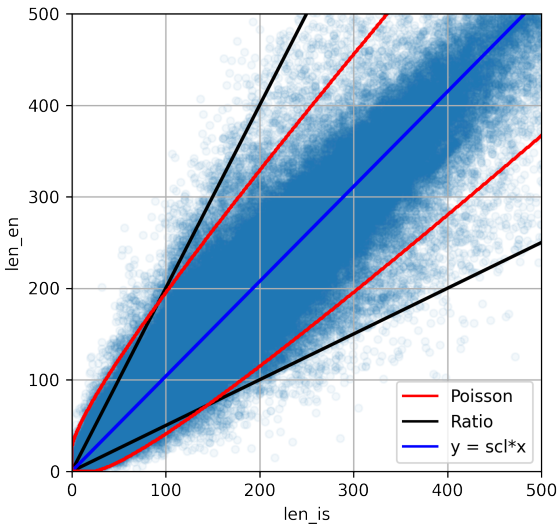

Figure 1: Distribution of lengths of parallel corpora. As depicted, Poisson-based heuristic allows more variation for shorter sentences and lower variation in length for longer ones.

2.4 Translation postprocessing

Our system has a tendency to generate the same quotation as in source text. Therefore, before submitting our translations for evaluation, we applied simple regular expressions to fix quoting. We made sure that only (" ") for English submission was used and for Icelandic we made sure that („ “) was used.

3 System overview

All of our models are based on the Transformer big architecture, as described in Vaswani et al. (2017). For training we used OpenNMT-py framework (Klein et al., 2017) together with sentencepiece tokenizer (Kudo and Richardson, 2018) unigram model of size 32k with full character coverage. We trained models on A100 GPU for 210k steps with a batch of 8192 tokens which amounts to around 12h per model. We used half-precision and tied embeddings. For optimization we used Adam (Kingma and Ba, 2014), with a linear warmup for learning rate for 15k steps up to 0.0005 and inverse square root decay afterwards. Additionally, all of our models were randomly initialized.

4 Results

Results are presented in Table 3. We trained a tokenizer on a cleaned ParIce corpus. A baseline model we trained on all available parallel corpora and achieved 18.1 BLEU in English→Icelandic direction and 24.0 BLEU in Icelandic→English direction.

4.1 Data filtering impact

We ran 4 variants with the same parameters as described at the beginning of section 3, but only for 100k steps. We compared the translation quality of models trained with filtered training corpus and the impact of cleaning data used in training tokenizer. We used the aforementioned cleaned ParIce corpus (Jónsson et al., 2020) to train the tokenizer. Table 2 presents the results of this comparison.

4.2 Back-translation of monolingual corpora

We took 10M monolingual sentences for each language and filtered them as described in section 2.2. For English we took only News Crawl from 2020, for Icelandic we used News Crawl 2020 and also Icelandic Gigaword to obtain full 10M sentences. We translated the English source to Icelandic, then translated it back to English. Then we compared those second translations to source by GLEU score

	clean tokenizer	raw tokenizer
clean corpus	16.6/22.6	14.0/19.4
raw corpus	16.2/22.2	14.2/18.9

Table 2: Comparison of impact of filtering data. Values reported are BLEU scores for en→is/is→en direction for newsdev2021. We can easily see that training tokenizer on clean data has a big impact. Also we can notice that removing 35% of parallel corpora can improve the quality of the model given the same amount of compute.

(Wu et al., 2016) and filtered the best 40% of pairs of original source and first translation based on that. GLEU score is a variation on the BLEU score. It is claimed to be a more accurate measure of single sentence translation quality. We repeated this procedure for 10M Icelandic monolingual sentences. It is interesting to note that 4.4% and 2.0% of second translations were the same as the original source, for English and Icelandic respectively. We then created English biased corpus which consisted of:

- 4M of clean parallel corpus
- 4M of English based back-translation where we used original source as target
- 4M of Icelandic base forward translation where we used our first translations as target

Then we used this corpus to train a new model, it achieved 26.8 BLEU in Icelandic→English direction.

4.3 Back-translation of parallel corpora

We used this newly acquired model to translate the Icelandic side of clean parallel corpus to English and likewise filtered by GLEU score for the English side of the corpus, finally we extracted 75% of most similar pairs. It is interesting to note that 11% of translations were the same as the English side of the parallel corpus. We then created a corpus for training English→Icelandic model, this time with typical setup for back-translation where original sentences were used as a target:

- 4M of clean parallel corpus
- 4M backtranslated monolingual corpus
- 3M backtranslated parallel corpus

Then we used this corpus to train a new model. It achieved 23.6 BLEU in English→Icelandic direction and that was our final model for this direction.

	newsdev2021	
Model	**En→Is**	**Is→En**
baseline	18.1	24.0
BT and FT mono	-	26.8
BT mono and parallel	23.6	-
BT mono and parallel	-	27.2
final models	23.6	27.4
	newstest2021	
final submission	22.7	33.3

Table 3: Comparison of forward-translation (FT) and back-translation (BT) model trained on monolingual and parallel corpora

Then, analogously, we used this model to translate the other side of the clean parallel corpora and filter by GLEU score. It is interesting to note that also 11% of translations was the same as the Icelandic side of the parallel corpus. We then created a corpus and trained Icelandic→English model which achieves 27.2 BLEU on the development set. For this direction our final system was an ensemble of this new model and previous best.

4.4 Denoising

As it has been recently demonstrated by Raffel et al. (2020), transfer learning can be successfully applied to sequence-to-sequences models. Therefore, we tried doing unsupervised de-noising pre-training based on provided monolingual data. We experimented with three different denoising schemes:

- Token-based masked language modeling (Devlin et al., 2019)
- Whole Word Masking objective inspired by BERT models released in May 2019
- BART-like denoising with text infilling and sentence permutation (Lewis et al., 2020)

We tried it in two regimes. One where we pretrain model and then finetune it on translation downstream task. The other where we train both denoising and translation objectives simultaneously. However, we didn't observe any benefits from doing this. The reason for this is unknown.

5 Conclusion

This paper describes Allegro.eu submission for the WMT21 news translation shared task. We took part in constrained track for the English↔Icelandic language pair only. Participation in this task allowed

us to deepen the understanding of filtering methods common in NMT. The experiments demonstrated the importance of data filtering in medium-resource regime machine translation. In this regime, less data but of higher quality can lead to superior results.

References

Starkaður Barkarson and Steinþór Steingrímsson. 2019. Compiling and filtering ParIce: An English-Icelandic parallel corpus. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 140–145, Turku, Finland. Linköping University Electronic Press.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.

Junxian He, Jiatao Gu, Jiajun Shen, and Marc'Aurelio Ranzato. 2019. Revisiting self-training for neural sequence generation. *CoRR*, abs/1909.13788.

Haukur Páll Jónsson, Haukur Barri Símonarson, Vésteinn Snæbjarnarson, Steinþór Steingrímsson, and Hrafn Loftsson. 2020. Experimenting with different machine translation models in medium-resource settings. In *Text, Speech, and Dialogue - 23rd International Conference, TSD 2020, Brno, Czech Republic, September 8-11, 2020, Proceedings*, volume 12284 of *Lecture Notes in Computer Science*, pages 95–103. Springer.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

Mateusz Krubiński, Marcin Chochowski, Bartłomiej Boczek, Mikołaj Koszowski, Adam Dobrowolski, Marcin Szymański, and Paweł Przybysz. 2020. Samsung R&D institute Poland submission to WMT20 news translation task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 181–190, Online. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *CoRR*, abs/1808.06226.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67.

Víctor M. Sánchez-Cartagena, Marta Bañón, Sergio Ortiz-Rojas, and Gema Ramírez. 2018. Prompsit's submission to WMT 2018 parallel corpus filtering shared task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 955–962, Belgium, Brussels. Association for Computational Linguistics.

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

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*, volume 30. Curran Associates, Inc.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.