Miðeind's WMT 2021 submission

Haukur Barri Símonarson, Vésteinn Snæbjarnarson,
Pétur Orri Ragnarsson, Haukur Páll Jónsson and Vilhjálmur Þorsteinsson
Miðeind ehf., Reykjavík, Iceland
`{haukur, vesteinn, petur, haukurpj, vt}@mideind.is`

Abstract

We present Miðeind's submission for the English→Icelandic and Icelandic→English subsets of the 2021 WMT news translation task. Transformer-base models are trained for translation on parallel data to generate back-translations iteratively. A pretrained mBART-25 model is then adapted for translation using parallel data as well as the last backtranslation iteration. This adapted pretrained model is then used to re-generate backtranslations, and the training of the adapted model is continued.

1 Introduction

Our work on machine translation for Icelandic has been going on for a couple of years as a part of the government sponsored Icelandic Language Technology Programme (Nikulásdóttir et al., 2020). By building on state-of-the-art solutions we have developed an open and effective translation system between Icelandic and English.

To achieve this, we collect parallel Icelandic and English texts which are filtered for good quality alignments. We also collect monolingual text for backtranslations. We follow tried and tested methods in neural machine translation using iterative backtranslation (Edunov et al., 2018) and adapt the multilingual denoising autoencoder model mBART-25 (Liu et al., 2020) for translation.

2 Datasets

We used several parallel and monolingual datasets, both publicly available and created in-house.

2.1 Parallel data

The parallel data used are ParIce (Steingrímsson et al., 2020) and the JW300 corpus (Agić and Vulić, 2019). In addition we used a parallel student theses and dissertation abstracts corpus, IPAC, generated in-house and sourced from the Skemman reposi-

tory[1] as described in (Símonarson and Snæbjarnarson, 2021). A breakdown of the data is shown in Table 1.

Corpus	#Sentences
The Bible	33k
EEA regulatory texts	1,700k
EMA	404k
ESO	12.6k
OpenSubtitles	1,300k
Tatoeba	10k
Jehova's Witnesses (JW300)	527k
Other*	93k
IPAC	64k

Table 1: Parallel corpora used. #Sentences are the number of sentence pairs. *Other** denotes software localizations, Project Gutenberg literature and the Icelandic sagas.

Following (Pinnis, 2018) we apply simple heuristic filters to the parallel data, mainly for capturing OCR and PDF errors, and correcting or removing character encoding errors after deduplication. Filters include but are not limited to: empty sequence removal, length cut-offs, character whitelists, mismatch in case and symbols between languages, edit-distances between source and target, normalizing of punctuation, and ad-hoc regular expressions for Icelandic specific OCR/PDF errors. For a more in-depth description see (Jónsson et al., 2020).

Other potential parallel datasets are ParaCrawl (Bañón et al., 2020) and CCMatrix (Schwenk et al., 2021). Manual review of a couple of hundred randomly chosen lines from ParaCrawl revealed that the data quality is quite low for Icelandic, many lines are machine translated or badly aligned. We therefore did not include ParaCrawl. CCMatrix did not exist when the project started and we have not

---

[1] `https://skemman.is`

taken the time to review and integrate it although a quick inspection indicates that the quality is higher than in ParaCrawl.

2.2 Data used for backtranslation

We collected and translated monolingual data for backtranslations, made available in (Símonarson et al., 2020), mostly building on the work in (Edunov et al., 2018). The English sentences (44.7m) are retrieved from the Wikipedia, Newscrawl and Europarl corpora. The Icelandic sentences (31.3m) are sourced from the Icelandic Gigaword Corpus (Steingrímsson et al., 2018).

Lang.	Name	#Sentences
IS	Court rulings	1.8M
IS	Supreme court rulings	2M
IS	Laws	814k
IS	Web of Science	268k
IS	Wikipedia	405k
IS	Parliamentary proc.	6.2M
IS	Misc	350k
IS	Newspaper (Mbl)	13.6M
IS	Newspaper (Visir)	4.9M
IS	Radio transcripts	1M
EN	Newscrawl	33.4M
EN	Wikipedia	9.3M
EN	Europarl	2.0M

Table 2: Monolingual data used for backtranslation.

3 Training of small transformer models

Our earlier models were trained using the transformer-base configuration described in (Vaswani et al., 2017) as implemented in Google's Tensor2Tensor (TensorFlow-based) (Vaswani et al., 2018) library. For later models we switched to Facebook's Fairseq (Ott et al., 2019) library. An improved translation task was implemented in Fairseq to include BPE dropout; it is available in the `greynirseq`[2] library.

The transformer-base models were trained iteratively and used to generate new backtranslations. We stopped when each language direction had been trained on backtranslations that were produced by a model that had itself seen backtranslations at training time. We compared tagged and untagged backtranslations, sampling versus beam search and different mixing ratios (upsampling rate) between parallel and backtranslated data. Using tagged

backtranslations as opposed to no tag showed an improvement from 16.5 to 17.5 BLEU[3] after the first iteration over the IPAC development set, while using no backtranslations gave 15.0, so we proceeded to use tagged translations.

Model	BLEU
Transformer-base	16.5
Transformer-base + bt	17.5
Transformer-base + iterative-bt	18.5
mBART (first run)	23.1
mBART (continued)	23.6

Table 3: BLEU scores over IPAC for the EN-IS direction.

We use the IPAC test set to measure BLEU since it was available, has a large range of topics (although maybe not a large range of style) and is very unlikely to be accidentally included in the training data. The IPAC data is out of distribution with the rest of the training data but we do not consider that to be a problem since our goal is a general purpose model. The WMT dev set did not exist at the time.

We used a joint BPE vocabulary of size 16k and shared input-output embedding matrices. We pre-tokenized the input using `tokenizer` [4] for the Icelandic side and spaCy (Honnibal et al., 2020) for the English side. A beam width of 4 was used for beam search during backtranslation. Each training iteration took approximately one week on a single GTX 1080 graphics card. We were pleasantly surprised with how far we got with only this modest hardware.

3.1 Translation mixing ratio selection and beam noise

We assessed the impact of the ratio of synthetic backtranslation data to authentic parallel data on translation performance. Best results were obtained with a 1:2 ratio of authentic to synthetic data, using IPAC (held out from training) for evaluation.

For noising the backtranslation beam outputs, we follow (Edunov et al., 2018) and used within-k permutation of whole words (with k=3), whole-word masking, and word dropout. Using sampling and noised beam outputs yielded comparable results.

---

[2] https://github.com/mideind/greynirseq

[3] SacreBLEU signature: `BLEU+case.mixed+ lang.en-is+numrefs.1+smooth.exp+tok.13a+ version.1.5.1`

[4] https://github.com/mideind/Tokenizer

4 Adapting mBART-25 for translation

The mBART-25 (Liu et al., 2020) (610M parameters) language model is far larger than the Transformer-base model (110M parameters). It was pretrained on 25 languages, including English and Swedish, but not Icelandic. We adapt it for translation from Icelandic to English and vice versa, using the same human-derived parallel translation data as for the transformer-base model along with the synthetic backtranslated corpus in a ratio of 1:2. We do not use any pre-tokenization and inherit the BPE sententencepiece vocabulary from mBART-25 (of size 250k) with the addition of an Icelandic language marker that was randomly initialized. We use the same hyperparameters as in (Liu et al., 2020) and the implementation from Fairseq (Ott et al., 2019). The models are trained until their performance on the development sets plateaus.

The initial learning rate was set to 3e-4. Sixteen 32GB nVidia V100 GPUs connected with Infiniband were used for training. The effective batch size was around 10k sequences and the training took around 4 days of wall clock time per model.

Subsequently, these trained models were used to generate improved backtranslations. We then continued training the first iteration of our models with the new backtranslated data for another 30,000 steps for the Icelandic-English direction, and 36,000 steps for the English-Icelandic direction. The same training configurations were maintained as for the earlier runs.

Dir.	Steps	'21 test	'21 dev	EEA
En-Is	40k	22.7	25.9	54.5
En-Is	40k + 36k	24.3	27.8	57.6
Is-En	36k	32.9	30.4	61.0
Is-En	36k + 30k	33.5	31.8	63.2

Table 4: BLEU scores for the mBART-25 adapted translation models over the newstest2021 and EEA evaluation sets.

The benefit of continuing training of the mBART-derived models ranges from 0.6 to 3.1 BLEU as shown in Table 4. BLEU performance is shown for both the newstest2021 development set as well as our cleaned-up dataset with sentence pairs from the EEA regulation corpus. Note that we do not fine-tune prior to evaluation nor do we perform checkpoint averaging.

5 Conclusion

We have shown how a small team with modest resources can adapt state-of-the-art methods to a medium resource language and achieve competitive results on machine translation between English and Icelandic.

The trained models are available for translation at `https://velthyding.is` and will be made available at the open `CLARIN-IS`[5] repository. While a formal human comparison of the current models to the popular Google Translate service has not been performed, hundreds of monthly active users choose our solutions for translation between Icelandic and English.

6 Future work

We note the relatively small training time of the mBART adaptation and the lack of Icelandic data in the pretraining task for mBART as primary factors that can be addressed for improving results. Additionally online (or semi-online) self-training instead of train-then-translate would also improve results, especially with selective loss truncation as described in (Zhou et al., 2021). The data selected for backtranslation should also be expanded for greater diversity of both genre and vocabulary. Finally, extending the translation context beyond the current sentence level is likely to improve results.

References

Željko Agić and Ivan Vulić. 2019. JW300: A wide-coverage parallel corpus for low-resource languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3204–3210, Florence, Italy. Association for Computational Linguistics.

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis,

[5]`https://repository.clarin.is`

Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. 2020. ParaCrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Online. 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.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

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*, pages 95–103, Cham. Springer International Publishing.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

A. B. Nikulásdóttir, Jón Guðnason, Anton Karl Ingason, H. Loftsson, Eiríkur Rögnvaldsson, Einar Freyr Sigurðsson, and Steinþór Steingrímsson. 2020. Language Technology Programme for Icelandic 2019-2023. In *LREC*.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A Fast, Extensible Toolkit for Sequence Modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Mārcis Pinnis. 2018. Tilde's parallel corpus filtering methods for WMT 2018. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 939–945, Belgium, Brussels. Association for Computational Linguistics.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. 2021. CCMatrix: Mining billions of high-quality parallel sentences on the web. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6490–6500, Online. Association for Computational Linguistics.

Haukur Barri Símonarson, Vésteinn Snæbjarnarson, and Vilhjálmur Þorsteinsson. 2020. En-Is Synthetic Parallel Corpus. CLARIN-IS.

Steinþór Steingrímsson, Sigrún Helgadóttir, Eiríkur Rögnvaldsson, Starkaður Barkarson, and Jón Guðnason. 2018. Risamálheild: A Very Large Icelandic Text Corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, LREC 2018, Miyazaki, Japan.

Steinþór Steingrímsson, Hrafn Loftsson, and Andy Way. 2020. Effectively aligning and filtering parallel corpora under sparse data conditions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 182–190, Online. Association for Computational Linguistics.

Haukur Barri Símonarson and Vésteinn Snæbjarnarson. 2021. Icelandic Parallel Abstracts Corpus.

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 Track)*, 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, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Chunting Zhou, Graham Neubig, Jiatao Gu, Mona Diab, Francisco Guzmán, Luke Zettlemoyer, and Marjan Ghazvininejad. 2021. Detecting hallucinated content in conditional neural sequence generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1393–1404, Online. Association for Computational Linguistics.