Miðeind's submission to WMT 2021

News translation between Icelandic and English

Haukur Barri Símonarson, Vésteinn Snæbjarnarson, Pétur Orri Ragnarsson, Haukur Páll Jónsson and Vilhjálmur Þorsteinsson - **Miðeind ehf, Reykjavík**

1. Machine translation for Icelandic

As a part of the Icelandic Government's three-year Language Technology Programme we have been implementing NMT systems for translation between Icelandic and English for the last two years.

All project data and tooling is open and public, including our submission to WMT 2021. We host a website at https://velthyding.is where our models can be tested.

Icelandic is morphologically rich low-to-medium resource language in the Germanic language family, most related to Norwegian, Danish and Swedish.

2. Data

Parallel data

Dataset	Sentences
Bible	33k
Jehovah's Witnesses	527k
EEA regulations	1,700k
EMEA	404k
ESO	12.6k
Open Subtitles	1,300k
Tatoeba	10k
Other	93k
IPAC - abstracts	64k

Backtranslation data (monolingual)

Language	Dataset	Sentences
IS	Legal	11M
IS	Encyclopaedic	0.7M
IS	News	20M
EN	Legal	2M
EN	Encyclopaedic	9M
EN	News	33M

3. Backtranslation and data cleaning

Backtranslation mixing ratio
We tested different mixing ratios when training. The best results were obtained with a ratio of 1:2 of authentic to synthetic data.

Data cleaning
Datasets are thoroughly cleaned, i.e. normalized, duplicates removed, OCR and PDF errors removed, and various parallel sentence sanity checks applied.

4. Training approach

We follow tried and tested methods using an iterative approach:

1. Train Transformer-base translation model
2. Generate backtranslation data
3. Fine tune mBART-25 for translation
4. Generate new backtranslation data
5. Continue training translation models

16 x 32GB V100 GPUs were used for training. The final models were trained for 4 days each.

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

Evaluated using the IPAC abstracts test set in the En-Is direction

5. Results

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

Our submission ended in the 3rd-9th place in the Is-En direction in the human evaluation.

The human evaluation of the En-Is direction is still pending.

These results are prior to temporal ensembling or fine-tuning.

6. Conclusion and future work

The resulting models are well capable of translation between Icelandic and English.

Note that despite the task being described as sentence-level, parts of the test set contained two concatenated sentences. This caused our model to poorly model the latter sentence, sometimes even omitting it, resulting in an artificially lower score.

We are currently looking into the benefits of including monolingual Icelandic data in the pre-training phase of the mBART model. Base mBART-25 does not include Icelandic.

The data selected for backtranslation will be expanded and the translation context extended.