State-of-the-Art Polish–Ukrainian Machine Translation

Helena Skowrońska, Paweł Wnuk
ShopAI sp. z o.o.

Introduction

Machine Translation between Polish and Ukrainian is a pressing need due to the large (and still rising) number of Ukrainians living in Poland.

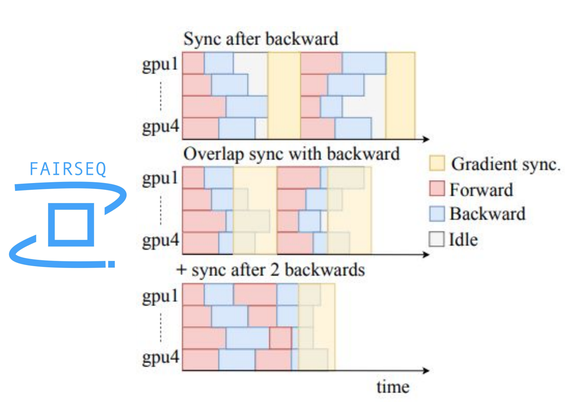
Unfortunately, the MT of this language pair still remains to be solved as an AI problem. For instance, Google Translate achieves just 12.9 BLEU points for PL > UK and 12.8 for UK > PL, both results calculated on 2 k sentences from the PL–UK part of the OpenSubtitles dataset.
Such low BLEU scores indicate that the results are largely unusable.

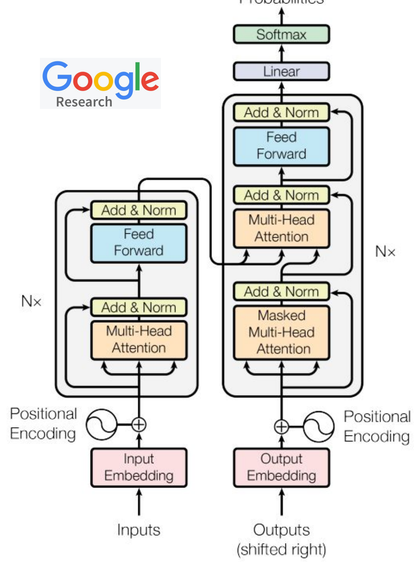
Framework

Based on our current experience with training MT, we selected the ModernMT framework because of its superb results, efficiency and ease of use.

ModernMT is a neural MT system implementing the Fairseq Transformer model. It optimizes the hyperparameters and runs the entire experiment. Training times vary from below 1 day to 2.5 days on a single GPU, depending on the dataset size.

Fairseq is a toolkit for various language generation tasks, implementing many neural models (including several Transformers) and reducing the training time by the utilization of gradient accumulation.





Experiments

To improve the results, we experimented with tokenization, corpus augmentation, and pivot translation.

As to pivot translation, we tried translating through:
- English (PL>EN>UK and UK>EN>PL), because English data are widely available;
- Russian (PL>RU>UK, UK>RU>PL), because Russian data are also widely available, and the language is linguistically related to both UK and PL, even mostly sharing the Cyrillic alphabet with UK.

We achieved success with pivot translation through the Russian language, even surpassing the results provided by Google Translate.

BLEU scores

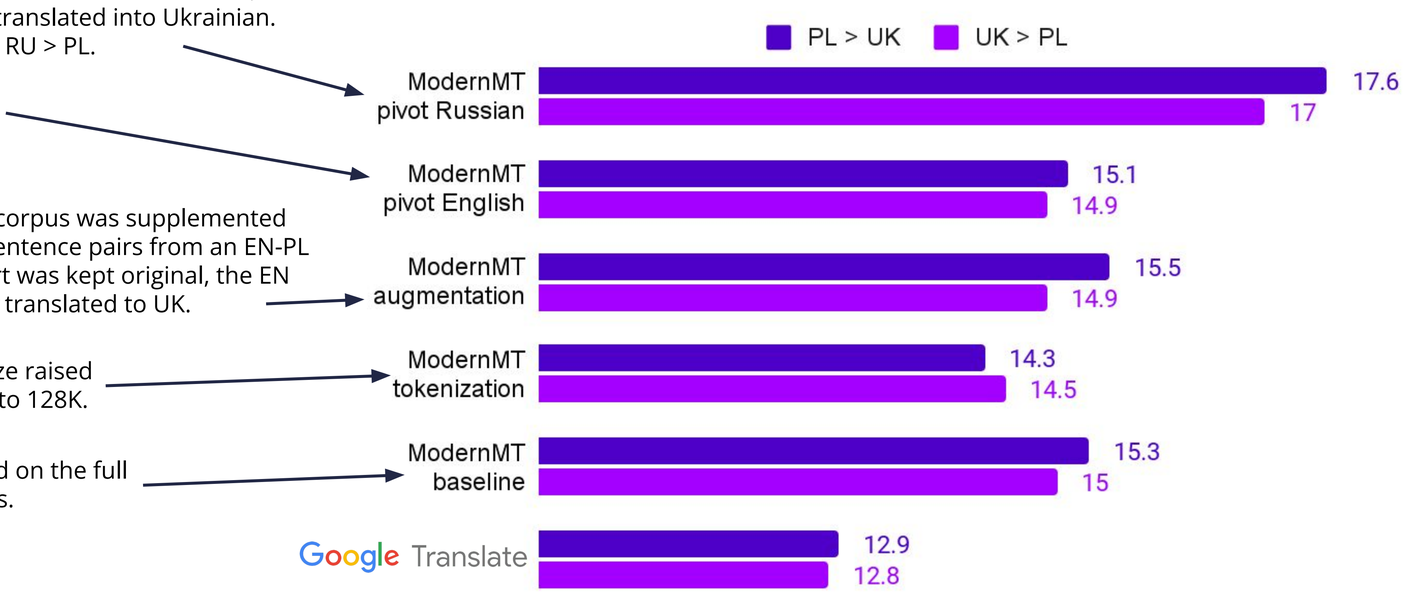
Polish text is first translated into Russian, and the result is then translated into Ukrainian. Analogously, UK > RU > PL.

PL > EN > UK
UK > EN > PL

The 10M training corpus was supplemented with 35M UK-PL sentence pairs from an EN-PL corpus; the PL part was kept original, the EN was automatically translated to UK.

BPE vocabulary size raised from original 33K to 128K.

ModernMT trained on the full 10M UK–PL corpus.



BLEU on OpenSubtitles

UK > PL example

Original: Вам пощастило, бар відкритий для вас.
Golden: Ma pan szczęście, że wpuścimy pana do baru.

Macie szczęście, że bar jest otwarty.

Masz szczęście, bar jest otwarty dla ciebie.

Masz szczęście, że bar jest otwarty dla Ciebie.

Masz szczęście, bar jest otwarty dla Ciebie.

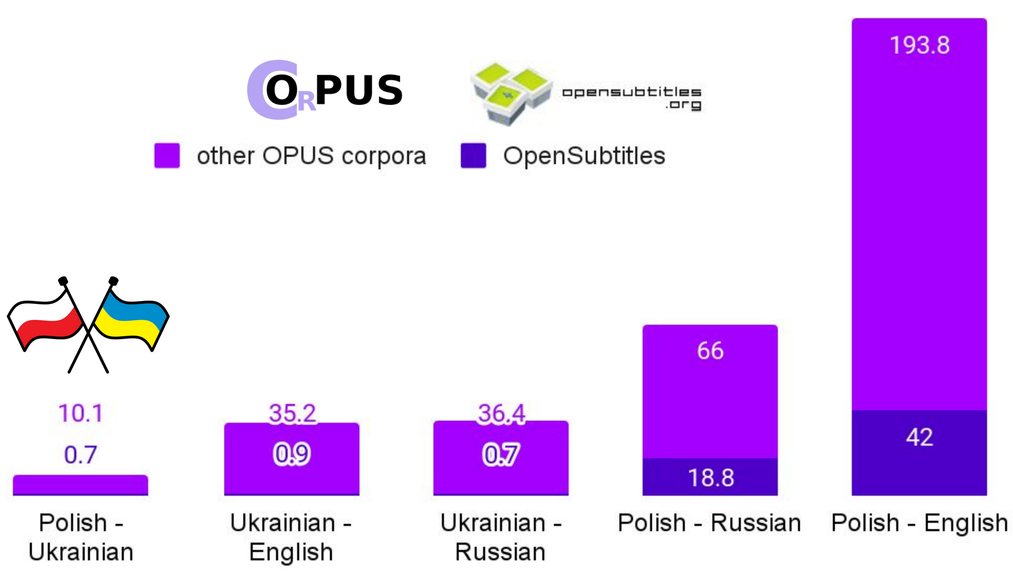
Masz szczęście, bar jest dla Ciebie otwarty.

Na szczęście dla ciebie bar jest dla ciebie otwarty.

Datasets

The initial low quality of the translations has been partially caused by the relative unavailability of training corpora, as both Polish and Ukrainian are low-resource languages. We managed to use only about 10 M PL–UK sentence pairs from open corpora (mainly OPUS). This was almost 20 times fewer than what we used in our MT experiments for the Polish–English language pair.

Pivot translation through Russian allowed us to leverage the potential of datasets that are a few times larger than the 10 M PL–UK corpus:
- 36 M UK–RU corpora,
- 66 M PL–RU corpora.



Conclusions

Among all our approaches to bidirectional PL–UK translation, we achieved success only with using Russian as a pivot language. This was thanks to the wide availability of Russian corpora, as well as the close linguistic proximity between the three languages, especially UK and RU. Our results have surpassed those provided by Google Translate.
To improve the results even further, we're experimenting with the following ideas:
- expanding the PL–UK datasets,
- domain adaptation,
- various approaches to subword tokenization,
- improved dataset preprocessing.

References

ModernMT: github.com/modernmt/modernmt; www.modernmt.eu
Fairseq: M. Ott, S. Edunov, A. Baevski et al. 2019. fairseq: A Fast, Extensible Toolkit for Sequence Modeling. *NAACL-HLT*; github.com/facebookresearch/fairseq
Transformer: A. Vaswani, N. Shazeer, N. Parmar et al. 2017. Attention Is All You Need. *NIPS*.
OPUS: J. Tiedemann. 2012. Parallel Data, Tools and Interfaces in OPUS. *LREC*; opus.nlpl.eu

Acknowledgements

This work was supported by the NCBR grant number POIR.01.01.01-00-1598/20.