Neural Translation for the European Union (NTEU) Project

L. Bié*, A. Cerdà-i-Cucó*, H. Degroote*, A. Estela*,
M. García-Martínez*, M. Herranz*, A. Kohan*, M. Melero‡,
T. O'Dowd§, S. O'Gorman§, M. Pinnis†, R. Rozis†,
R. Superbo§, A. Vasiļevskis†*

‡Barcelona Supercomputing Center maite.melero@bsc.es

§KantanMT{sineadog,tonyod,riccardos}@kantanmt.com

*Pangeanic-PangeaMT{l.bie,a.cerda,h.degroote,a.estela,m.garcia,m.herranz,a.kohan}@pangeanic.com

†Tilde{roberts.rozis,marcis.pinnis,arturs.vasilevskis}@tilde.lv

Abstract

The Neural Translation for the European Union (NTEU) project aims to build a neural engine farm with all European official language combinations for eTranslation,[1] without the necessity to use a high-resourced language as a pivot. NTEU started in September 2019 and will run until August 2021.

1 Introduction

Normally, data for translation are available in English from or to another language. With a few exceptions, all eTranslation MT engines include English as either source or target. Thus, to translate between two non-English languages, English must be used as a pivot.

The NTEU partners, Pangeanic,[2] Tilde,[3] KantanMT[4] and SEAD[5], have been awarded EU funds to build direct machine translation (MT) engines between any of the 24 EU official languages (e.g. Spanish to German, Croatian to Italian, Greek to Polish, etc.) without pivoting through English (around 550 translation engines in total).

2 Approach

NTEU will provide a capacity service to eTranslation by building a near-human-professional-quality neural engine farm which includes all EU language combinations. State-of-the-art technologies such as the transformer (Vaswani et al., 2017) architecture will be implemented. Moreover, lower-resourced languages (for example, Irish or Maltese) will be a challenge, and more effort will be required to obtain well-performing engines for them. In order to obtain the best results, we will experiment with techniques to supplement the original data, such as generating synthetic data by doing back-translation (Sennrich et al., 2016), checking of sentence alignments, transfer learning (Zoph et al., 2016) and unsupervised learning on a monolingual corpus (Artetxe et al., 2019).

In addition to providing the trained engines, the NTEU consortium will gather and clean data from all language combinations so that the engines can be retrained with other technologies in the future. As part of the national digital data gathering efforts, NTEU will also act as a bridge between previous efforts, putting to work the results of the ELRC[6] repository and other European data gathering efforts such as the NEC TM[7] and ParaCrawl[8] projects. Therefore, the project will promote the free flow of data between public administrations themselves and EU bodies.

3 Evaluation methodology

The results of the MT will be manually evaluated in an open-source platform created by the consortium. Following industry and WMT[9] practices for human evaluation, the evaluation dataset has been carefully chosen so as to represent real-world, whole, human-translated documents, purposely excluded from the training data. Human

*All authors have contributed equally to this work.

[1] https://ec.europa.eu/info/resources-partners/machine-translation-public-administrations-etranslation_en

[2] https://pangeamt.com/

[3] https://www.tilde.com/

[4] https://kantanmt.com/

[5] Spanish Agency for Digital Advancement, dependant on the Ministry of Economy

[6] https://elrc-share.eu/

[7] https://www.nec-tm.eu/

[8] https://paracrawl.eu/

[9] http://www.statmt.org/wmt19/

evaluators will score the MT for fluency, adequacy and fit-for-purpose, in a range between 0 and 100. The human evaluator —who will be a native speaker of the target language, but not necessarily knowledgeable of the source language— will be presented with the reference translation in context, plus two automatic translations (one from our engines and one from a third party state-of-the-art system), both unidentified. Each translation will be evaluated by two different human evaluators. The aim of the evaluation platform is to make human evaluation of MT faster, more efficient and more consistent. Automatic evaluation using a larger dataset, also excluded from training, is foreseen as well. The objective is to extensively reduce the time required to build and approve production-ready MT engines with continuous evaluator feedback that reinforces data selection and training procedures.

4 Conclusion

In conclusion, the NTEU project will yield a state-of-the-art neural MT engine farm, which will include all EU language combinations without pivoting through English for eTranslation. Special emphasis will be put on the low-resourced language combinations. The collected data will be provided so it can be used in future technologies. Moreover, an open-source platform will be developed to facilitate human evaluation of the MT engines. Finally, the project will be promoted between public administrations and will put to work previous European data-gathering initiatives.

Acknowledgements

The work reported in this paper was conducted during the NTEU project, which was funded by Innovation and Network Executive Agency (INEA) through grant N^{o} INEA/CEF/ICT/A2018/1816500 as part of the EU's CEF Telecommunications Program.

References

Artetxe, Mikel, Gorka Labaka, and Eneko Agirre. 2019. Unsupervised neural machine translation, a new paradigm solely based on monolingual text. *Procesamiento del Lenguaje Natural*, 63:151–154.

Sennrich, Rico, 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, August. Association for Computational Linguistics.

Vaswani, Ashish, 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*, pages 5998–6008.

Zoph, Barret, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas, November. Association for Computational Linguistics.