Man against Machine: Qualitative Comparison of Original, Translated and Post-edited Wikipedia Articles

Kadri Vare
University of Tartu, Estonia
Kadri.vare@ut.ee

Mark Fišel
University of Tartu, Estonia
mark.fishel@ut.ee

Martin Luts
Tilde Eesti OÜ
martin.luts@tilde.ee

Arvi Tavast
qlaara
arvi@tavast.ee

Sirli Zupping
University of Tartu, Estonia
sirli.zupping@ut.ee

Abstract

In 2017 Estonia started a nationwide project Miljon+[1], to get the Estonian Wikipedia among the Wikipedia versions with over 1 million articles. The project focuses on increasing the size of new articles in Estonian as well as increasing the size of translated articles. This paper describes one possible way how machine translation (MT) could speed up and support the process of reaching to a million Wikipedia articles in Estonian.

1 Introduction

In April 2017 the Center of Estonian Language Resources (CELR) organized user involvement workshops in conjunction with the Estonian Language Technology Conference 2017. One of the events was a joint hackathon with the M+ Project – machine translating Wikipedia articles from English to Estonian, which brought together wikipedists (among them many experts from different fields), translators, linguists and language technologists.

One of CELR's missions is to develop Estonian language technology, also CELR is responsible for archiving all digital resources and language technology tools created within the National Programme for Estonian Language Technology (NPELT). CELR is a CLARIN centre in Estonia and all accomplishments (both software and resources) of NPELT are thereby usable to all users of CLARIN.

The main motivation behind this hackathon was to test whether machine translation can create an article which could be easily post-edited to conform to Wikipedia's quality standards.

2 Overview of the machine translation hackathon

The M+ Project expects every Estonian citizen to contribute new Wikipedia articles and to edit those already present, so we didn't invite only professional translators, but rather simulated the ideal M+ situation. There were around 30 translators with different backgrounds and 5 Wikipedia-experienced judges who were asked to blindly assess articles. The hackathon was organized as follows: wikipedists selected a set of articles on similar topics (bibliographical articles about researchers and scientists) and with approximately similar length (one page) in English which hadn't already been translated to Estonian. Then translators were divided equally between two participating MT systems and each translator was able to translate one article from scratch and post-edit one of the machine translated articles. The participants recorded the time in minutes each article took them.

Judges had two assignments – to guess whether an article was human or machine translated and to give each article a subjective quality score from 1 to 5. As the control group, judges were randomly

[1] http://www.miljonpluss.ut.ee/

given articles that were already published in Wikipedia and had undergone the normal quality procedures there.

3 Overview of the machine translation hackathon

During the event outputs of two machine translation systems were used for post-editing: a custom one tuned on Wikipedia parallel texts and a general-domain one. Both MT systems are based on the encoder-decoder neural machine translation approach with an attention mechanism (Bahdanau et al., 2015), using byte-pair encoding (Sennrich et al., 2016b) to keep the vocabulary size fixed and get rid of unknown input and output tokens.

3.1 Machine translation system from University of Tartu

The tuned system was created at the Institute of Computer Science, University of Tartu. It is trained on a mix of publicly available general-domain corpora as well as in-domain parallel texts extracted from English-Estonian Wikipedia articles.

Wikipedia articles are sources of comparable texts and thus to use them as material for training SMT/NMT systems parallel texts have to be extracted. We used the LEXACC tool (Ştefănescu et al., 2012) to do the extraction and as a result we had an in-domain parallel corpus of 300 000 English-Estonian sentence pairs.

The general-domain data was taken from multilingual corpora, the biggest of which are Europarl (Koehn, 2005), OPUS OpenSubtitles (Tiedemann, 2009) and DCEP (Hajlaoui et al., 2014). The resulting general-domain data consisted of 15.4 million sentence pairs (183.5 million English / 143.5 million Estonian tokens).

We used Nematus (REF) for training the NMT system and AmuNMT for running it in production. After initial training on the combination of in-domain and general-domain data we tuned the system to the Wikipedia domain by continuing training on in-domain data only, a method introduced by Luong and Manning (2015).

3.2 Machine translation system from Tilde

Within the scope of The National Programme for Estonian Language Technology[2], a language technology company Tilde[3] trained English-Estonian neural machine translation (NMT) system using the following training data: Sentence pairs in parallel corpora (filtered & unique): 62,611,065 (21,900,622).

For training of the NMT systems we use the sub-word neural machine translation toolkit Nematus[4] (Sennrich et al., 2016a) that is based on the toolkit dl4mt-tutorial[5]. The toolkit allows training attention-based encoder-decoder models with gated recurrent units. For word segmentation in sub-word units, we use the byte pair encoding (BPE) tools from the toolkit subword-nmt[6] (Sennrich et al., 2016b).

Prior to training of the NMT models, we pre-processed the training data using the following data pre-processing techniques: corpora cleaning, filtering, nontranslatable token identification, tokenisation and truecasing. Finally, we trained NMT models using a vocabulary size of 100,000 segments (99,500 for byte pair encoding). All other parameters were set to the default parameters used by the developers of Nematus for their WMT2016 submissions[7].

[2] https://www.keeletehnoloogia.ee/en
[3] https://tilde.com/
[4] https://github.com/rsennrich/nematus
[5] https://github.com/nyu-dl/dl4mt-tutorial
[6] https://github.com/rsennrich/subword-nmt
[7] https://github.com/rsennrich/wmt16-scripts/blob/master/sample/config.py

4 Results

The judges were not able to guess the provenance of articles: 61 of 120 assignments between the three available categories (MT, human translation, existing article) were wrong.

As expected, the control group of existing articles ended up at near the top of the quality scale in blind assessment. However, the best articles from both human and machine translation reached a comparable level. The placement of machine translated articles at the lower end of the scale is also somewhat expectable, but overall the two translation methods gave very similar results.

One striking difference between the translation methods is in the correlation between effort and quality. In human translation, quality seems to have a cost in time: translating the best article took about ten times more than the fastest article. The time spent on post-editing MT, however, was not correlated at all with the quality of the resulting article. Post-editing the best machine translated articles was faster than translating from scratch with the same quality. At the other end of the quality scale, it was faster to produce a bad translation from scratch than to use MT.

5 Conclusion and Future Work

Bad translations are unsuitable for publication regardless of whether MT is used or not, and MT actually makes their production slower. Publication-grade articles can be obtained using both translation methods, with post-editing MT being slightly faster than making a good translation from scratch.

The MT systems were not customised to Wikipedia texts, including the Wiki syntax and links, therefore the current results can be taken as a lower bound to the practical usability of MT in Wikipedia.

Most of the participants had no experience or knowledge in MT post-editing, apart from the 5-minute introduction given before the test. Using professionals may give different results.

We did not control for the quality of the translation source texts. It is possible that pre-editing or even pre-selection of material for machine translation would have a visible effect on the results.

References

Bahdanau, D., Cho, K., & Bengio, Y. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. – Proceedings of the International Conference on Learning Representations (ICLR). Retrieved from http://arxiv.org/abs/1409.0473.

Hajlaoui, N., Kolovratnik, D., Väyrynen, J., Steinberger, R., and Varga, D. 2014. DCEP – Digital Corpus of the European Parliament. – LREC, pp. 3164–3171.

Junczys-Dowmunt, M., Dwojak, T., and Hoang H. 2016. Is neural machine translation ready for deployment? A case study on 30 translation directions. – Proceedings of IWSLT, Seattle, WA, USA.

Koehn, P. 2005. Europarl: A parallel corpus for statistical machine translation. – Proceedings of the 10th MT Summit (pp. 79–86), Phuket, Thailand.

Luong, M.-T., and Manning, C. D. 2015. Stanford neural machine translation systems for spoken language domains. – Proceedings of the International Workshop on Spoken Language Translation.

Sennrich, R., Haddow, B., & Birch, A. 2016a. Edinburgh Neural Machine Translation Systems for WMT 16. – arXiv preprint arXiv:1606.02891.

Sennrich, R., Haddow, B., & Birch, A. 2016b. Neural Machine Translation of Rare Words with Subword Units. – Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016). Berlin, Germany: Association for Computational Linguistics.

Ştefănescu, D., Ion, R., and Hunsicker, S. 2012. Hybrid Parallel Sentence Mining from Comparable Corpora. – Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT 2012), pp. 28–30, Trento, Italy.

Tiedemann, J. 2009. News from OPUS-A collection of multilingual parallel corpora with tools and interfaces. – Recent advances in natural language processing (Vol. 5, pp. 237–248). Borovets, Bulgaria.