TraMOOC - Translation for Massive Open Online Courses: Recent Developments in Machine Translation

Rico Sennrich and **Antonio Valerio Miceli Barone**
University of Edinburgh
rico.sennrich@ed.ac.uk, amiceli@inf.ed.ac.uk

Joss Moorkens and **Sheila Castilho** and **Andy Way** and **Federico Gaspari**
ADAPT Centre
{joss.moorkens, sheila.castilho}@adaptcentre.ie, {away, fgaspari}@computing.dcu.ie

Valia Kordoni and **Markus Egg** and **Maja Popovic**
Humboldt-Universität zu Berlin
{evangelia.kordoni, markus.egg}@anglistik.hu-berlin.de, popovicm@hu-berlin.de

Yota Georgakopoulou and **Maria Gialama**
Deluxe Media Europe
{yota.georgakopoulou, maria.gialama}@bydeluxe.com

Menno van Zaanen
Tilburg University
mvzaanen@uvt.nl

Abstract

Massive open online courses have been growing rapidly in size and impact. TraMOOC[1] aims at developing high-quality translation of all types of text genre included in MOOCs from English into eleven European and BRIC languages that are hard to translate into and have weak MT support.

1 Recent developments

In TraMOOC, we have developed machine translation prototypes for 11 target languages, from English into German, Italian, Portuguese, Dutch, Bulgarian, Greek, Polish, Czech, Croatian, Russian, and Chinese. The translation systems are based on phrase-based SMT and neural machine translation. The latter has achieved state-of-the-art performance in recent evaluation campaigns (Bojar, 2016). We use the Nematus toolkit (Sennrich, 2017) for training; the translation server is based on the amuNMT toolkit (Junczys-Dowmunt et al., 2016). The translation systems have been adapted to MOOC texts via fine-tuning of the model parameters on in-domain training data to maximize translation quality on this domain.

We have also completed a comparative human evaluation of phrase-based SMT and NMT for four language pairs to compare educational domain output from both systems using a variety of metrics. These include automatic evaluation, human rankings of adequacy and fluency, error-type markup, and technical and temporal post-editing effort. The results show a preference for NMT in side-by-side ranking for all language pairs, texts, and segment lengths. In addition, perceived fluency is improved and annotated errors are fewer in the NMT output. However, results are mixed for some error categories. Despite far fewer segments requiring post-editing, document-level post-editing performance was not found to have significantly improved when using NMT in this study, suggesting that NMT may not show an enormous improvement over SMT when used in a production scenario. We have subsequently prepared data and a slightly amended quality evaluation methodology to apply to all TraMOOC NMT systems later in 2017.

References

Bojar, Ondřej et al. 2016. Findings of the 2016 Conference on Machine Translation. In *Proceedings of the First Conference on Machine Translation*, pages 131–198, Berlin, Germany. Association for Computational Linguistics.

Junczys-Dowmunt, Marcin, Tomasz Dwojak, and Hieu Hoang. 2016. Is neural machine translation ready for deployment? a case study on 30 translation directions. In *Arxiv*.

Sennrich, Rico et al. 2017. Nematus: a Toolkit for Neural Machine Translation. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 65–68, Valencia, Spain.

[1]TraMOOC is a H2020 Innovation Action project funded by the European Commission (H2020-ICT-2014-1-ICT-17-2014/644333) and runs from February 2015 to February 2018. For more details on the project, please, visit http://www.tramooc.eu