MTUOC: easy and free integration of NMT systems in professional translation environments

Antoni Oliver
Universitat Oberta de Catalunya
aoliverg@uoc.edu

Abstract

In this paper the MTUOC project, aiming to provide an easy integration of neural and statistical machine translation systems, is presented. Almost all the required software to train and use neural and statistical MT systems is released under free licences. However, their use is not always easy and intuitive and medium-high specialized skills are required. MTUOC project provides simplified scripts for preprocessing and training MT systems, and a server and client for easy use of the trained systems. The server is compatible with popular CAT tools for a seamless integration. The project also distributes some free engines.

1 Introduction

MTUOC is a project from the Arts and Humanities department at the Universitat Oberta de Catalunya (UOC) to facilitate the use and integration of neural and statistical machine translation systems.

Most of the software needed for training and using such systems is distributed under free permissive licences. So this technology is, in principle, freely available for any professional, company or organization. The use of MT toolkits presents some problems:

- *Technological skills*: medium-high technological skills are required. Knowledge of some programming (as Python, for example) and scripting (as Bash, for example) languages are necessary. On the other hand, the documentation of these toolkits is not always detailed enough and some time in trial and error is spent.

- Integration: the resulting systems are not easily integrable in existing workflows. Most of the toolkits provide access through some kind of API, usually using a server-client configuration. Some CAT tools offer plugins to access some existing systems. But not all CAT Tool - MT system combinations are available.

- *Hardware*: relatively high hardware requirements are present, especially for training the systems. For training SMT systems lots of RAM memory is required. For training NMT systems one or more powerful GPU units are compulsory.

MTUOC tries to offer solutions for the first two problems. Regarding the *technological skills problem*, it provides a series of easy-to-use and easy-to-understand Python and Bash scripts for corpus pre-processing and training. All these scripts are well documented and can be adapted and extended in an easy way. Regarding the *integration problem*, a fully configurable server and client are provided. The server can mimic the behaviour of several kinds of servers, so it can be used with a large range of CAT tools. For example, the server can use a Marian engine but behave as a Moses server so it can be directly integrated with OmegaT. The client can deal with several widely used file formats (as XLIFF, for example) and generate TMX translation memories that can be used in any CAT tool. Regarding the *hardware problem*, several facts should be borne in mind. Firstly, hardware requirements for training are much harder than for translating. Once an engine is trained, it can be

used in any consumer computer. So many potential users can benefit from the freely available engines. Several providers can offer the service of training tailored machine translation systems. UOC can sign technology-transfer agreements with companies and organizations to train tailored systems at very competitive rates. This service is free for NGOs. Secondly, the price of hardware is getting lower over time and powerful GPU units are now available at affordable prices.

2 Components

The MTUOC project offers six main components:

- *Python modules*: providing several functionalities as tokenization, truecasing, etc.
- *Scripts* written in Python and Bash and several configuration files:
  - Corpus pre-processing scripts: to process the training corpora for training the systems
  - Training scripts and configuration files: for several SMT and NMT toolkits
  - Evaluation scripts: providing some widely used MT evaluation metrics (as BLEU, NIST, WER, TER, edit distance)
- *MTUOC Server*: the component that receives a segment to translate from the client or CAT tool, process it (tokenization, truecasing and so on) and sends it to the translation server. After receiving the translation this component post-processes it (detruecasing, detokenization and so on) and sends it back to the client or CAT tool.
- *MTUOC Client*: this component can handle several translation formats, send segments to the server, receive the translations and create the translated file.
- *MTUOC Virtual Machine*: as most toolkits work under Linux, this virtual machine is useful for Windows users to run all the required components.
- *Pre-trained translation engines* that can be freely used with MTUOC

3 MT Toolkits

MTUOC-server can be used with the following MT toolkits.

- Moses[1] (Koehn et al., 2007)
- Marian[2] (Junczys-Dowmunt et al., 2018)
- OpenNMT[3] (Klein et al., 2017)
- ModernMT[4] (Bertoldi et al., 2018)

4 Obtaining MTUOC

All the components of MTUOC can be downloaded from its SourceForge page.[5] The documentation of the systems is available in the Wiki space of the project page. All the MTUOC components are released under a free licence, namely GNU GPL version 3.

Acknowledgements: The training of the neural MT systems distributed by MTUOC has been possible thanks to the NVIDIA GPU grant programme.

References

Bertoldi, Nicola, Davide Caroselli, and Marcello Federico. 2018. The ModernMT project.

Junczys-Dowmunt, Marcin, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia, July. Association for Computational Linguistics.

Klein, Guillaume, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proc. ACL*.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions*, pages 177–180.

[1] http://www.statmt.org/moses/
[2] https://marian-nmt.github.io/
[3] https://opennmt.net/
[4] https://github.com/modernmt/modernmt
[5] https://sourceforge.net/projects/mtuoc/