OPUS-MT – Building open translation services for the World

Jörg Tiedemann
Department of Digital Humanities
HELDIG
University of Helsinki

Santhosh Thottingal
Wikimedia Foundation

1 Introduction

Equality among people requires, among other things, the ability to access information in the same way as others independent of the linguistic background of the individual user. Achieving this goal becomes an even more important challenge in a globalized world with digital channels and information flows being the most decisive factor in our integration in modern societies. Language barriers can lead to severe disadvantages and discrimination not to mention conflicts caused by simple misunderstandings based on broken communication. Linguistic discrimination leads to frustration, isolation and racism and the lack of technological language support may also cause what is known as the *digital language death* (Kornai, 2013).

Machine translation (MT) has developed into a useful tool that diminishes and partially removes such language barriers. Modern MT engines enable people to communicate, to access information in foreign languages and to build efficient resources for new communities. The mission of OPUS-MT[1] is to provide open translation services and tools that are free from commercial interests and restrictions. The idea is to make automatic translation accessible for anyone in a transparent and secure way without exploitation plans and hidden agendas compromising privacy and placing marketing strategies. We also want to focus on the support of minority and low-resource languages with the aim to introduce a community effort for the benefit of all.

OPUS-MT has successfully launched its first pilot system and currently collaborates with the Wikimedia foundation in the setup of translation services for the production of Wikipedia content in new languages based on more elaborated resources available in, e.g. English. Currently, the project provides over 1,000 pre-trained translation models that are free to download and use. OPUS-MT also contains open-source software for launching translation services as web applications. The on-going effort focuses on the improvement of translation quality, language coverage and emphasizes specific test cases to study the applicability of the approach. More details about the implementation and current status of the project are given below.

2 OPUS-MT models

The models that we train are based on state-of-the-art transformer-based neural machine translation (NMT). We apply Marian-NMT[2] in our framework, a stable production-ready NMT toolbox with efficient training and decoding capabilities (Junczys-Dowmunt et al., 2018). Our models are trained on freely available parallel corpora collected in the large bitext repository OPUS[3] (Tiedemann, 2012). The architecture is based on a standard transformer setup with 6 self-attentive layers in both, the encoder and decoder network with 8 attention heads in each layer. The hyper-parameters follow the general recommendations given in the documentation of the software. All the details can be seen in the training procedures that we also release as open source in our GitHub repository.[4]

OPUS-MT supports both, bilingual as well as multilingual models. For the latter, we apply the language label approach proposed by (Johnson et al., 2017). Our package implements generic

[1] `https://github.com/Helsinki-NLP/Opus-MT`

[2] `https://marian-nmt.github.io`

[3] `http://opus.nlpl.eu`

[4] `https://github.com/Helsinki-NLP/Opus-MT-train`

model	BLEU	chrF_2
English–Finnish	22.9	0.548
+ back-translation	23.7	0.562
+ fine-tuning	25.7	0.578

Table 1: Test results for the English–Finnish OPUS-MT model based on the news translation task from WMT 2019. Fine-tuning was done using the English–Finnish news translation test sets from earlier years.

procedures that make it easy to train a large number of translation models from the existing data in the OPUS collection. The procedures take care of proper pre-processing and training setups to enable batch-processes without the immediate need for further adjustments. We try to reduce the burden of time-consuming optimization and focus on rather generic models for the time being in order to quickly achieve a good language coverage without significantly compromising translation quality that can be achieved.

We use common benchmarks and test sets that are extracted on the fly from held-out data to monitor the quality of the NMT models. Test sets and results are released together with the models, pre- and post-processing scripts and basic information about their usage. The table of currently supported language pairs can be accessed on-line.[5]

We also develop generic fine-tuning and data augmentation procedures that can be used to further improve the translation models. We implemented a pipeline for backtranslation of Wikimedia content (coming from Wikipedia, Wikibooks, Wikisource, etc.) to augment existing training data. Backtranslation is known to significantly boost performance and to enable simple domain adaptation based on in-domain target language data. Furthermore, we also provide procedures for fine-tuning that can adjust model parameters according to some small in-domain data set, another successful strategy for domain adaptation. The impact of fine-tuning and backtranslation can be seen on the example of the English–Finnish OPUS-MT model listed in Table 1.

3 OPUS-MT servers

Finally, we also provide simple web applications that can be used to launch translation services based on the pre-trained models. The most straightforward setup is implemented as a dockerized Tornado-besed web application that can be set up with a few simple commands. The configuration can be adjusted and extended to serve any bilingual translation model that we provide. Each service can accommodate several language pairs and may connect multiple servers. The current implementation is based on CPU-based decoding as a cost-efficient setup for every-day users but it should be adjustable to a GPU-based setup without major changes. A running service demonstrating the app is hosted by the Wikimedia foundation at `https://opusmt.wmflabs.org`.

Another websocket based application is also provided, which enables the support of multilingual models, a simple translation cache and the retrieval of token alignment information, which is supported by most models that we train with the guided alignment feature of Marian-NMT. Further improvements of the web applications are planned once we have finished our tests of the current implementation in a production environment for selected test cases.

4 Conclusions

This paper presents OPUS-MT, a project that focuses on the development of free resources and tools for machine translation. The current status is a repository of over 1,000 pre-trained neural MT models that are ready to be launched in on-line translation services. For this we also provide open-source implementations of web applications that can run efficiently on average desktop hardware with a straightforward setup and installation.

The work is supported by the European Union's Horizon 2020 projects FoTran and MeMAD as well as by the Finnish IT Center for Science, CSC.

References

Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., et al. (2017). Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association of Computational Linguistics*, 5(1):339–351.

Junczys-Dowmunt, M., Grundkiewicz, R., Dwojak, T., Hoang, H., Heafield, K., Neckermann, T., Seide, F., Germann, U., Aji, A. F., Bogoychev, N., Martins, A. F. T., and Birch, A. (2018). Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121.

Kornai, A. (2013). Digital language death. *PloS one*, 8(10). :e77056.

Tiedemann, J. (2012). Parallel data, tools and interfaces in OPUS. In *Proceedings of LREC*, Istanbul, Turkey.

[5] `http://opus.nlpl.eu/Opus-MT/`