Chapter 24
Open Translation Models, Tools and Services

Jörg Tiedemann, Mikko Aulamo, Sam Hardwick, and Tommi Nieminen

Abstract The ambition of the Open Translation Models, Tools and Services (OPUS-MT) project is to develop state-of-the art neural machine translation (NMT) models that can freely be distributed and applied in research as well as professional applications. The goal is to pre-train translation models on a large scale on openly available parallel data and to create a catalogue of such resources for streamlined integration and deployment. For the latter we also implement and improve web services and computer-assisted translation (CAT) tools that can be used in on-line interfaces and professional workflows. Furthermore, we want to enable the re-use of models to avoid repeating costly training procedures from scratch and with this contribute to a reduction of the carbon footprint in MT research and development. The ELG pilot project focused on European minority languages and improved translation quality in low resource settings and the integration of MT services in the ELG infrastructure.

1 Overview and Objectives of the Pilot Project

OPUS-MT (Tiedemann and Thottingal 2020) provides ready-made server solutions that can be deployed on regular desktop machines to run translations using any NMT model that has been released through the project.[1] The service is powered by Marian-NMT[2] (Junczys-Dowmunt et al. 2018), an efficient open-source framework written in pure C++ with implementations of state-of-the-art neural machine translation architectures. OPUS-MT provides two implementations that can be deployed on regular Ubuntu servers or through containerised solutions using docker images. Both solutions can easily be configured using JSON and can be deployed with a wide range of OPUS-MT models. Multiple translation services and nodes can be combined in one access point through a lightweight API. The coverage is constantly growing and

Jörg Tiedemann · Mikko Aulamo · Sam Hardwick · Tommi Nieminen
University of Helsinki, Finland, jorg.tiedemann@helsinki.fi, mikko.aulamo@helsinki.fi, sam.hardwick@helsinki.fi, tommi.nieminen@helsinki.fi

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

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

G. Rehm (ed.), *European Language Grid*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-17258-8_24

improved models are continuously released through our repository as a result of our on-going model training efforts.

A dockerised web app is implemented using the Tornado Python framework, which we adapted for the integration into the European Language Grid environment providing an interface that can seamlessly be deployed in the ELG infrastructure. The essential metadata records for the ELG service catalogue are generated from pre-defined templates using information available from released translation models. The routines support bilingual as well as multilingual models and can also be used to set up access points that serve several translation services. Appropriate docker images are compiled using installation recipes and scripts. We host them on Docker Hub from where they can be pulled by ELG requests to serve translation requests directly through the online APIs. Detailed deployment documentation is available from the repository.[3]

At the time of writing, OPUS-MT provides 89 registered MT services within ELG including a wide variety of bilingual and multilingual models. Registered services can be tested online and can also be accessed through the web API and ELG Python SDK. The translation runs on regular CPUs with minimal resource requirements thanks to the efficient decoder implementation in Marian-NMT. Multilingual models are handled in a special way: multiple source languages can be handled by a single access point whereas multiple target languages require separate access points. Metadata records include the relevant information to describe the service provided.

We also developed plugins for professional translation workflows under the label of OPUS-CAT[4] (Nieminen 2021). Our tools include a local MT engine that can run on regular desktop machines making MT available without the security and confidentiality risks associated with online services. OPUS-CAT integrates with popular translation software such as Trados Studio, memoQ, OmegaT and Memsource. It also provides an integrated fine-tuning procedure for domain adaptation. All OPUS-MT models can be downloaded and used locally with the MT engine, some of the plugins can also fetch translations directly from the OPUS-MT services in ELG.

2 Increasing Language Coverage

The general goal of OPUS-MT is to increase language coverage of freely available machine translation solutions. The project already provides over a thousand pre-trained translation models covering hundreds of languages in various translation directions. The ongoing effort is documented by public repositories and regular updates and we omit further details here as this is a quickly moving target.

Within our ELG pilot project, we further developed our pipelines and recipes to systematically train additional NMT models. The effort resulted in the model de-

[3] https://github.com/Helsinki-NLP/Opus-MT/tree/master/elg

[4] https://helsinki-nlp.github.io/OPUS-CAT/

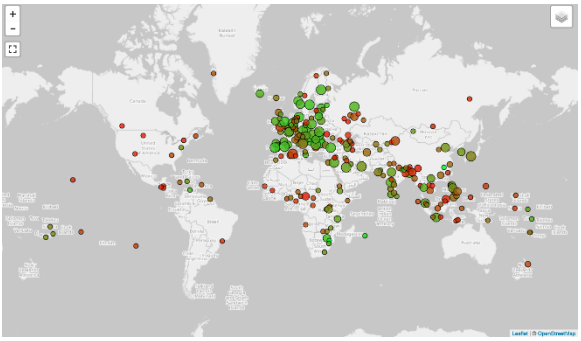

Fig. 1 OPUS-MT map: A visualisation of language coverage and model quality according to automatic evaluation metrics and the Tatoeba MT challenge benchmarks; here: models that translate from a source language mapped on their glottolog location to English; larger circles indicate bigger benchmark test sets and the color scale goes from green (high quality) to red (poor quality)

velopment framework OPUS-MT-train[5] with support for bilingual and multilingual models that can be trained on data provided by OPUS[6] and the Tatoeba translation challenge[7] (Tiedemann 2020).

In order to keep track of the development, we heavily rely on the Tatoeba benchmarks and we implemented an interactive tool to visualize the current state of our released models. Figure 1 shows an example screenshot.

The geographic distribution of released models is an appealing way to uncover blind spots in the NLP landscape. The lack of appropriate data resources is one of the major bottlenecks that block the development of proper MT solutions for most language pairs of the world. Another issue is the narrow focus of research that typically overemphasises well established tasks for reasons of comparability and measurable success. OPUS-MT does not have a strict state-of-the-art development focus based on major benchmarks but rather emphasises language coverage and the focus on under-researched translation directions. The OPUS-MT map and the Tatoeba MT challenge try to make this work visible and more attractive.

The main strategy to tackle issues with *limited data resources* is to apply transfer learning and some type of data augmentation. In OPUS-MT we are constantly facing the problem of limited training data and noise and the ELG pilot project specifically focused on low-resource scenarios and European minority languages.

The idea of transfer learning is based on the ability of models to pick up valuable knowledge from other tasks or languages. In MT, the main type of transfer learning is based on cross-lingual transfer where multilingual translation models can be used to push the performance in low-resource settings (Fan et al. 2021). The effect is typically pronounced with closely related languages where strong linguistic similarities can lead to big improvements across language boundaries (Tiedemann 2021).

[5] https://github.com/Helsinki-NLP/OPUS-MT-train

[6] https://opus.nlpl.eu

[7] https://github.com/Helsinki-NLP/Tatoeba-Challenge/

In OPUS-MT, we therefore focused on multilingual models of typologically related languages. In our setup, we rely on language groups and families established within the ISO 639-5 standard. A dedicated tool for mapping languages to language groups and connecting them with the hierarchical language tree has been developed to allow a systematic development of multilingual NMT models based on typological relationships.[8] The procedures have been integrated in the OPUS-MT training recipes and can be applied to arbitrary datasets from the Tataobea MT Challenge.

Table 1 illustrates the effect of cross-lingual transfer with multilingual models on the example of the Belarusian-English translation benchmark from the Tatoeba MT Challenge. All models apply the same generic transformer-based architecture (Vaswani et al. 2017) with identical hyper-parameters and training recipes.

NMT model	Belarusian ⟶ English	English ⟶ Belarusian
Belarusian – English	10.0	8.2
East Slavic – English	38.7	20.8
Slavic – English	**42.7**	**22.9**
Indo-European – English	41.7	18.1

Table 1 Machine translation between Belarusian and English with different NMT models; scores refer to BLEU scores measured on the Tatoeba MT Challenge benchmark

The bilingual baseline model is very poor due to the limited training data that is available from the Tatoeba dataset (157,524 sentence pairs). Augmenting the training data with closely related languages such as other (East) Slavic languages leads to significant improvements, which is not very surprising. The effect can be seen in both directions. Note that the multi-target models need to be augmented by language tokens to indicate the output language to be generated. The importance of systematic benchmarks is also shown in the table where we can see that Indo-European language model struggles and the effect of positive transfer diminishes due to the capacity issues of such a complex model setup.

Finally, we also tested a novel type of data augmentation using a rule-based system (RBMT) for back-translation (Sennrich et al. 2016) to produce additional data for the translation from Finnish to Northern Sámi (Aulamo et al. 2021). Our results revealed that knowledge from the RBMT system can effectively be injected into a neural MT model significantly boosting the performance as shown in Table 2.

We use two benchmarks in our evaluations: the UiT set[9], and the YLE set of 150 sentence pairs from news stories about Sámi culture.[10] Preliminary manual evaluation revealed that the NMT-based model was often unable to correctly translate proper names. Adding copies of monolingual data as suggested by Currey et al. (2017) helps to alleviate that issue. Furthermore, we also added experiments with subword regularisation (Kudo 2018) and data tagging (Caswell et al. 2019) to bet-

[8] https://github.com/Helsinki-NLP/LanguageCodes

[9] 2,000 sentence pairs sampled from the Giellatekno Free corpus https://giellatekno.uit.no

[10] Collected from https://yle.fi

	Training Data	UiT	YLE
Baseline	25,106	18.9	4.3
+ NMT-bt	422,596	34.0	9.8
+ RBMT-bt	378,567	36.3	15.5
+ NMT-bt + RBMT-bt	885,301	**40.1**	10.8
+ NMT-bt + copy	845,192	35.7	12.5
+ RBMT-bt + copy	757,134	35.7	**18.6**
+ NMT-bt + RBMT-bt + SR + TB	885,301	40.0	17.2

Table 2 Training data sizes (sentence pairs) and results (BLEU) for the Finnish-Northern Sámi translation models using original parallel data (Baseline), augmented data with back-translations from NMT and RBMT systems (NMT-bt, RBMT-bt), added monolingual data (copy), subword regularisation (SR) and tagged back-translations (TB) evaluated on the UiT and YLE test sets

ter exploit the distributions in the training data and to distinguish between sources with different noise levels. Preliminary results are encouraging and deserve further investigations. In future work, we plan to add pivot-based translation and multilingual models to further improve the performance of the system, to support additional input languages and to include other Sámi language varieties, too.

3 Conclusions and Results of the Pilot Project

OPUS-MT is an on-going effort to make MT widely available for open research and development with an extensive language coverage and well established deployment and integration procedures. Our ELG pilot project made it possible to strengthen the focus on minority languages and to further exploit transfer and data augmentation strategies to improve the quality of MT for under-resourced language pairs.

Acknowledgements The work described in this article has received funding from the EU project European Language Grid as one of its pilot projects. We would also like to acknowledge the support by the FoTran project funded by the European Research Council (no. 771113) and CSC, the Finnish IT Center for Science, for computational resources.

References

Aulamo, Mikko, Sami Virpioja, Yves Scherrer, and Jörg Tiedemann (2021). “Boosting Neural Machine Translation from Finnish to Northern Sámi with Rule-Based Backtranslation”. In: *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*. Reykjavik, Iceland: Linköping University Electronic Press, pp. 351–356. URL: https://aclanthology.org/2021.nodalida-main.37.

Caswell, Isaac, Ciprian Chelba, and David Grangier (2019). “Tagged Back-Translation”. In: *Proc. of the Fourth Conf. on Machine Translation*, pp. 53–63.

Currey, Anna, Antonio Valerio Miceli Barone, and Kenneth Heafield (2017). "Copied Monolingual Data Improves Low-Resource Neural Machine Translation". In: *Proceedings of the Second Conference on Machine Translation*. Copenhagen, Denmark: ACL, pp. 148–156. DOI: 10.18653/v1/W17-4715. URL: https://aclanthology.org/W17-4715.

Fan, Angela, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Michael Auli, and Armand Joulin (2021). "Beyond English-Centric Multilingual Machine Translation". In: *Journal of Machine Learning Research* 22.107, pp. 1–48. URL: http://jmlr.org/papers/v22/20-1307.html.

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*. Melbourne, Australia: ACL, pp. 116–121. URL: http://www.aclweb.org/anthology/P18-4020.

Kudo, Taku (2018). "Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates". In: *Proc. of the 56th Annual Meeting of the Association for Computational Linguistics*, pp. 66–75.

Nieminen, Tommi (2021). "OPUS-CAT: Desktop NMT with CAT integration and local fine-tuning". In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*. ACL, pp. 288–294. DOI: 10.18653/v1/2021.eacl-demos.34. URL: https://aclanthology.org/2021.eacl-demos.34.

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)*, pp. 86–96.

Tiedemann, Jörg (2020). "The Tatoeba Translation Challenge – Realistic Data Sets for Low Resource and Multilingual MT". In: *Proceedings of the Fifth Conference on Machine Translation (WMT)*. ACL, pp. 1174–1182. URL: https://aclanthology.org/2020.wmt-1.139.

Tiedemann, Jörg (2021). "The Development of a Comprehensive Data Set for Systematic Studies of Machine Translation". In: *Multilingual Facilitation*. Ed. by Mika Hämäläinen, Niko Partanen, and Khalid Alnajjar. Finland: University of Helsinki, pp. 248–262. DOI: 10.31885/9789515150257.

Tiedemann, Jörg and Santhosh Thottingal (2020). "OPUS-MT – Building open translation services for the World". In: *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation (EAMT)*. Lisboa, Portugal: European Association for Machine Translation, pp. 479–480. URL: https://helda.helsinki.fi/bitstream/handle/10138/327852/2020.eamt_1_499.pdf.

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: *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 6000–6010.