OpenNMT: Open-Source Toolkit for Neural Machine Translation

Guillaume Klein†, Yoon Kim*, Yuntian Deng*, Jean Senellart†, Alexander M. Rush*
Harvard University*, SYSTRAN †

Technologies

OpenNMT is an ecosystem based on multiple technologies and frameworks:

- OpenNMT: the original full-featured project in *LuaTorch*, focusing on maintainability, user support, and production.
- OpenNMT-py: a *PyTorch* clone of OpenNMT, focusing on research and modularity.
- CTranslate: an inference engine for OpenNMT models in C++ and *Eigen*, focusing on embedded and production environments.

Features

OpenNMT implements many additional features on top of the baseline model:

- factored translation for richer text representation;
- tokenization and data preparation tools;
- model variants: bidirectional encoder, convolutional encoder, variational dropout, etc.;
- learning rate decay strategies;
- advanced model retraining and adaptation;
- beam search normalization;
- . . . and many more!

Tasks

OpenNMT supports other tasks than machine translation:

- Sequence tagging.
- Language modeling.
- Speech-to-text, using a pyramidal RNN encoder.
- Image-to-text, using a combination of CNN and RNN layers.

For example, *Im2Text* (github.com/OpenNMT/Im2Text) is an extension that can be used for image captioning, optical character recognition, or LaTeX decompilation:

$$Q = (b + 1/b)\rho, \qquad \rho = \frac{1}{2}\sum_{\alpha>0} \alpha,$$

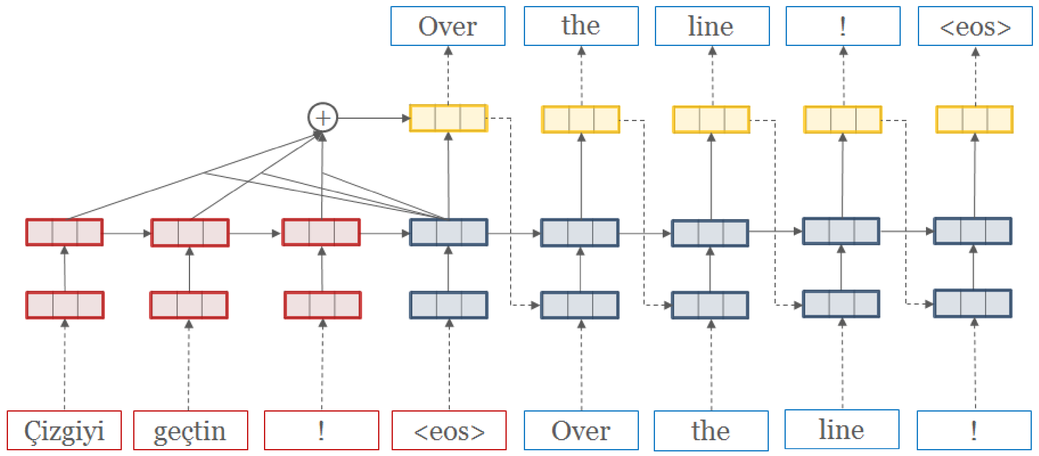

Figure: The standard sequence-to-sequence model.

OpenNMT is an industrial-strength and open-source neural machine translation ecosystem featuring:

- Ready-to-use and highly configurable implementations in *Torch* and *PyTorch*.
- State-of-the-art translation accuracy and competitive training efficiency.
- Extensive set of model and training options covering a large set of needs of academia and industry.
- Extensions to allow other sequence generation tasks such as summarization, image-to-text, and speech-to-text.
- Standalone and dependency-free inference engine in C++.

Neural Machine Translation

Neural machine translation (NMT) is a new methodology for machine translation that has led to remarkable improvements, particularly in terms of human evaluation, compared to rule-based and statistical machine translation (SMT) systems.

OpenNMT implements the attention-based encoder-decoder architecture that models the probability of a target sentence $w_{1:T}$ given a source sentence $x_{1:S}$ as:

$$p(w_{1:T}|x) = \prod_{1}^{T} p(w_t|w_{1:t-1}, x; \theta)$$

This modeling is usually achieved using LSTM recurrent networks which allows long term dependency learning.

State-of-the-art system

OpenNMT implements models and training procedures that achieve competitive results in system comparison, e.g. in the recent WMT 2017 translation task:

System	BLEU-cased
uedin-nmt-ensemble	28.3
LMU-nmt-reranked-wmt17-en-de	27.1
SYSTRAN-single	**26.7**

Table: Best scores on English-German *newstest2017*.

More generally, OpenNMT produces strong baselines with optimized training time and memory requirements.

Additional resources

OpenNMT also provides additional resources including:

- A complete documentation portal (opennmt.net/OpenNMT) for beginners to advanced users describing data preparation, models, training strategies, command line options, etc.
- Visualization tools for debugging or understanding.

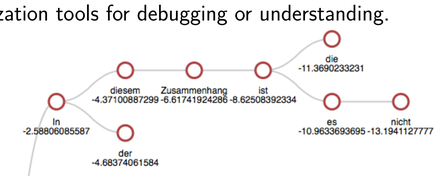

Figure: Beam search visualization

Community

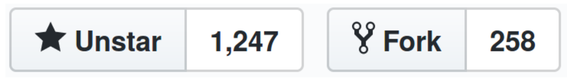

Figure: GitHub statistics

OpenNMT is also a community around machine translation and language modeling. The forum (forum.opennmt.net) counts more than 200 users with daily questions on how to improve or adapt their system and training procedure.

Production environment

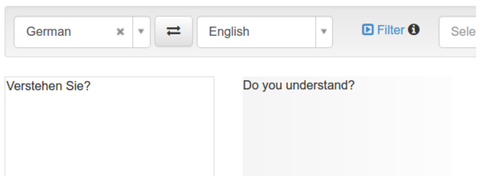

Figure: Live demo of OpenNMT

OpenNMT has proved to be adapted to production settings. SYSTRAN—a major translation services provider—is using OpenNMT for its Pure Neural™ Machine Translation offering which enables higher translation quality in existing services.