OpenNMT: Open-Source Toolkit for Neural Machine Translation

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

Competitive System

OpenNMT achieves competitive results against other systems, 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: Top 3 on English-German *newstest2017*.

OpenNMT is optimized for training time and memory efficiency.

System	Training (WPS)	Inference (WPS)	BLEU
Nematus	3221	252	18.25
OpenNMT	5254	457	19.34

Table: Comparison of source words per second (WPS) on similar model size.

Features

OpenNMT implements many additional features on top of the standard sequence-to-sequence model:

- Model variants: bidirectional encoder, convolutional encoder, variational dropout, attention parameterization, copying mechanisms, etc.
- Factored input representation for richer features.
- Tokenization and data preparation tools.
- Multi-GPU training.
- Beam search length normalization.
- ... and many more!

Extensions

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

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

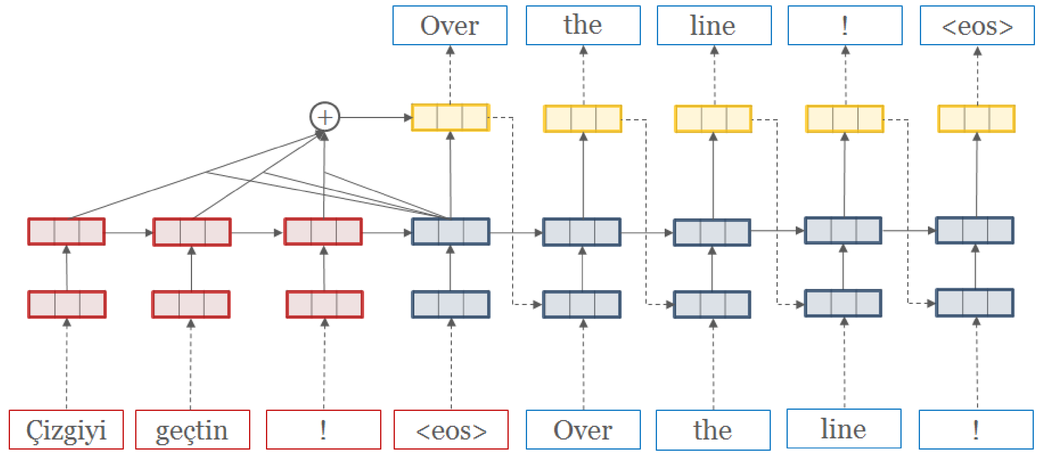

Figure: Sequence-to-sequence model with attention.

OpenNMT is an industrial-strength, 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.
- Implementations of latest advances in neural machine translation and sequence-to-sequence learning.
- Standalone and dependency-free inference engine in C++.
- Rich 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.

Approach

Neural machine translation (NMT) is simple approach for machine translation based on neural networks 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:

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

The encoder and decoder are usually parameterized with LSTM recurrent neural networks.

Platforms

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.

Additional Resources

OpenNMT provides additional resources including:

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

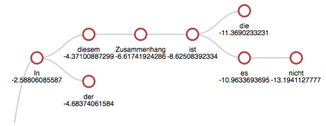

Figure: Beam search visualization

Industry Deployment

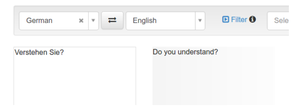

Figure: Live demo of OpenNMT

OpenNMT is robust and has been deployed for production by SYSTRAN, a major translation services provider. SYSTRAN is using OpenNMT for its Pure Neural™ Machine Translation offering which enables higher translation quality in existing services.

Active Community

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 systems and training procedures.

Figure: GitHub statistics as of July 18th, 2017.

A user survey showed that users are equally coming from industry and academia, suggesting that the framework is well-suited for many use cases.

Open-source Landscape

Landscape for neural machine translation (and sequence-to-sequence learning in general) software has been growing and several excellent open-source alternatives exist, such as:

- *Nematus* (github.com/EdinburghNLP/nematus)
- *Marian NMT* (marian-nmt.github.io)
- *Neural Monkey* (github.com/ufal/neuralmonkey)
- *Google seq2seq* (google.github.io/seq2seq)