



# OpenNMT: Open-Source Toolkit for Neural Machine Translation

Guillaume Klein<sup>†</sup>, Yoon Kim\*, Yuntian Deng\*, Jean Senellart<sup>†</sup>, Alexander M. Rush\*  
Harvard University\*, SYSTRAN<sup>†</sup>



## Competitive 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
<b>SYSTRAN-single</b>	<b>26.7</b>

Table: Top 3 on English-German *newstest2017*.

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

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

Table: Comparison of source words per second (WPS).

## Features

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

## Other 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](https://github.com/OpenNMT/Im2Text)) is an extension that can be used for image captioning, optical character recognition, or  $\LaTeX$  decompilation:

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

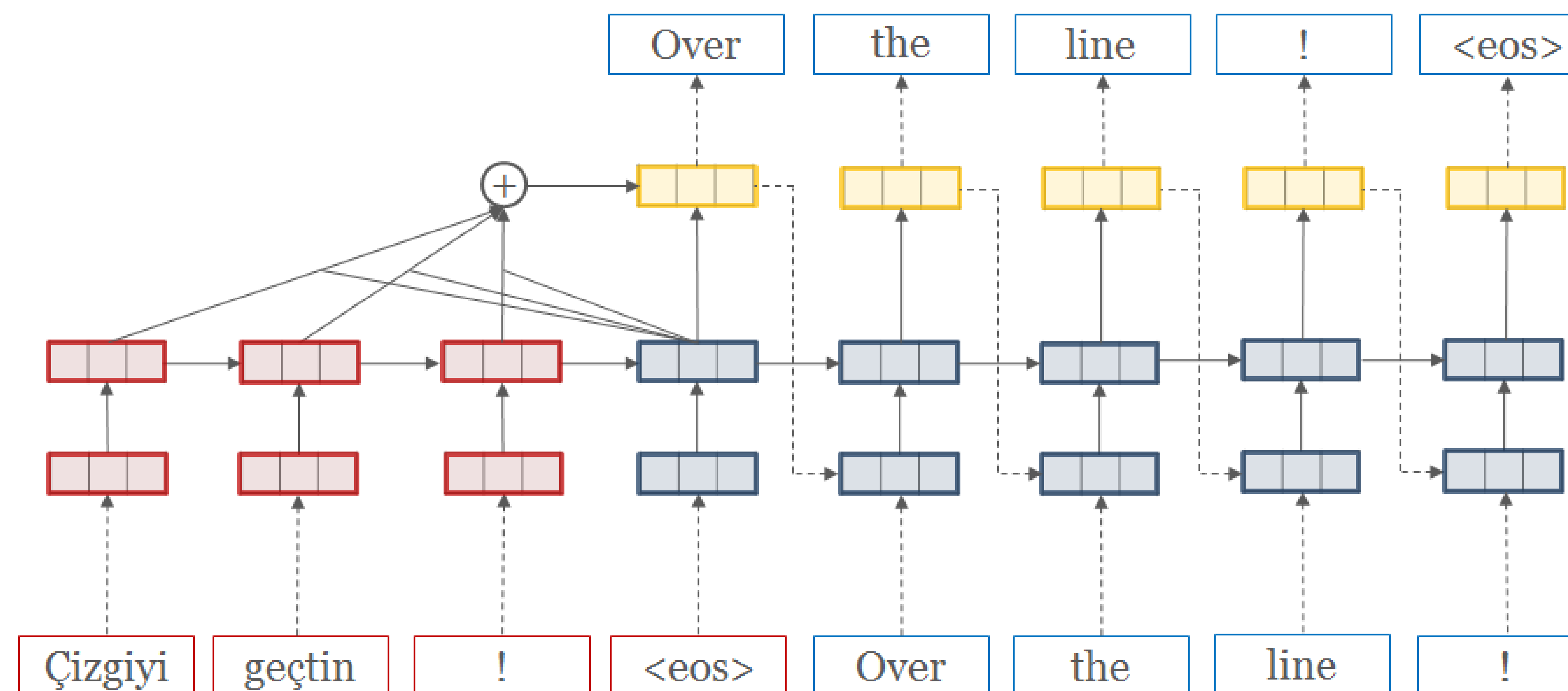


Figure: The standard sequence-to-sequence model.

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++.

## Approach

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:

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

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

## Technologies

OpenNMT is an ecosystem based on multiple technologies and frameworks:

- ▶ **OpenNMT**: the original full-featured project in *Lua Torch*, 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 complete documentation portal ([opennmt.net/OpenNMT](https://opennmt.net/OpenNMT)) for beginners to advanced users describing data preparation, models, training strategies, command line options, etc.
- ▶ Visualization tools for debugging or understanding.

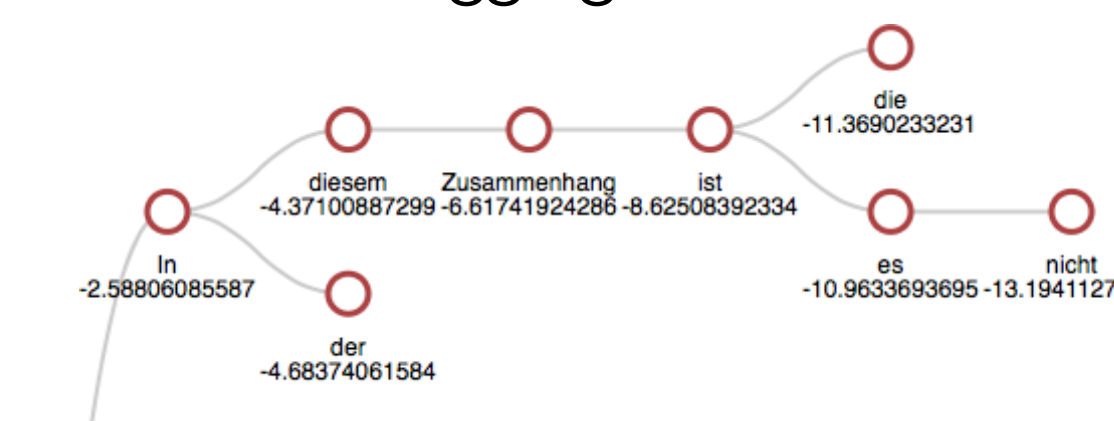


Figure: Beam search visualization

## Production

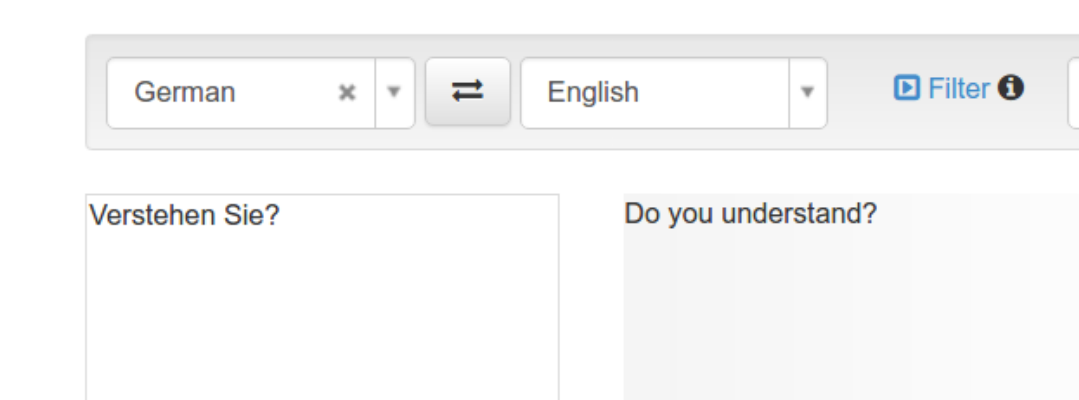


Figure: Live demo of OpenNMT

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.

## Community

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

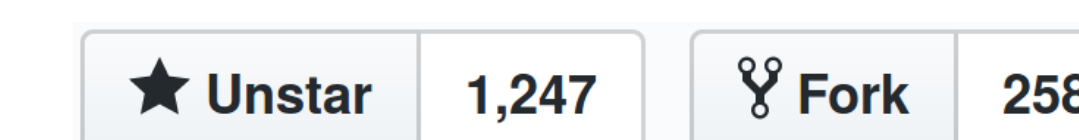


Figure: GitHub statistics as of July 18<sup>th</sup>, 2017.

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

## Open-source landscape

There exist several open-source alternatives that maintain a healthy competition and collaboration. Notable mentions are:

- ▶ *Nematus* ([github.com/EdinburghNLP/nematus](https://github.com/EdinburghNLP/nematus))
- ▶ *Marian NMT* ([marian-nmt.github.io](https://marian-nmt.github.io))
- ▶ *Neural Monkey* ([github.com/ufal/neuralmonkey](https://github.com/ufal/neuralmonkey))
- ▶ *seq2seq* ([google.github.io/seq2seq](https://google.github.io/seq2seq))