BLOG ›

A Neural Network for Machine Translation, at Production Scale

TUESDAY, SEPTEMBER 27, 2016

Posted by Quoc V. Le & Mike Schuster, Research Scientists, Google Brain Team

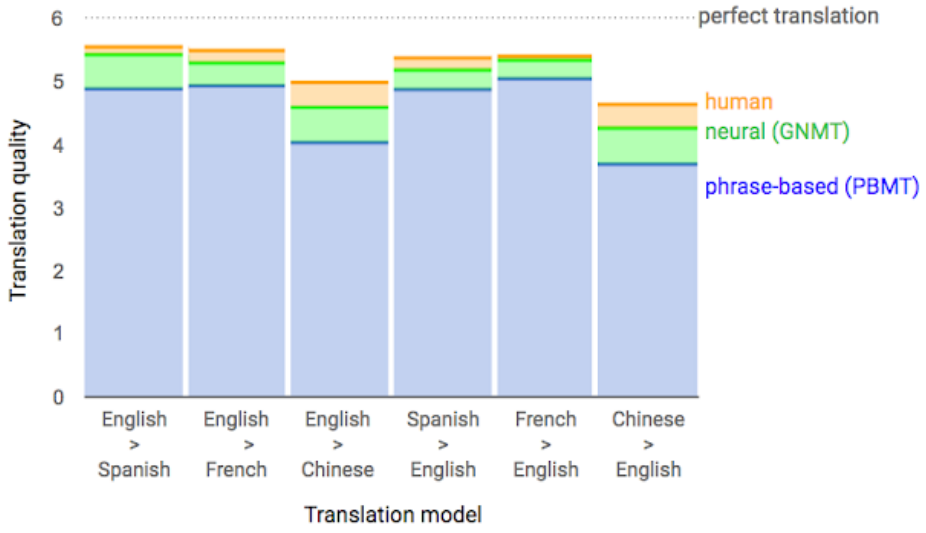
Ten years ago, we announced the **launch of Google Translate**, together with the use of **Phrase-Based Machine Translation** as the key algorithm behind this service. Since then, rapid advances in machine intelligence have improved our **speech recognition** and **image recognition** capabilities, but improving machine translation remains a challenging goal.

Today we announce the Google Neural Machine Translation system (GNMT), which utilizes state-of-the-art training techniques to achieve the largest improvements to date for machine translation quality. Our full research results are described in a new technical report we are releasing today: "*Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation*" [1].

A few years ago we started using **Recurrent Neural Networks** (RNNs) to directly learn the mapping between an input sequence (e.g. a sentence in one language) to an output sequence (that same sentence in another language) [2]. Whereas Phrase-Based Machine Translation (PBMT) breaks an input sentence into words and phrases to be translated largely independently, Neural Machine Translation (NMT) considers the entire input sentence as a unit for translation.The advantage of this approach is that it requires fewer engineering design choices than previous Phrase-Based translation systems. When it first came out, NMT showed equivalent accuracy with existing Phrase-Based translation systems on modest-sized public benchmark data sets.

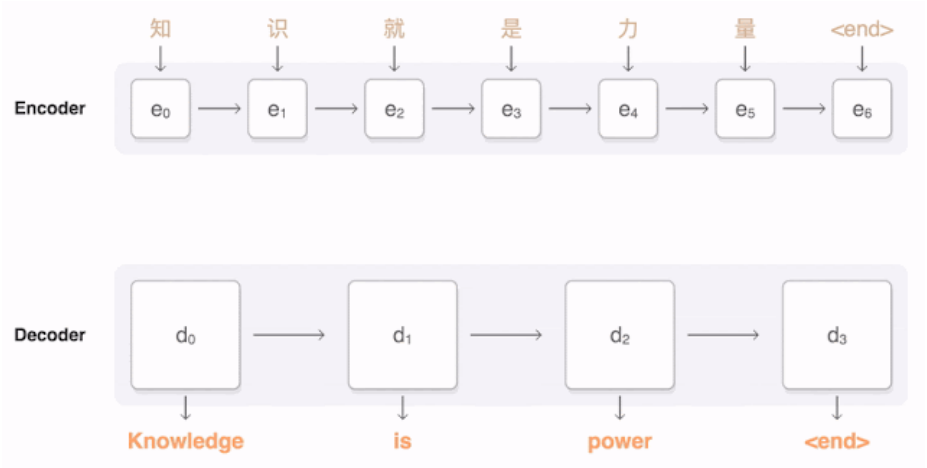
Since then, researchers have proposed many techniques to improve NMT, including work on handling rare words by mimicking an external alignment model [3], using attention to align input words and output words [4] and breaking words into smaller units to cope with rare words [5,6]. Despite these improvements, NMT wasn't fast or accurate enough to be used in a

data sets and built a system that is sufficiently fast and accurate enough to provide better translations for Google's users and services.



Data from side-by-side evaluations, where human raters compare the quality of translations for a given source sentence. Scores range from 0 to 6, with 0 meaning "completely nonsense translation", and 6 meaning "perfect translation."

The following visualization shows the progression of GNMT as it translates a Chinese sentence to English. First, the network encodes the Chinese words as a list of vectors, where each vector represents the meaning of all words read so far ("Encoder"). Once the entire sentence is read, the decoder begins, generating the English sentence one word at a time ("Decoder"). To generate the translated word at each step, the decoder pays attention to a weighted distribution over the encoded Chinese vectors most relevant to generate the English word ("Attention"; the blue link transparency represents how much the decoder pays attention to an encoded word).



phrase based production system. GNMT reduces translation errors by more than 55%-85% on several major language pairs measured on sampled sentences from Wikipedia and news websites with the help of bilingual human raters.

Input sentence:	Translation (PBMT):	Translation (GNMT):	Translation (human):
李克強此行將啟動中加總理年度對話機制，與加拿大總理杜魯多舉行兩國總理首次年度對話。	Li Keqiang premier added this line to start the annual dialogue mechanism with the Canadian Prime Minister Trudeau two prime ministers held its first annual session.	Li Keqiang will start the annual dialogue mechanism with Prime Minister Trudeau of Canada and hold the first annual dialogue between the two premiers.	Li Keqiang will initiate the annual dialogue mechanism between premiers of China and Canada during this visit, and hold the first annual dialogue with Premier Trudeau of Canada.

*An example of a translation produced by our system for an input sentence sampled from a news site. Go **here** for more examples of translations for input sentences sampled randomly from news sites and books.*

In addition to releasing this research paper today, we are announcing the launch of GNMT in production on a notoriously difficult language pair: Chinese to English. The Google Translate mobile and web apps are now using GNMT for 100% of machine translations from Chinese to English—about 18 million translations per day. The production deployment of GNMT was made possible by use of our publicly available machine learning toolkit **TensorFlow** and our **Tensor Processing Units** (TPUs), which provide sufficient computational power to deploy these powerful GNMT models while meeting the stringent latency requirements of the Google Translate product. Translating from Chinese to English is one of the more than 10,000 language pairs supported by Google Translate, and we will be working to roll out GNMT to many more of these over the coming months.

Machine translation is by no means solved. GNMT can still make significant errors that a human translator would never make, like dropping words and mistranslating proper names or rare terms, and translating sentences in isolation rather than considering the context of the paragraph or page. There is still a lot of work we can do to serve our users better. However, GNMT represents a significant milestone. We would like to celebrate it with the many researchers and engineers—both within Google and the wider community—who have contributed to this direction of research in the past few years.

Acknowledgements:
We thank members of the **Google Brain team** and the **Google Translate team** for the help with the project. We thank Nikhil Thorat and the **Big Picture team** for the visualization.

Human and Machine Translation, *Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, Jeffrey Dean. Technical Report, 2016.*

[2] **Sequence to Sequence Learning with Neural Networks**, *Ilya Sutskever, Oriol Vinyals, Quoc V. Le. Advances in Neural Information Processing Systems, 2014.*

[3] **Addressing the rare word problem in neural machine translation**, *Minh-Thang Luong, Ilya Sutskever, Quoc V. Le, Oriol Vinyals, and Wojciech Zaremba. Proceedings of the 53th Annual Meeting of the Association for Computational Linguistics, 2015.*

[4] **Neural Machine Translation by Jointly Learning to Align and Translate**, *Dzmitry Bahdanau, Kyunghyun Cho, Yoshua Bengio. International Conference on Learning Representations, 2015.*

[5] **Japanese and Korean voice search**, *Mike Schuster, and Kaisuke Nakajima. IEEE International Conference on Acoustics, Speech and Signal Processing, 2012.*

[6] **Neural Machine Translation of Rare Words with Subword Units**, *Rico Sennrich, Barry Haddow, Alexandra Birch. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 2016.*

Labels: Google Brain Google Translate Machine Learning Machine Translation TensorFlow

Previous posts

SEP 23, 2016

Show and Tell: image captioning open sourced in

SEP 21, 2016

The 280-Year-Old Algorithm Inside Google Trips

SEP 20, 2016

The 2016 Google Earth Engine User Summit: Turning