重磅 | 谷歌翻译整合神经网络：机器翻译实现颠覆性突破（附论文）

机器之心 2016-09-28

选自Google Research

作者：Quoc V. Le、Mike Schuster

机器之心编译

参与：吴攀

> *昨日，谷歌在 ArXiv.org 上发表论文《Google`s Neural Machine Translation System: Bridging the Gap between Human and Machine Translation》介绍谷歌的神经机器翻译系统（GNMT），当日机器之心就对该论文进行了摘要翻译并推荐到网站（www.jiqizhixin.com）上。今日，谷歌 Research Blog 发布文章对该研究进行了介绍，还宣布将 GNMT 投入到了非常困难的汉语-英语语言对的翻译生产中，引起了业内的极大的关注。*

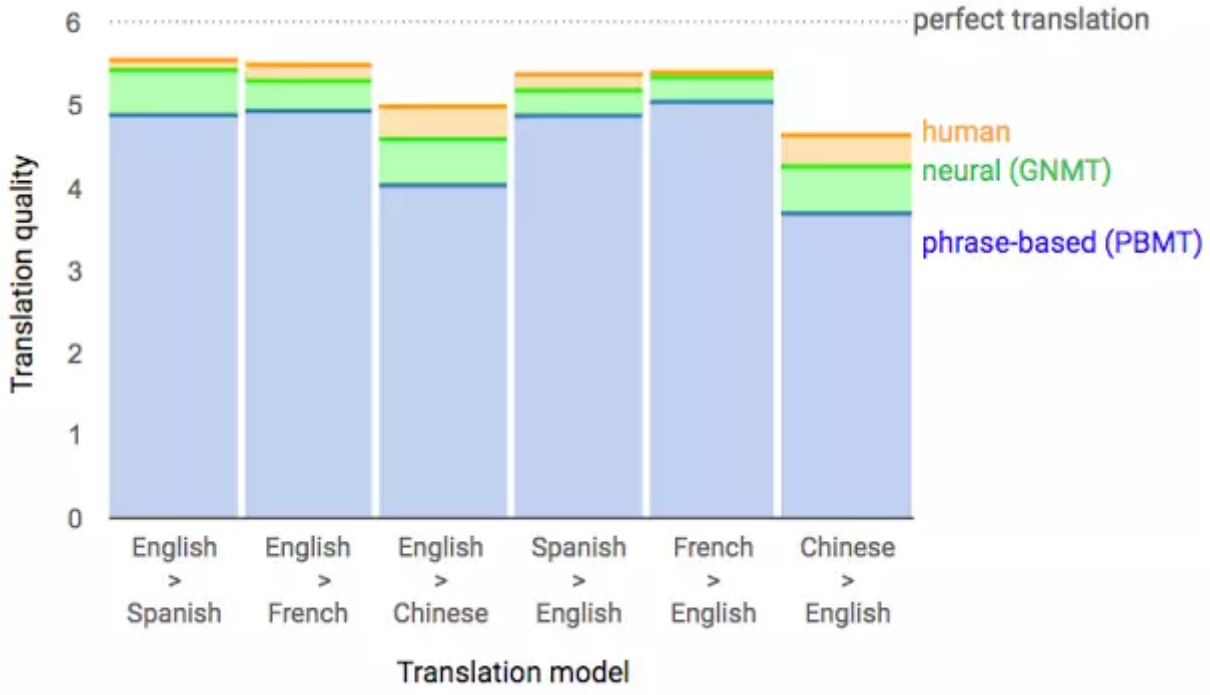
十年前，我们发布了 Google Translate（谷歌翻译），这项服务背后的核心算法是基于短语的机器翻译（PBMT:Phrase-Based Machine Translation）。自那时起，机器智能的快速发展已经给我们的语音识别和图像识别能力带来了巨大的提升，但改进机器翻译仍然是一个高难度的目标。

今天，我们宣布发布谷歌神经机器翻译（GNMT：Google Neural Machine Translation）系统，该系统使用了当前最先进的训练技术，能够实现到目前为止机器翻译质量的最大提升。我们的全部研究结果详情请参阅我们的论文《Google`s Neural Machine Translation System: Bridging the Gap between Human and Machine Translation》（见文末）[1]。

几年之前，我们开始使用循环神经网络（RNN：Recurrent Neural Networks）来直接学习一个输入序列（如一种语言的一个句子）到一个输出序列（另一种语言的同一个句子）的映射 [2]。其中基于短语的机器学习（PBMT）将输入句子分解成词和短语，然后很大程度上对它们进行独立地翻译，而神经机器翻译（NMT）则将整个输入句子视作翻译的基本单元。这种方法的优点是：相比于之前的基于短语的翻译系统，这种方法所需的工程设计更少。当其首次被提出时，NMT 在中等规模的公共基准数据集上就达到了可与基于短语的翻译系统媲美的准确度。

自那以后，研究者已经提出了很多改进 NMT 的技术，其中包括模拟外部对准模型（external alignment model）来处理罕见词 [3]，使用注意（attention）来对准输入词和输出词 [4] 以及将词分解成更小的单

元以应对罕见词 [5,6]。尽管有这些进步，但 NMT 的速度和准确度还没能达到成为 Google Translate 这样的生产系统的要求。我们的新论文 [1] 描述了我们怎样克服了让 NMT 在非常大型的数据集上工作的许多挑战，以及我们如何打造了一个在速度和准确度上都已经足够能为谷歌的用户和服务带来更好的翻译的系统。



来自对比评估的数据，其中人类评估者对给定源句子的翻译质量进行比较评分。得分范围是 0 到 6，其中 0 表示「完全没有意义的翻译」，6 表示「完美的翻译」。

下面的可视化图展示了 GNMT 将一个汉语句子翻译成英语句子的过程。首先，该网络将该汉语句子的词编码成一个向量列表，其中每个向量都表征了到目前为止所有被读取到的词的含义（「编码器（Encoder）」）。一旦读取完整个句子，解码器就开始工作——一次生成英语句子的一个词（「解码器（Decoder）」。为了在每一步都生成翻译正确的词，解码器重点注意了与生成英语词最相关的编码的汉语向量的权重分布（「注意（Attention）」，蓝色链接的透明度表示解码器对一个被编码的词的注意程度）。

使用人类评估的并排比较作为一项标准，GNMT 系统得出的翻译相比于之前的基于短语的生产系统实现了极大的提升。在双语人类评估者的帮助下，我们在来自维基百科和新闻网站的样本句子上测定发现：GNMT 在多个主要语言对的翻译中将翻译误差降低了 55%-85% 以上。

我们的系统产出一个翻译案例，其输入句子采样自一个新闻网站。这个地址（https://drive.google.com/file/d/0B4-Ig7UAZe3BSUYweVo3eVhNY3c/view?usp=sharing）可以看到更多随机采样自新闻网站和书籍的输入句子翻译样本。

今天除了发布这份研究论文之外，我们还宣布将 GNMT 投入到了一个非常困难的语言对（汉语-英语）的翻译的生产中。现在，移动版和网页版的 Google Translate 的汉英翻译已经在 100% 使用 GNMT 机器翻译了——每天大约 1800 万条翻译。GNMT 的生产部署是使用我们公开开放的机器学习工具套件 TensorFlow 和我们的张量处理单元（TPU：Tensor Processing Units），它们为部署这些强大的 GNMT 模型提供了足够的计算算力，同时也满足了 Google Translate 产品的严格的延迟要求。汉语到英语的翻译是 Google Translate 所支持的超过 10000 种语言对中的一种，在未来几个月，我们还将继续将我们的 GNMT 扩展到远远更多的语言对上。

机器翻译还远未得到完全解决。GNMT 仍然会做出一些人类翻译者永远不出做出的重大错误，例如漏词和错误翻译专有名词或罕见术语，以及将句子单独进行翻译而不考虑其段落或页面的上下文。为了给我们的用户带来更好的服务，我们还有更多的工作要做。但是，GNMT 代表着一个重大的里程碑。我们希望与过去几年在这个研究方向上有所贡献的许多研究者和工程师一起庆祝它——不管是来自谷歌还是更广泛的社区。

Google Brain 团队和 Google Translate 团队都参与了该项目。Nikhil Thorat 和 Big Picture 也帮助了该项目的可视化工作。

- **论文：Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation**

摘要：神经机器翻译（NMT: Neural Machine Translation）是一种用于自动翻译的端到端的学习方法，该方法有望克服传统的基于短语的翻译系统的缺点。不幸的是，众所周知 NMT 系统的训练和翻译推理的计算成本非常高。另外，大多数 NMT 系统都难以应对罕见词。这些问题阻碍了 NMT 在实际部署和服务中的应用，因为在实际应用中，准确度和速度都很关键。我们在本成果中提出了 GNMT——谷歌的神经机器翻译（Google's Neural Machine Translation）系统来试图解决许多这些问题。我们的模型由带有 8 个编码器和 8 个解码器的深度 LSTM 网络组成，其使用了注意（attention）和残差连接（residual connections）。为了提升并行性从而降低训练时间，我们的注意机制将解码器的底层连接到了编码器的顶层。为了加速最终的翻译速度，我们在推理计算过程中使用了低精度运算。为了改善对罕见词的处理，我们将词分成常见子词（sub-word）单元（词的组件）的一个有限集合，该集合既是输入也是输出。这种方法能提供「字符（character）」-delimited models 的灵活性和「词（word）」-delimited models 的有效性之间的平衡、能自然地处理罕见词的翻译、并能最终提升系统的整体准确度。我们的波束搜索技术

（beam search technique）使用了一个长度规范化（length-normalization）过程，并使用了一个覆盖度惩罚（coverage penalty），其可以激励很可能能覆盖源句子中所有的词的输出句子的生成。在 WMT'14 英语-法语和英语-德语基准上，GNMT 实现了可与当前最佳结果媲美的结果。通过在一个单独的简单句子集合的人类对比评估中，它相比于谷歌已经投入生产的基于短语的系统的翻译误差平均降低了 60%。

参考文献：

[1] Google's Neural Machine Translation System: Bridging the Gap between 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.

阅读原文