# **机器翻译重大突破：中英翻译已达人类水平**

由微软亚洲研究院与雷德蒙研究院的研究人员组成的团队昨晚宣布，其研发的****机器翻译系统在通用新闻报道测试集****

****newstest2017http://statmt.org/wmt17/translation-task.html）的中-英测试集上，达到了可与人工翻译媲美的水平****。

和语音识别相比，判断机器翻译的准确率是一件非常困难的事，即使是两位专业的翻译人员对于完全相同的句子也会有略微不同的翻译，而且两个人的翻译都不是错的，因为****表达同一个句子的“正确的”方法不止一种****。

为了确保翻译结果准确且达到人类的翻译水平，微软研究团队还特意邀请了双语语言顾问将微软的翻译结果与两个独立的人工翻译结果进行比较评估。

****据悉，这是首个在新闻报道的翻译质量和准确率上可以比肩人工翻译的翻译系统。****

****▌微软的“秘诀”****

虽然学术界和产业界的科研人员致力于机器翻译研究很多年，但近两年深度神经网络的使用让机器翻译的表现取得了很多实质性突破，翻译结果相较于以往的统计机器翻译结果更加自然流畅。为了能够取得中-英翻译的突破，****来自微软亚洲研究院和雷德蒙研究院的三个研究组，采用了对偶学习、推敲网络、联合训练和一致性规范这四种技术****。

****其中，微软亚洲研究院机器学习组将他们的最新研究成果——对偶学习（Dual Learning）和推敲网络（Deliberation Networks）应用在了此次取得突破的机器翻译系统中****。

对偶学习利用的是人工智能任务的天然对称性。当我们将其应用在机器翻译上时，效果就好像是通过自动校对来进行学习——当我们把训练集中的一个中文句子翻译成英文之后，系统会将相应的英文结果再翻译回中文，并与原始的中文句子进行比对，进而从这个比对结果中学习有用的反馈信息，对机器翻译模型进行修正。

而推敲网络则类似于人们写文章时不断推敲、修改的过程。通过多轮翻译，不断地检查、完善翻译的结果，从而使翻译的质量得到大幅提升。

****在此次的系统模型中增加了另外两项新技术：联合训练（JointTraining）和一致性规范（AgreementRegularization），以提高翻译的准确性。****

联合训练可以理解为用迭代的方式去改进翻译系统，用中英翻译的句子对去补充反向翻译系统的训练数据集，同样的过程也可以反向进行。一致性规范则让翻译可以从左到右进行，也可以从右到左进行，最终让两个过程生成一致的翻译结果。

****▌四大技术详解****

****对偶学习（Dual Learning）****：对偶学习的发现是由于现实中有意义、有实用价值的人工智能任务往往会成对出现，两个任务可以互相反馈，从而训练出更好的深度学习模型。例如，在翻译领域，我们关心从英文翻译到中文，也同样关心从中文翻译回英文；在语音领域，我们既关心语音识别的问题，也关心语音合成的问题；在图像领域，图像识别与图像生成也是成对出现。此外，在对话引擎、搜索引擎等场景中都有对偶任务。

一方面，由于存在特殊的对偶结构，两个任务可以互相提供反馈信息，而这些反馈信息可以用来训练深度学习模型。也就是说，即便没有人为标注的数据，有了对偶结构也可以做深度学习。另一方面，两个对偶任务可以互相充当对方的环境，这样就不必与真实的环境做交互，两个对偶任务之间的交互就可以产生有效的反馈信号。因此，****充分地利用对偶结构，就有望解决深度学习和增强学习的瓶颈——训练数据从哪里来、与环境的交互怎么持续进行等问题。****

论文地址：https://papers.nips.cc/paper/6469-dual-learning-for-machine-translation.pdf



对偶无监督学习框架

****推敲网络（Deliberation Networks）****：“推敲”二字可以认为是来源于人类阅读、写文章以及做其他任务时候的一种行为方式，即任务完成之后，并不当即终止，而是会反复推敲。

微软亚洲研究院机器学习组将这个过程沿用到了机器学习中。推敲网络具有两段解码器，其中第一阶段解码器用于解码生成原始序列，第二阶段解码器通过推敲的过程打磨和润色原始语句。后者了解全局信息，在机器翻译中看，它可以基于第一阶段生成的语句，产生更好的翻译结果。

论文地址：https://www.microsoft.com/en-us/research/publication/deliberation-networks-sequence-generation-beyond-one-pass-decoding/



推敲网络的解码过程

****联合训练（Joint Training）：****这个方法可以认为是从源语言到目标语言翻译（Source toTarget）的学习与从目标语言到源语言翻译（Target to Source）的学习的结合。

中英翻译和英中翻译都使用初始并行数据来训练，在每次训练的迭代过程中，中英翻译系统将中文句子翻译成英文句子，从而获得新的句对，而该句对又可以反过来补充到英中翻译系统的数据集中。同理，这个过程也可以反向进行。****这样双向融合不仅使得两个系统的训练数据集大大增加，而且准确率也大幅提高。****

论文地址：https://arxiv.org/pdf/1803.00353.pdf



联合训练：从源语言到目标语言翻译（Source to Target）P(y|x) 与从目标语言到源语言翻译（Target to Source）P(x|y)

****一致性规范（Agreement Regularization）：****翻译结果可以从左到右按顺序产生，也可以从右到左进行生成。

该规范对从左到右和从右到左的翻译结果进行约束。如果这两个过程生成的翻译结果一样，一般而言比结果不一样的翻译更加可信。****这个约束，应用于神经机器翻译训练过程中，以鼓励系统基于这两个相反的过程生成一致的翻译结果。****



一致性规范：从左到右和从右到左

最后，虽然此次微软在机器翻译领域的突破让人印象深刻，****但是研究人员也提醒大家，这并不代表人类已经完全解决了机器翻译的问题****，只能说明我们离终极目标又更近了一步，未来仍有很多挑战需要解决。