机器翻译的原理是什么

机器翻译已经取得了非常好的效果，图 1 给出 Google 翻译的一个例子。

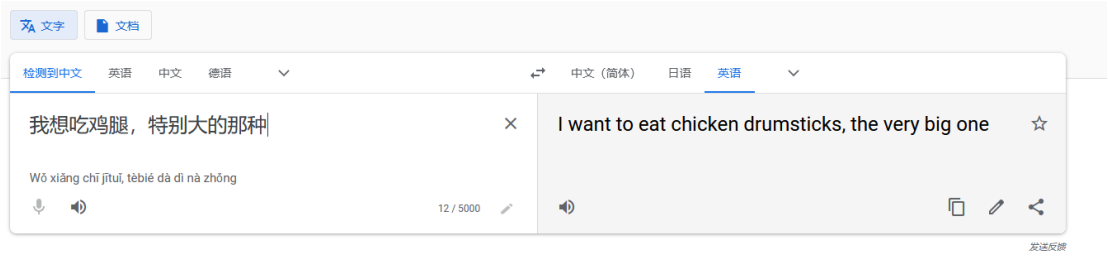

图 1：Google Translate 的翻译样例

如果不细想，好像机器做翻译没什么困难的，拿一本词典一个词一个词地翻过去不就可以了吗？这也是人工智能发展初期研究者的想法。那时恰好是冷战期间，美国和苏联为了情报工作的需要，都投入重金发展英-俄互译。然而，人们很快发现语言现象不是那么简单的，语序、多义、上下文相关性等很多问题都不是靠查字典能解决的。因此，早期努力基本以失败告终，资金支持断了，导致机器翻译研究在 60 年代后陷入低谷。

1. 统计机器翻译

突破发生在 1993 年，IBM 的 Brown 和 Della Pietra 等人提出了基于词对齐的翻译模型，标志着现代统计机器翻译（SMT）方法的诞生。这一方法首先从平行语料中发现对应的语言片段，并给每个对应片段计算一个概率，表示该片段出现的可能性，如图 2 所示。

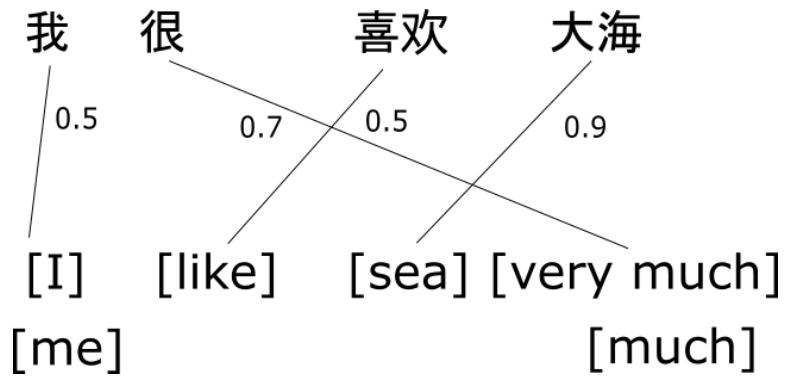

图 2：统计机器翻译中的对应片段

可以把这些对应片段理解发为一个翻译词典，只不过这个词典是从语料中自动总结出来的，并考虑了各种可能。有了这个词典，我们就知道了源语言中的每个词大约可能被翻译成目标语言中哪个词。

那么，如果选出原句子中每个词对应的概率最大的词，是否就可以得到翻译结果呢？还是不行。一是单个词概率最大并不意味着组成的句子就是合理的，二是不同语言的词序不同，按原句子的词序得到的结果很可能是不对的。因此，我们需要另一个模型，用来判断得到的句子是否通顺流畅，这一模型称为**语言模型**。比如，“I like sea very much”就比”I very much like sea”更通顺，因此在语言模型里的得到的分数也更高。有了语言模型，就可以得到合理、流畅的翻译结果了。

2. 神经机器翻译

基于统计的翻译方法取得了巨大成功，但在翻译性能上还有明显的``直译''痕迹，词语组合生硬的情况比较严重。这是因为这一方法仅以词间概率为准则做机械搜索，并没有对句子进行理解。

2014 年，Google 提出了一种序列对序列的神经网络模型[2]，用来处理机器翻译任务，从此，神经机器翻译（NMT）开始成为主角。如图 3 所示，待翻译的句子被一个递归神经网络（RNN）``压缩''成一个句子向量 S，这一过程称为编码。基于这一句子向量，另一个 RNN 通过迭代方式生成目标翻译，这一过程称为解码。可以将这里的一编码---解码过程形象理解为一个语义打包过程：在编码时将每个词的语义层层打包起来形成句子向量，在解码时再把这个包层层打开。由于打包的是语义而不是单词本身，因此可实现跨语言的语义重现。这一重现的语义用目标语言表示出来，即实现了翻译。后来，研究者对这一模型进行了很多改进，进一步提高了神经机器翻译的性能。

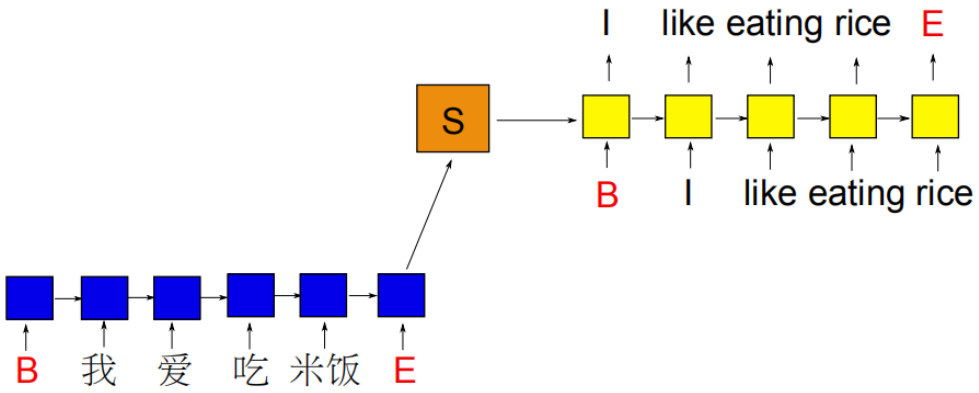

图 3：用于机器翻译的序列到序列模型[3]

2016 年 11 月，基于神经模型的翻译引擎在 Google 上线，标志着机器翻译进入神经网络时代。

[1] Brown P F, Della Pietra S A, Della Pietra V J, et al. The mathematics of statistical machine translation: Parameter estimation[J]. Computational linguistics, 1993, 19(2): 263-311.
[2] Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with

neural networks[J]. arXiv preprint arXiv:1409.3215, 2014.
[3] 王东， 人工智能，清华大学出版社，2019.10.