**机器翻译**

# 简介

机器翻译（machine translation, MT）是指使用机器（计算机）自动地将一种自然语言（源语言）语句转化为相同含义的另一种自然语言（目标语言）语句的过程。这里的自然语言是指日常使用的人类语言（如英语、汉语、日语等），它区别于人工为某些特定目的而创造的语言（如计算机编程语言）。[1]

机器翻译方法主要可以分为基于规则的机器翻译方法（rule-based machine translation, RBMT）和基于语料库的机器翻译方法（corups-based machine translation, CBMT）。后者基于建模方法不同，又可以细分为基于实例的机器翻译方法（example-based machine translation, EBMT）、统计机器翻译方法（statistical machine translation, SMT）和神经机器翻译（neural machine translation, NMT）。

随着互联网的普及和硬件技术的飞速发展，语料数据的收集规模和机器计算能力得到了提高。机器翻译系统的研究和开发从基于规则的方法发展到了基于语料库的方法，又从统计机器翻译方法发展到了神经机器翻译方法。目前，几大网络在线多语言翻译系统已经采用神经机器翻译方法，如微软翻译、谷歌翻译和百度翻译等。[1]

样例数据：

输入：There is a book on the table.

输出：桌上放着一本书。

# 算法描述（我们实现的算法）

## 机器翻译算法

### 算法一：基于注意力网络的编码器-解码器模型

传统的编码器-解码器框架包含两个主要的部分，一部分是编码器，负责将源语言句子编码成语义空间中的一个向量，期望该向量包含源语言句子的主要信息；另一部分是解码器，它基于编码器提供的语义向量，产生在语义上等价的目标语言句子。

基于循环神经网络的编码器解码器仅仅使用编码器的最后一个隐含状态来作为源语言句子的上下文向量，这种做法给编码器的信息抽取能力提出了比较大的挑战。同时，由于隐含状态向量的大小是固定的，固定长度的向量的记忆信息能力有限，特别是当句子比较长时，固定长度的向量对句子的语义编码能力可能存在问题，模型的翻译性能会随着句子变长而急剧下降。

为了克服仅仅使用最后的隐含状态作为源语言句子的上下文表示导致的问题，Bahdanau等人于2015年提出了基于注意力网络的编码器-解码器框架（Bahdanau et al., 2014）[2]。同传统的编码器-解码器框架相比，基于注意力网络的编码器-解码器框架主要有两点不同，如图1所示。

[1]. 编码器采用双向的循环神经网络。正向和反向循环神经网络生成的对应时刻的隐含状态进行拼接形成该时刻的最终隐含状态，拼接之后的隐含状态包含了以第个词为中心的整个源语言句子的信息。

[2]. 使用注意力网络将编码器生成的所有隐含状态加权。注意力机制通过比较解码器隐含状态和编码器隐含状态来计算编码器每个隐含状态的权重，然后用该权重将编码器的所有隐含状态按位加权得到该时刻的源语言句子的上下文向量。[1, 2]

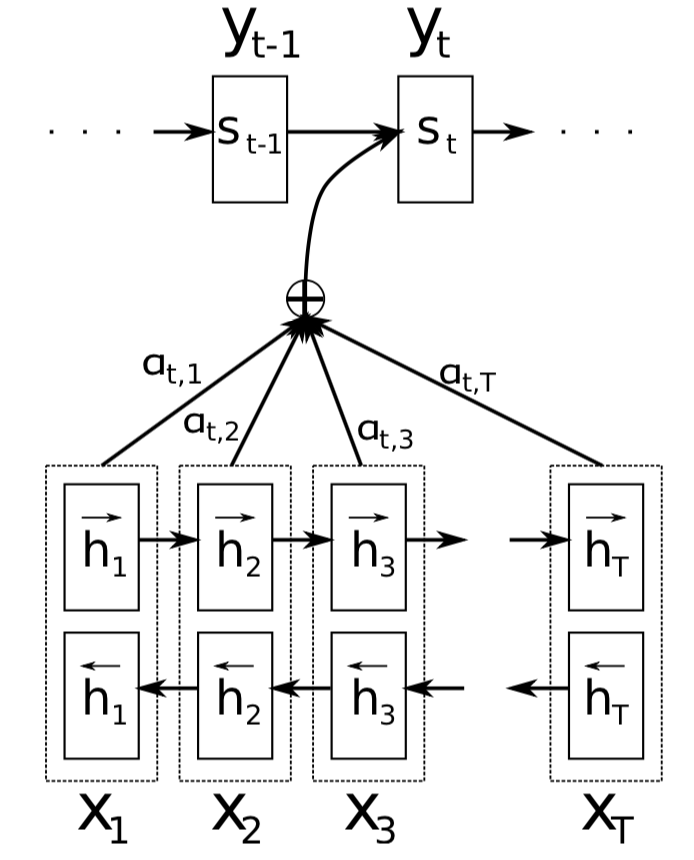


图1 基于注意力网络的编码器-解码器框架

### 算法二：完全基于注意力网络的神经翻译模型

基于循环神经网络的编码器和解码器，每个词的隐含状态都依赖于前一个词的信息，所有编码状态是顺序生成的，这严重影响了模型的并行能力。另一方面，尽管基于门的循环神经单元可以解决梯度消失或者爆炸的问题，然而相距太远的词的信息仍然不能保证被考虑进来。为了同时解决这些问题，可以将两个额外的注意力网络引入编码器和解码器的内部，分别用于解决源语言句子和目标语言句子内部词语之间的依赖关系。基于这样的考虑，Vaswani等人提出了完全基于注意力网络的神经翻译模型[3]。

如图2所示，编码器由N个同构的网络层堆叠而成，每一个网络层包含两个子网络层：第一个是多头自注意力网络，用于将同层的源语言句子里的其他词的信息考虑进来以生成当前词的上下文向量；第二个是一个全联通的前馈神经网络，将自注意力网络生成的源语言句子内的上下文向量同当前词的信息整合，从而生成考虑了整个句子上下文的当前时刻的隐含状态。为提高模型的训练速度，残差连接和层规范化被用于这两个子网络层。

解码器同样包含堆叠的N个同构网络层，每个网络层包含三个子网络层：第一个同编码器的第一个子网络层类似，负责将同层的目标语言句子里的其他词的信息考虑进来生成一个目标语言句子内的上下文向量。因为解码器在解码的时候只能看到已经生成的词的信息，对于未生成的内容，使用掩码机制将其屏蔽掉。第二个是多头注意力网络，负责将源语言句子的隐含状态同目标语言的隐含状态进行比较生产源语言句子的上下文向量。第三个子网络层同编码器的第二个子网络层类似，是一个全联通的前馈神经网络，将自注意力网络生成的目标语言句子内的上下文向量、注意力网络生成的源语言句子的上下文向量以及当前词的信息进行整合，从而更好地预测下一个目标语言词。同编码器类似，残差网络和层规范化也被用于解码器的三个子网络层。[1, 3]

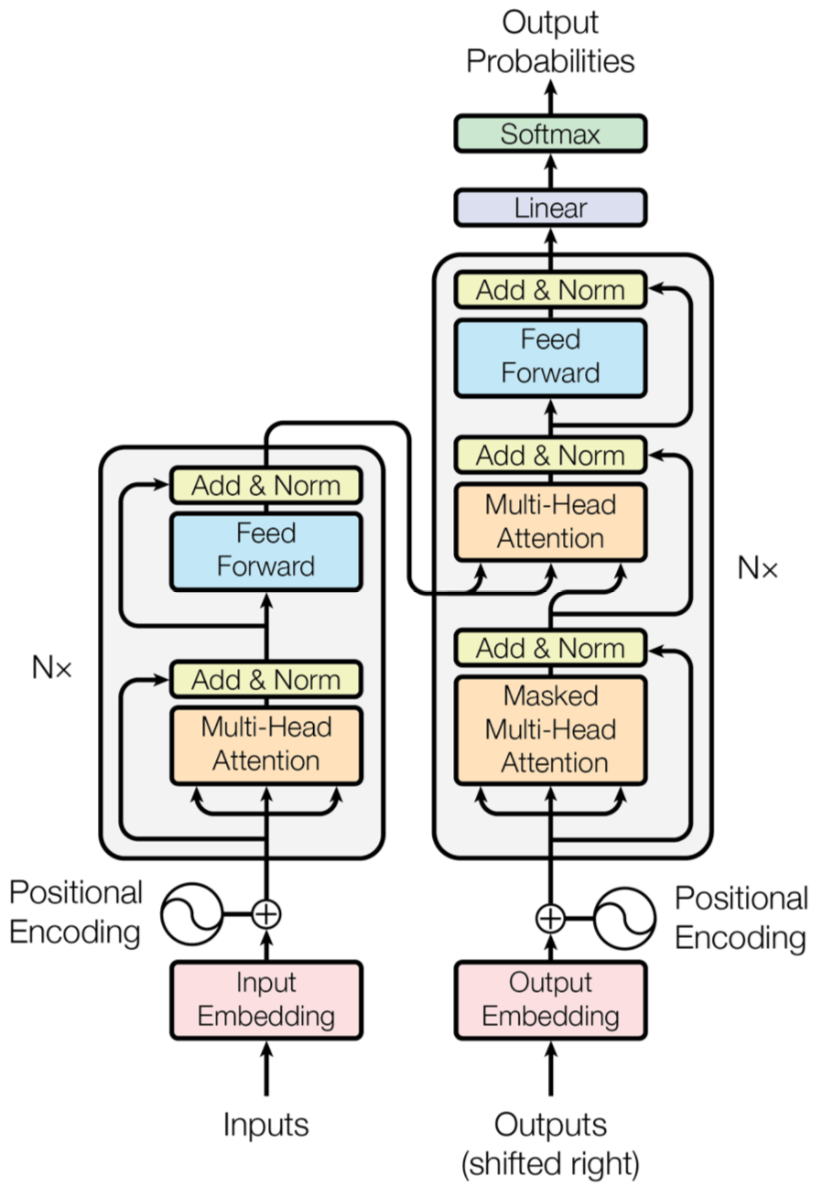


图2 完全基于注意力网络的神经翻译模型

# 公开数据集

## 数据集translation2019zh

**数据描述：**

中英文平行语料520万对。每一个对，包含一个英文和对应的中文。中文或英文，多数情况是一句带标点符号的完整的话。对于一个平行的中英文对，中文平均有36个字，英文平均有19个单词。

数据集划分：数据去重并分成三个部分。训练集：516万；验证集：3.9万；测试集，数万，不提供下载。

**结构：**

{"english": <english>, "chinese": <chinese>}

其中，english是英文句子，chinese是中文句子，中英文一一对应。

**例子：**

{"english": "In Italy, there is no real public pressure for a new, fairer tax system.", "chinese": "在意大利，公众不会真的向政府施压，要求实行新的、更公平的税收制度。"}

**下载地址：**https://storage.googleapis.com/nlp\_chinese\_corpus/translation2019zh.zip

# 性能指标

## 评价指标

BLEU是目前使用最广泛的自动评测指标。它的原理是衡量篇章机器译文和篇章参考译文之间不同长度的短语片段的匹配程度。具体计算方法是统计机器译文和参考译文之间n元文法匹配的数目占机器译文中所有n元文法总数的比例。在机器译文长度一定的情况下，匹配数目越多则代表候选的译文质量越高。[1]

## 性能（我们的算法在不同数据集上的性能）

选取translation2019zh验证集的前3000对中英语料作为验证集，选取translation2019zh验证集的后2000对中英语料作为测试集，验证集上的BLEU值为16.09，测试集上的BLEU值为15.78。

# 相关工作

统计机器翻译几十年来一直是主导的翻译范式，其实际实现通常是基于短语的系统，可以翻译不同长度的单词或短语序列。在神经机器翻译出现之前，神经网络已被用作统计机器翻译系统中一个组件并取得了一些成功。然而这种方法仍然使用基于短语的翻译系统作为核心，因此继承了它的缺点。

2013年，Nal Kalchbrenner和Phil Blunsom提出了一种用于机器翻译的新型端到端编码器-解码器结构[4]。该模型可以使用卷积神经网络（CNN）将给定的一段源文本编码成一个连续的向量，然后再使用循环神经网络（RNN）作为解码器将该状态向量转换成目标语言。他们的研究成果可以说是神经机器翻译的诞生。

由于梯度消失/爆炸问题让RNN难以处理长距离依存，因此，NMT模型一开始的表现并不友好。2014年，Cho et al.和Sutskever et al.提出了seq2seq学习的方法[5, 6]，将RNN既用于编码器也用于解码器，并且引入了LSTM，在门机制的帮助下，梯度消失/爆炸问题得到了控制，从而让模型可以更好地获取句子中的长距依存。

LSTM 的引入解决了长距离重新排序问题，同时将 NMT 的主要难题变成了固定长度向量问题：不管源句子多长，神经网络都需要将其压缩成一个固定长度的向量，这会在解码过程中带来更大的复杂性和不确定性，尤其是当源句子很长时。自 2014 年 Yoshua Bengio 的团队为 NMT 引入了注意力机制后，固定长度向量的问题也开始得到解决[2]。

2016年9月份，谷歌发布了基于神经网络的翻译系统（GNMT），致力于解决翻译速度、处理低频词、处理模型漏翻等问题[7]。2017年，Facebook 人工智能研究院（FAIR）宣布了他们使用 CNN 实现 NMT 的方法，其可以实现与基于 RNN 的 NMT 近似的表现水平，但速度却快 9 倍[8, 9]。作为回应，谷歌在 6 月份发布了完全基于注意力（attention）的 NMT 模型，这个模型既没有使用 CNN，也没有使用 RNN，而是完全基于注意力机制[3]。

# 引用

[1] 李沐，刘树杰，张冬冬，周明. 机器翻译[M].北京：高等教育出版社，2018.8.

[2] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate[J]. arXiv preprint arXiv:1409.0473, 2014.

[3] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Advances in Neural Information Processing Systems. 2017: 5998-6008.

[4] Kalchbrenner N, Blunsom P. Recurrent continuous translation models[C]//Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. 2013: 1700-1709.

[5] Cho K, Van Merriënboer B, Gulcehre C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[J]. arXiv preprint arXiv:1406.1078, 2014.

[6] Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks[C]//Advances in neural information processing systems. 2014: 3104-3112.

[7] Wu Y, Schuster M, Chen Z, et al. Google's neural machine translation system: Bridging the gap between human and machine translation[J]. arXiv preprint arXiv:1609.08144, 2016.

[8] Gehring J, Auli M, Grangier D, et al. A convolutional encoder model for neural machine translation[J]. arXiv preprint arXiv:1611.02344, 2016.

[9] Gehring J, Auli M, Grangier D, et al. Convolutional sequence to sequence learning[J]. arXiv preprint arXiv:1705.03122, 2017.