神经机器翻译模型演进的综述

张帅

（ 浙大城市学院计算机 1803 班 31801150 ）

摘要：机器翻译是指将一种自然语言转化为另一种自然语言的过程，是计算机语言学的分支之一，也是当前自然语言处理的领域的一个重要研究方向。Encoder－Decoder 模型的提出，标志着机器翻译进入了深度学习时代，神经机器翻译也成为了机器翻译研究的主流方向。经过最近几年的发展，神经机器翻译的模型逐步演进改良，本文选取了神经机器翻译提出后的标志性模型，包括 Encoder－Decoder 模型，RNNSearch 模型，Transformer 模型和 Bert 模型以及其改良方案，进行简单介绍，并阐释其发展历程。

关键词：自然语言处理　机器翻译　Attention 机制　模型演变

1 机器翻译的发展

机器翻译是指将一种自然语言转化为另一种自然语言的过程，是计算机语言学的分支之一，也是当前自然语言处理的领域的一个重要研究方向，具有重要的科学研究价值。同时，机器翻译已经广泛运用到了经济、政治、文化、生活的各个领域，在跨国交易，同声传译，文献翻译和在线翻译软件等方面具有较高的实用性。神经机器翻译的发展大致可以分为三个阶段：在 1990 年之前，大多采用基于规则的翻译法，包含了转化法（transfer-based）、中间语法（interlingual）、以及辞典法（dictionary-based）等；1990 年 Peter Brown 等人提出了基于噪声信道模型的统计机器翻译模型[1-2]，机器翻译领域开始使用基于统计的翻译方式，利用数学统计规律进行翻译；2013 年，Kalchbrenner 和 Blunsom[3]提出利用神经网络进行机器翻译，随 后 一 两 年 内，Suteskever[4]、Cho[5-6]、Bahdanau[7]等人提出了基于编码器—解码器结构的神经机器翻译模型，标志着机器翻译进入深度学习的时代。

2 神经机器翻译

神经机器翻译的优势是采用连续空间表示方法. 在翻译时省略了词语对齐、翻译规则抽取等步骤，语言的映射完全采用神经网络完成，可以训练一张从一个序列映射到另一个序列的神经网络，实现端到端的学习，当前神经机器翻译通常采用编码器－解码器（encoder-decoder）模型，其核心是通过一个编码器将输入的源语言编码成一个固定的向量,并提取源语言中信息,在神经网络中学习后，利用解码器对该向量进行解码，最终得到目标语言，其输入输出可以是一个变长的序列，因而在翻译、对话和文字概括方面能够获得非常好的表现。接下来的内容中，我将对神经机器的翻译的几种主流模型的演进和改良进行简要介绍。

3 经典的神经机器翻译模型以及其改进

3.1 基于循环神经网络（RNN）的 Encoder－Decoder 神经机器翻译模型

循环神经网络（RNN）是很长一段时间内，神经机器翻译所采用的主流网络结构，最早由 Cho[5-6]等人提出。其提出了一种新颖的神经网络架构，它可以将变长的序列表示为固定长度的向量，并将固定长度的向量解码为可变长度的序列。其编码器是一个 RNN 结构，输入一个变长的序列(x1,...,xt)，每次读入一个字符时，RNN 的隐层状态会发生改变，变化公式如下 ，其中 f 为一个非线性激活函数。

$$h\langle t\rangle = f(h\langle t-1\rangle, xt)$$

当读取到序列结束的标记时，RNN 隐层状态将会是整个输入序列的表说 c：

$$\boldsymbol{c} = q(\{h_1, \cdots, h_I\})$$

模型的解码器是另一个 RNN,解码器根据源语言的表示 c 和前驱输出序列(Y1，Y2…….)，生成目标词语 Yt。输出 yt 和解码器的隐层状态 ht 都受制于 yt-1 和输入序列的表示 c，ht 的计算方式可以定义为

$$h\langle t\rangle = f(h\langle t-1\rangle, y_{t-1}, c)$$

通过 ht 可以计算得到输出 yt 条件分布为，其中 f 和 g 都是非线性激活函数

$$\mathrm{P}(y_t \mid y_{t-1}, y_{t-2}, \ldots, t_1, \mathrm{c}) = \mathrm{g}(\mathrm{h}\langle \mathrm{t}-1\rangle, y_{t-1}, \mathrm{c})$$

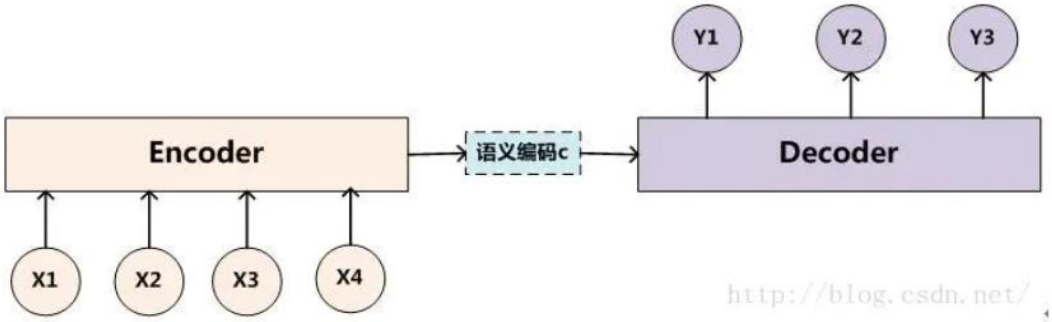

图 1 Encoder－Decoder 模型

3.2 使用长短时记忆神经网络（LSTM）改进的机器翻译模型

循环神经网络虽然在机器翻译中进行了有效利用，但是实际过程中，由于循环神经网络采用反向传播算法（BPTT）[8]进行误差的传递，会产生梯度消失的问题[9]，长短时记忆神经网络（LSTM）[10]是循环神经网络的变形结构，引入了门控的概念，通过遗忘门、输入门和输出门进行信息选择和更新，具有与循环神经网络相似的结构和优点，且性能更好。原理如下：

当信息输入时，首先由“遗忘门”决定哪些信息是要扔掉的，例如在探究一个新的对象的性别时，需要忘记上一个对象的性别。该门会读取 h_{t-1} 即 t-1 时刻的的输出结果和 x_t 即 t 时刻神经网络看到的信息，并输出一个在 0 到 1 之间的数值给每个在细胞状态，1 表示“完全保留”，0 表示“完全舍弃”。即 LSTM 以多大程度去忘记之前的信息，取决于上一次输出的结果和现在看到的东西

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right)$$

之后由“输入门”和 tanh 层决定哪些信息是要存进记忆单元。“输入门” 决定什么值将要更新。tanh 层创建一个新的候选值向量 $\tilde{G}_t$。 $\tilde{G}_t$ 为新学习到，并且要加入“记忆细胞”的候选者，i_t 可以选择性的从 $\tilde{G}_t$ 中选择“记忆细胞”中

没有的东西，计算如下：

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$$

$$\tilde{G}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C)$$

之后通过将 f_t 和旧状态相乘，丢弃掉我们确定需要丢弃的信息，在将i_t 和 $\tilde{G}_t$ 相乘，得到需要增加的信息。此时每个记忆细胞为以多大程度去忘记之前的信息 * 之前的记忆细胞 + 以多大程度去接受新的信息 * 新学习到的东西，即

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

最后，通过“输出门”确定要输出的内容，原理是运行一个 sigmoid 层来确定哪个部分将输出出去。把细胞状态通过 tanh 进行处理后，将它和 sigmoid 门的输出相乘，最终仅仅会输出确定输出的那部分数据。

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o)$$

$$h_t = o_t \times \tanh(C_t)$$

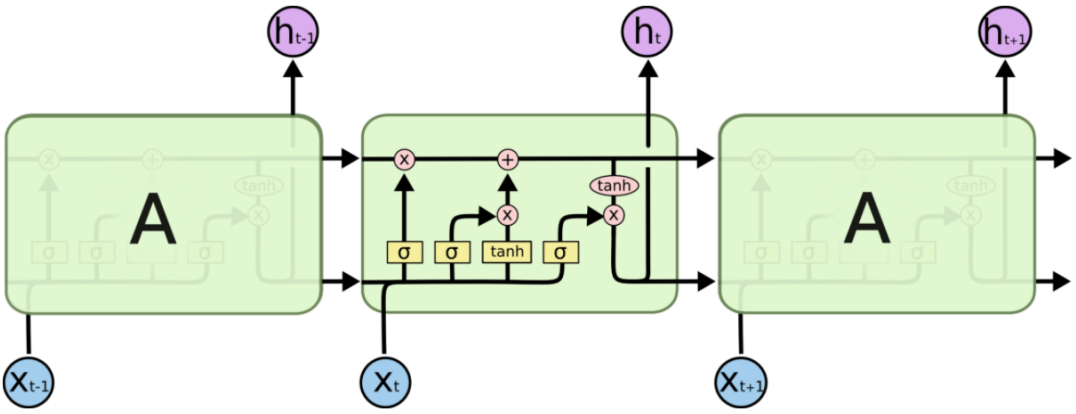

图 2 LSTM 神经网络模型

在国内，Liu 等人层进行了基于 LSTM 的蒙汉机器翻译的研究[11]。其根据蒙古语属于主语、宾语、谓语结构，而汉语属于主语、谓语、宾语结构的语言特点，在翻译过程中会发生词语调序这一现象，随着翻译句子长度的增加，长距离的语调现象不可避免，选择 LSTM 神经网络进行机器翻译一定程度上缓解了使用 RNN 神经网络后句子过长导致的梯度消失现象，该论文将正向的 LSTM 和逆向的 LSTM

进行组合构成双向 LSTM 神经网络来获得更多的语义，其构建的神经网络结构中，输入层将向量传递给正向和逆向的隐藏层节点，在 t 时刻，编码器包含 t-1 时刻和 t+1 时刻的向量信息，充分保证获取蒙古语句的上下文信息。

3.3 通过门限循环单元（GRU）进行的机器翻译模型改进

门限循环单元（GRU），由 Cho[5,12]等人提出，该结构是对长短时记忆神经网络的简化，其将 LSTM 的输入门和遗忘门合并成更新门（Update Gate），又引入了重置门（Reset Gate）， 更新门控制选择遗忘的内容和新接受的内容，用复位门控制候选状态中有多少信息是从历史信息中得到，效果与 LSTM 相近，但是降低了计算量。GRU 的计算过程如下：

首先，复位门的计算过程如下,其中 σ 为 logistic sigmoid 函数，$[\wedge]_j$代表一个向量的第 j 个元素。X 和 h_{t-1}分别为输入和前一个隐层状态。 W_r 和 U_r为已学习的权重矩阵。

$$r_j = \sigma\left([W_r x]_j + [U_r h_{<t-1>}]_j\right)$$

类似的， 更新门由以下公式计算：

$$z_j = \sigma\left([W_z x]_j + [U_z h_{<t-1>}]_j\right)$$

单元 h_j的激活由以下公式计算:

$$h_j^{<t>} = z_j h_j^{<t-1>} + \left(1 - z_j\right)\tilde{h}_j^t$$

$$\tilde{h}_j^{<t>} = \phi\left([Wx]_j + [U(r \odot h_{<t-1>})]_j\right)$$

在公式中，当复位门接近 0 时，隐层状态将会强制忽略前一个状态并且只复位当前输入。这允许隐层状态可有效地丢弃在将来会被发现不相关的信息，从而使表示更加紧凑。

在 Junyoung Chung[12] 等人的实验中，对 LSTM 和 GRU 在神经机器翻译上的表现进行了对比，结果表明，新型的门单元（GRU）的比传统的循环单元（LSTM）

表现更优，在不同的数据集上，GRU 的速度经常更快，最终的解更佳。但是不能断定 LSTM 和 GRU 何者更优，其选择严重取决于数据集和与之对应的任务。

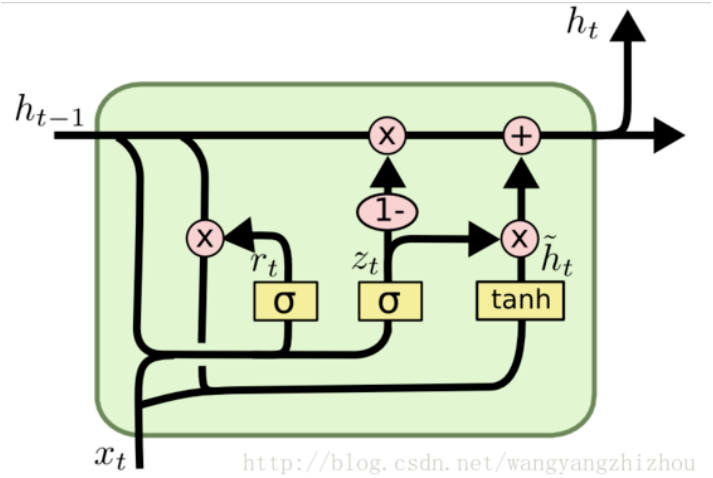

图 3 门限循环单元(GRU)

3.4 使用简单循环单元（SRU）的进行机器翻译模型的改进

GRU 单元和 LSTM 神经网络解决了 RNN 神经网络模型的梯度消失问题，但是 Zhang Wen 等[13]发现增加 GRU 单元神经网络中编码器解码器深度来优化模型性能时，翻译质量并没有上升，同时严重影响了训练速度，其推测是由于增加模型深度导致的梯度消失问题使得模型难以收敛。

因而，其引入了简单循环单元（simple recurrent unit）代替了编码器解码器中的门限循环单元。主要思想是避免了 GRU 中的状态计算和复杂的控制机制，消除门状态对前一步隐状态的依赖性，提高门计算单元的可并行性，从而加快训练速度，SRU 的计算过程如下：

$$f_t = \sigma\left(W_f x_t + b_f\right)$$

$$z_t = \sigma(W_z x_t + \boldsymbol{b}_\varepsilon)$$

$$c_t = f_t * c_{t-1} + (1 - f_t) * (W x_t)$$

$$\boldsymbol{h}_t = (1 - z_t) * \tanh(c_t) + z_t * x_t$$

其更新门和复位门的计算仅依赖输入序列，因而可以在序列长度的维度上实现并行计算。

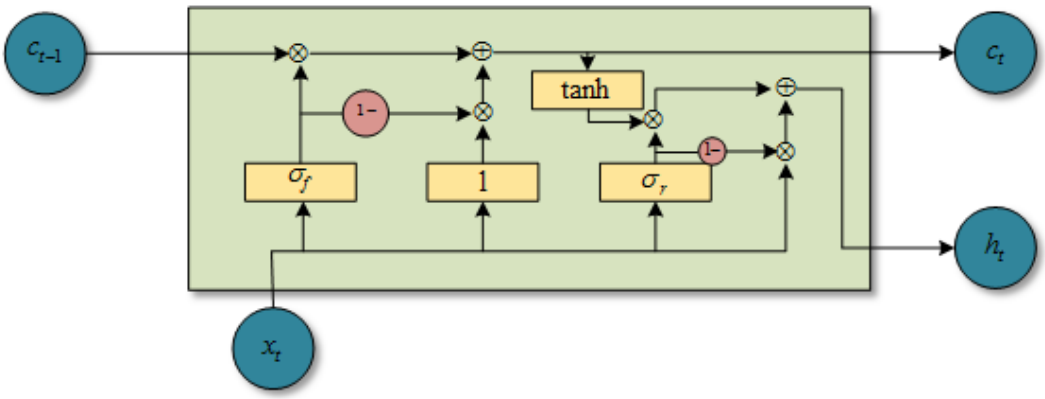

图 4 简单循环单元（SRU）

3.5 使用双向循环网络（BRNN）的进行机器翻译模型的改进

标准的 RNN 在时刻 t 的状态只能从过去的序列以及当前的输入中捕获信息，无法使用到后面序列的信息，但在机器翻译问题中，当前时刻的输出不仅和之前的状态有关，还可能和未来的状态有关系，BRNN 结合时从序列起点开始移动的 RNN 和另一个从序列末尾开始移动的 RNN 。输出单元能够计算同时依赖于过去和未来且对时刻 t 的输入值的表示。在机器翻译任务中，充分保证获取语句的上下文信息，使获得语义更完整，效果更好。

模型	特点
循环神经网络（RNN）	首次提出了并实现了神经网络机器翻译模型，缺点是采用反向传播算法进行误差的传递，会产生梯度消失的问题
长短时记忆神经网络（LSTM）	引入了门控的概念，通过遗忘门、输入门和输出门确定输出的内容，进行信息选择和更新，具有与循环神经网络相似的结构和优点，且性能更好，但计算量较大
门限循环单元（GRU）	对长短时记忆神经网络的简化，其将 LSTM 的输入门和遗忘门合并成更新门，并引入了重置门，降低计算量，
简单循环单元（SRU）	避免了 GRU 中的状态计算和复杂的控制机制，消除门状态对前一步隐状态的依赖性，提高门计算单元的可并行性，从而加快训练速度
双向循环网络（BRNN）	克服了 RNN 只能从过去的序列中获取信息的缺陷，充分获取上下文信息，是的获取语义更完整。

表 1　　经典神经机器翻译各模型对比

4 带注意力机制的神经机器翻译模型

4.1 使用注意力机制（attention）进行经典神经机器翻译模型的完善

基于循环神经网络和编码器—解码器结构的神经机器翻译模型在很长一段时间内都是神经机器翻译的主流模型。Bahdanau 等人[14]在 encoder-decoder 框架的基础上，提出了 RNNSearch 模型。 其认为传统的 encoder-decoder 模型在输入序列时，需要进行压缩，将信息固定在一个定长的向量中，这类模型在处理长句时，表现下降，为了解决这个问题，其引入了注意力机制，本质是引入了当前预测词对应输入词的上下文信息以及位置信息，在解码时从这个固定向量中抽取有用的信息来解码当前词。

RNNSearch 模型采用了双向的 BRNN，隐藏层状态包括正向与逆向输入的两部分，使得模型对输入序列由更好的表达.每个的输出词取决于当前隐状态及上一个输出词，但是文本向量 c 不再是那个固定纬度大小的向量，而是新的文本向量 c_i ，公式如下:

$$s_i = f(s_{i-1}, y_{i-1}, c_i)$$

$$p(y_i \mid y_1, y_2, \dots, y_{i-1}) = g(y_{i-1}, s_i, c_i)$$

其中新的文本向量c_i 是一个全部隐状态 h_1, h_2...h_t 的一个加权和，即

$$c_i = \sum_{j=1}^{T_x} \alpha_{ij} h_j$$

注意力权重参数也由一个神经网络训练得到:

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})}$$

$$e_{ij} = a(s_{i-1}, h_j)$$

α 是一个对齐模型，用来评估当前预测词与每一个输入词的相关度。在生成一个输出词的时候，会考虑每一个输入词与当前词的对齐关系，对齐越好的词，

应该享有更大的权重，对预测当前词会产生更大的影响。基于注意力的神经机器翻译在解码时能够动态 获取源语言相关信息，显著提升了翻译效果。

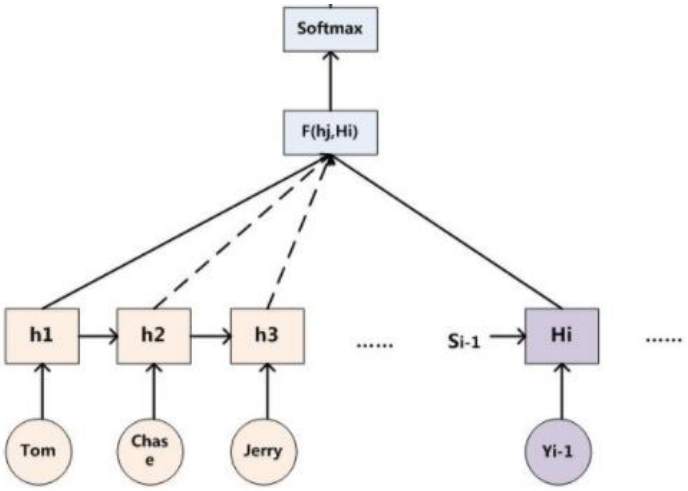

图 5 注意力的计算

4.2 通过局部注意力模型进行减少计算量的优化

RNNSearch 模型在求解注意力分配权重时，需要计算源语言句子中所有词语的权重，很耗费计算资源，Luong 等人[15]针对这个问题提出了局部注意力（LoaclAttention）模型，模型首先为当前目标单词预测一个对齐好位置 P_t ，以源位置 P_t 为中心，D 选取大小固定的窗口[P_t －D， P_t + D]的来计算上下文矢量 c_t ， c_t 通过窗口[P_t －D， P_t + D]内的源隐藏状态集合加权平均计算得到。局部注意力模型只关注一个小窗口内的信息，有效减少了计算量，适合长句子翻译，且随着句子长度增加，翻译质量并没有降低。

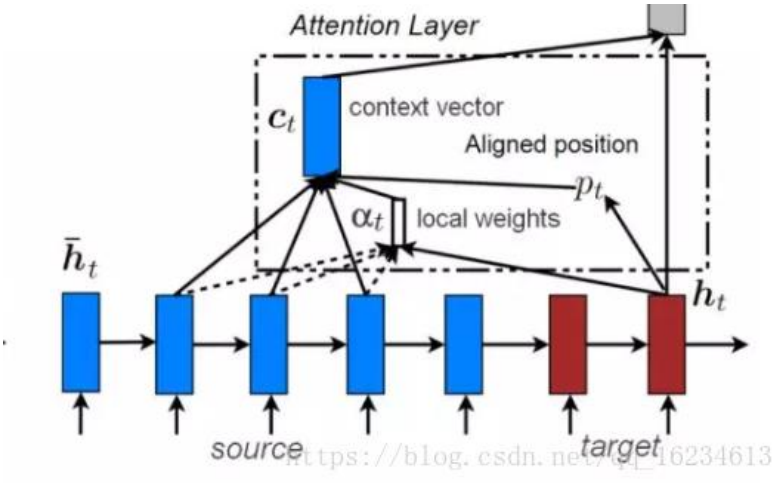

图 6 局部注意力的机制

4.3 通过监督注意力机制进行优化

RNNSearch 模型是在没有监督的情况下以无监督的方式学习的，词对齐质量较差。但是，在常规统计机器翻译（SMT）中，标准做法是在常规的对齐模型的指导下以监督的方式学习重排序模型。

受常规 SMT 中有监督重排序的启发，Liu 等人[16]提出了一种基于监督注意力的 NMT（SA-NMT）模型。似于常规 SMT，首先运行现成的对齐器（GIZA ++）[17]，以预先获得双语训练语料库的对齐。然后，将对齐结果作为注意力的监督。其采用试探法对硬对齐进行预处理：如果目标单词未与任何源单词对齐，则从最接近的对齐单词继承其从属关系，并优先选择右侧;如果一个目标词与多个源词对齐，则假定它与每个源词均等对齐。在实施翻译时，在源句和目标句中添加两个特殊标记“ eol”，假设它们彼此对齐。从而获得对注意力的最终监督，表示为α。学习过程如下：

$$-\sum_{i}\log p\left(\mathbf{y}^{i}\mid\mathbf{x}^{i};\theta\right)+\lambda\times\Delta\left(\alpha^{i},\hat{\alpha}^{i};\theta\right)$$

Δ是损失函数，处理 $\alpha^{i},\hat{\alpha}^{i}$ 之间的分歧，通过λ减轻过拟合；其次，通过在整个网络的中间层中添加监管，更容易解决消失的梯度问题。该模型提高了注意力机制的词对齐质量，但是与统计机器翻译词对齐相比仍有较大差距。

4.4 通过融合统计机器翻译词对齐信息进行优化

RNNSearch 对源语言和目标语言词语对应关系建模，无监督的模型，没有利用任何先验知识和约束机制[18]，而统计机器翻译包含了质量相对较高的词语对齐信息 。因而可以通过将统计机器翻译词对齐信息引入模型进行优化。国内外的研究大概有：

Feng 等人[19]将位变模型、繁衍模型思想引入基于注意力的神经机器翻译，提高了词对齐效果. 并且在一定程度上缓解了过度翻译问题.

Cohn 等人[20]则在注意力机制中融合了更多的结构化偏置（StructuralBiases 信息，包括位置偏置（PositionBias）、马尔可夫条件 （MarkovCondition）、繁衍模型、双语对称（Bilingual Symmetry）等信息. 实验在部分语言上产生了显

著的效果.

Zhang 等人[21]将位变模型显式地集成到注意力机制中，使得该机制同时获得源语言的词语信息和词语重排序 （Word Reordering）信息. 在较大规模的汉英语料上能够显著提高翻译质量和词对齐质量.

模型	特点
RNNSearch 模型	能够动态获取源语言相关信息，与经典的神经机器翻译相比显著提升了翻译效果，但是计算较为复杂，并且采用无监督的方式学习，词对齐质量较差。
局部注意力模型	只关注一个小窗口内的信息，有效减少了计算量，缓解了耗费计算资源的问题，适合长句子翻译
基于监督注意力的 SA-NMT 模型	通过的对齐器（GIZA ++），以预先获得双语训练语料库的对齐。然后，将对齐结果作为注意力的监督，提高了注意力机制的词对齐质量，但是与统计机器翻译词对齐相比仍有较大差距。
融合统计机器翻译词对齐	通过将统计机器翻译词对齐信息引入模型进行优化，提高了词对齐效果. 并且在一定程度上缓解了过度翻译问题

表 2　　注意力机制的神经机器翻译各模型对比

5 Transformer 模型

在前阶段机器翻译中，大部分的模型都采用了基于循环神经网络(RNN)的结构或者其改进方案，但 RNN 的明显缺点之一就是无法并行，因此速度较慢，这是其天然缺陷。而 Vaswani 等人[22]针对这个问题，提出了完全基于注意力机制的 Transformer 模型，抛弃了 RNN 结构来做机器翻译， 使用了自注意力机制来对序列进行编码，其编码器和解码器均由注意力模块和前馈神经网络构成。Transformer 模型具有高度并行化的模型结构，因此在训练速度上远超循环神经网络，且在翻译质量上也有大幅提升。Transformer 模型的结构大致如下：

编码器：

编码器由 n 个结构相同的层组成，每层主要包含两个模块：注意力模块和前馈神经网络。注意力模块以点乘注意力为基础，对输入的请求 Q、键 K 和值 V

做如下操作

$$\text{Attention}\ (Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

其中$\boldsymbol{Q} \in \mathrm{R}^{n\times d_k}, \boldsymbol{K} \in \mathrm{R}^{m\times d_k}, \boldsymbol{V} \in \mathrm{R}^{m\times d_v}$，事实上它就是三个$n\times d_k$，$m\times d_k$，$m\times d_v$的矩阵相乘，先计算 Q 与 K 的相关度，然后根据计算 Q 与 K 的相关度矩阵后，再使用这个相关度矩阵与 V 相乘，其中$\sqrt{d_k}$因子起到调节作用，使得内积不至于太大 ，最后的结果就是一个$n\times d_v$ 的矩阵，其 K 和 V 是同一个矩阵通过不同的全连接层计算得来的， Q，K 计算的目的就是为了计算出 Q 应该关注 V 中的哪些值，关注度达到多少。通过 Attention 层，模型将$n\times d_k$ 的序列 Q 编码成了一个新的$n\times d_v$ 的序列。

文中还提出了一个 multi-head 结构，为了防止模型只关注到一部分特征，却忽略了其他特征，所以 multi-head 增加模型的厚度，让模型拥有多层结构相同,其中采用权重不同的 Attention 模块,每一个 head 都关注到了不同的特征，那么模型整体就会关注到更多的特征。

在完成注意力的计算后，经过前馈神经网络进行空间变换。FFN 模块的加入引入了非线性(ReLu 激活函数)特点，从而增加了模型的表现能力，运行过程如下：

$$\text{FFN}\ (x) = max(0, xW_1 + b_1)W_2 + b_2$$

如上的模型存在一个问题，这样的模型并不能捕捉序列的顺序，如果将句子中的词序打乱，那么 Attention 的结果还是一样的，因此模型需要对词进行位置编码，以表示序列中不同词的位置关系。Transformer 模型将每个位置进行编号，然后每个编号对应一个向量，通过结合位置向量和词向量，给每个词都引入了一定的位置信息，这样 Attention 就可以分辨出不同位置的词了。

$$PE_{(\text{pos},2i)} = \sin\left(\ \text{pos}\ /10000^{2i/d_{\text{model}}}\right)$$

$$PE_{(pos,2i+1)} = \cos\left(\text{pos}/10000^{2i/d_{\text{matel}}}\right)$$

解码器

Transformer 模型的解码器与编码器结构基本相同，在解码时，因为译文在生成的时候是无法知道未来的信息的。因此，在训练时，解码器到自注意力模块里会引入一个单向的 mask 矩阵，使从前往后注意力结果被固定为 0 。另外，在解码器的自注意力模块与前馈神经网络模块之间，还有一个编码器－解码器注意力模块，这一个模块帮助解码器用编码器的输出信息来计算当前解码的输出。

Transformer 模型的关键优势至少可以归结为三点：

1.客服了基于 循环神经网络（RNN）的模型不能并行计算的缺点。

2.每个元素可以像卷积神经网络（CNN）一样和全局的信息进行交互，同时计算两个位置之间的关联所需的操作次数不随距离增长。

3.通过引入自注意力可以产生更具可解释性的模型。可以从模型中检查注意力分布。同时各个注意头可以学会执行不同的任务。

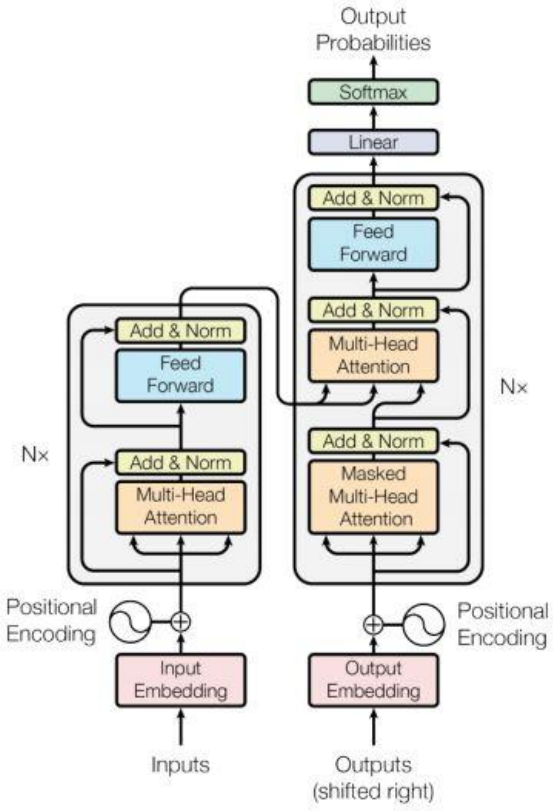

图 7 Transformer 模型

6 Bert 模型

在 Transformer 模型的基础上，J Devlin 等人提出了 BERT 模型[23]，该模型采用双向的 transformer 结构，并且主要通过预训练方法的创新，即使用 Masked LM 和 Next Sentence Prediction 两种预训练方法，建立了一个通用的 NLP 模型，BERT 模型在机器翻译等 11 个 NLP 任务上都有非常优异的表现，BERT 模型的核心是使用了有效的预训练机制，下面我将简要介绍：

BERT 的输入：

BERT 的输入编码向量是含有 3 个嵌入特征

1.WordPiece 嵌入：其将单词划分成一组有限的公共子词单元，能在单词的有效性和字符的灵活性之间取得一个折中的平衡。例如将 ‘playing’被拆分成了‘play’和‘ing’

2. 位置嵌入（Position Embedding）：位置嵌入是指将单词的位置信息编码成特征向量，概念在之前的 Transformer 模型中也有提到，位置嵌入的作用是向模型中引入单词位置关系。

3. 分割嵌入（Segment Embedding）：分割嵌入用于区分两个句子，判断上下文关系。用特征值 0 和 1 来区别前后句子。这在之后的 Next Sentence Prediction 中有所使用。

BERT 的预训练：

1.Masked LM

Masked LM 指在训练时，随机遮罩（mask）每一个句子中 15%的词，用其上下文来做预测，类似完型填空的任务，采用非监督学习的方法预测 mask 位置的词是什么，但是该方法有一个问题，因为是 mask15%的词，其数量已经很高了，这样就会导致某些词从未见过，为了解决这个问题，在训练时，确定要 Mask 掉的单词之后，80%的时候会直接替换为[Mask]，10%的时候将其替换为其它任意单词，这么做的原因是 Transformer 要保持对每个输入的分布式表征，否则模型就会记住这个[mask]的内容；剩下 10%的时候会保留原始数据。

2.Next Sentence Prediction

Next Sentence Prediction（NSP）的任务是判断句子的前后文关系，其选择一些句子对 A 与 B，其中 50%的数据 B 是 A 的下一条句子，剩余 50%的数据 B 是中随机选择的，学习其中的相关性，添加这样的预训练的目的是目前很多 NLP 的任务都需要理解两个句子之间的关系，从而能让预训练的模型更好的适应这样的任务

Bert 增大了整个模型的灵活性。如果你能拿到较好的预训练模型，就可以轻易实现一个较好的任务模型；同时，无监督的预训练模型，给了其他连续型数据问题很多想象的空间，其不仅在性能上相比 transformer 模型有所提升，而且其泛用性好，对自然语言处理的多个任务都有了历史性的突破。

模型	特点
经典的基于 RNN 的神经机器翻译模型	首次提出了使用 RNN 神经网络进行机器翻译的模型，很长一端时间内的主流模型，首先于 RNN 神经网络的缺陷，在翻译长句时，性能不佳，并且会产生梯度消失问题，通过使用 RNN 神经网络的变体（LSTM/GRU）等可以提高一定性能，但是具有 RNN 神经网络不可并行的天生缺点。
引入 Attention 机制的模型	通过引入 Attention 机制，克服了经典 RNN 模型将输入固定在一个定长的向量中，导致处理长句时，表现下降的问题，但是计算量较大，且为无监督的学习模型，通过引入局部注意力机制和进行有监督的学习可以进行改进，但是仍无法客服 RNN 神经网络的劣势
Transformer 模型	抛弃了 RNN 的结构来做机器翻译，完全基于注意力机制进行翻译任务，提出了使用自注意力机制来对序列进行编码。克服了基于 RNN 的模型不能并行计算的缺点，且模型更具可解释性。
Bert 模型	在 Transformer 模型的基础上，通过双向的 transformer 结构和预训练方法的创新，建立了一个通用的 NLP 模型，灵活性，泛用性强，是当前最具突破性的模型。

表 3　本文提到的各模型演变以及性能对比

7 总结与展望

机器翻译作为一个在经济、政治、文化、生活等各个领域都可以发挥重大作用的自然语言处理任务，具有重要的实用价值。从 2014 年 Encoder－Decoder 模型的提出起，神经机器翻译成为了机器翻译的主流研究方向，针对循环神经（RNN）网络的记忆能力有限，翻译长句时，性能不佳，并且会产生梯度消失等问题，LSTM，GRU，SRU 等 RNN 神经神经网络的变体被提出并运用于神经机器翻译，并取得了一定的成效。

Attention 机制的引入，克服了 Encoder－Decoder 模型将输入固定在一个定长的向量中，导致处理长句时，表现下降的问题，使得神经机器翻译有了巨大的突破，之后的研究中，针对引入 Attention 机制的 RNNSearch 模型，局部注意力机制和基于监督注意力的模型被提出，来改进 RNNSearch 模型运算量大，和作为无监督学习模型句子对齐较差的问题；同时，也有融合统计机器翻译（SMT），通过使用 SMT 中的处理方法将 SMT 优势融合进神经机器翻译的模型中方案提出，机器神经翻译得到了进一步发展。

Transformer 模型的提出，跳出了之前机器神经翻译大量基于 RNN 神经网络实现，带来的不可并行的天生缺点，提出了自注意力机制实现完全基于注意力机制进行翻译模型，为神经机器翻译领域带来了革命性的变革。

在 Transformer 模型的基础上，Bert 模型通过使用双向的 transformer 结构和创新性的预训练方法，建立了一个灵活性好，泛用性强的模型，并且在其他 NLP 任务上有出色的表现，成为神经机器翻译领域的集大成者，并且至今仍有不可磨灭的影响。

不可否认机器翻译技术正在飞速的进步，但是，机器翻译的质量远没有达到令人满意的水平，尤其是在追求“信、达、雅”层面，神经机器翻译还有所欠缺，并且极度依赖充足的训练语料。与此同时，在同声传译，篇章翻译，多语言翻译等领域，神经机器翻译仍在发展阶段，仍需要革命性的技术进步推动其发展。

我们可以预期，未来的机器翻译系统能够辅助人类实现高效率精准翻译，但要实现无须人工干预的高质量全自动翻译仍需要大量时间和等待其他领域技术发展的支持。

参考文献

[1] Brown P F ,Cocke J,Della Pietra S A,et al.Astatistical approach to machine tra--nslation[J]. Computational Linguistics, 1990, 16(2): 79-85.

[2] Brown, P.F., Della Pietra, S.A., Della Pietra, V.J. and Mercer, R.L. (1993) The Mathematics of Statistical Machine Translation: Parameter Estimation. Computational Linguistics, 19, 263-311.

[3] Kalchbrenner N, Blunsom P. Recurrent Continuous Translation Models. [C]. EMNLP, 2013:1700-1709

[4] Sutskever I , Vinyals O , Le Q V . Sequence to Sequence Learning with Neural Networks[C]// NIPS. MIT Press, 2014:3104-3112

[5] CHO K, van MERRIENBOER B, GULCEHRE C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation [J]. CoRR, 2014: abs/1406. 1078.

[6] Cho K , Van Merrienboer B , Bahdanau D , et al. On the Properties of Neural Machine Translation: Encoder-Decoder Approaches[J]. Computer Science, 2014.

[7] Bahdanau, D., Cho, K. and Bengio, Y. (2014) Neural Machine Translation by Jointly Learning to Align and Translate[J].arXiv preprint at arXiv:1409.1973,2014

[8] D. E. Rumelhart, G. E. Hinton and R. J. Williams. Learning repre-sentations by back-propagating errors. Nature, 1986, 323(9): 533- 536.

[9] Y. Bengio, P. Y. Simard, and P. Frasconi, Learning long-term dependencies with gradient descentis difficult, IEEE Transactions on Neural Networks, vol. 5, no. 2, pp. 157–166, 1994.

[10] Hochreiter S , Schmidhuber J . Long Short-Term Memory[J]. Neural Computat--ion, 1997, 9(8):1735-1780.

[11] 刘婉婉，苏依拉，乌尼尔,等. 基于 LSTM 的蒙汉机器翻译的研究[J]. 计算机工程与科学，2018，40(10):1890-1896.

[12]CHUNG J,GULCEHR E C,CHO K H,etal.Empirical evaluation of gated recur--ednt neuralnet works on sequence modeling[J].arXiv preprintar Xiv:1412.3555,2014.

[13] Lei T , Zhang Y , Wang S I , et al. Simple Recurrent Units for Highly Parallelizable Recurrence[C]// Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018.

[14] Bahdanau, D., Cho, K.H. and Bengio, Y. (2014) Neural Machine Translation by Jointly Learning to Align and Translate. ICLR 2015.

[15] Luong, M.T., Pham, H. and Manning, C.D. (2015) Effective Approaches to Attention-Based Neural Machine Translation.2015:1412-1421

[16] Liu L ,Utiyama M,Finch A,et al. Neural machine translation machine translation. arXiv preprint /1511.04586v1, 2015 with supervise dattention//Proceedin of COLING 2016, the 26th International Conference on Computational Linguistics, Osaka, Japan 2016:3093-3102.

［17］Zong Cheng-Qing. Statistical machine ttranslation. Second Edition. Beijing: Tsinghua University Press, 2013 (in Chinese)
［18］Cheng Y , Shen S , He Z , et al. Agreement-based Joint Training for Bidirectional Attention-based Neural Machine Translation[J]. 2015.
［19］Feng S, Liu S, Li M, et al. Implicit Distortion and Fertility Models for Attention-based Encoder-Decoder NMT Model[J]. arxiv, 2016
［20］Cohn T , Hoang C D V , Vymolova E , et al. Incorporating Structural Alignment Biases into an Attentional Neural Translation Model[C]// Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2016.
［21］Zhang J , Wang M , Liu Q , et al. Incorporating Word Reordering Knowledge into Attention-based Neural Machine Translation[C]// Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2017.
［22］Vaswani A , Shazeer N , Parma rN , etal.Attention is all you need [C] // NIPS2017: Proceedings of the 2012 International Conferenceon Neural Information Processing Systems.Long Beach, USA.2017:6000-6010
［23］Devlin J , Chang M W , Lee K , et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[J]. 2018.