计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.200200003

一种基于对数位置表示和自注意力的机器翻译新模型

纪明轩[1]　宋玉蓉[2]
1 南京邮电大学计算机学院　南京 210023
2 南京邮电大学自动化学院　南京 210023
(793271650@qq.com)

摘　要　在机器翻译任务中,自注意力机制由于高度可并行化的计算能力而减少了模型的训练时间,并且可以有效地捕捉到上下文中所有单词之间的语义相关度而受到了广泛的关注。然而,不同于循环神经网络,自注意力机制的高效源于忽略上下文单词之间的位置结构信息。为了使模型能够利用单词之间的位置信息,基于自注意力机制的机器翻译模型 Transformer 使用正余弦位置编码方式表示单词的绝对位置信息,然而,这种方法虽然能够反应出相对距离,但却缺乏方向性。文中将对数位置表示方法与自注意力机制相结合,提出一种机器翻译新模型。该模型不仅继承了自注意力机制的高效性,还可以保留单词之间的距离信息与方向性信息。研究表明,与传统的自注意力机制模型以及其它模型相比,文中所提新模型能够显著地提高机器翻译的准确性。

关键词:机器翻译;自注意力;位置信息;位置编码;对数位置表示
中图法分类号　TP181

New Machine Translation Model Based on Logarithmic Position Representation and Self-attention

JI Ming-xuan[1] and SONG Yu-rong[2]
1 College of Computer, Nanjing University of Posts and Telecommunications, Nanjing 210023, China
2 College of Automation, Nanjing University of Posts and Telecommunications, Nanjing 210023, China

Abstract　In the task of machine translation, self-attention mechanism has attracted widespread attention due to its highly parallelizable computing ability, which significantly reduces the training time of the model, and its ability to effectively capture the semantic relevance between all words in the context. However, unlike recurrent neural networks, the efficiency of self-attention mechanism stems from ignoring the position information between the words of the context. In order to make the model utilize the position information between the words, the machine translation model called Transformer, which is based on self-attention mechanism, represents the absolute position information of the words with sine function and cosine function. Although this method can reflect the relative distance, it lacks directionality. Therefore, based on the logarithmic position representation and self-attention mechanism, a new model of machine translation is proposed. This model not only inherits the efficiency of self-attention mechanism, but also retains distance and directionality between words. The results show that the new model can significantly improve the accuracy of machine translation compared with the traditional self-attention mechanism model and other models.

Keywords　Machine translation, Self-attention, Position information, Position encoding, Logarithmic position representation

1　引言

机器翻译是时下热门的自然语言处理(NLP)任务,循环神经网络(RNN)中的长短期记忆(LSTM)网络[1]和门控循环单元(GRU)网络[2]作为机器翻译中经典的建模方法,通常使用编码器-解码器体系作为模型的基本架构[3-6]。循环神经网络模型通常沿着输入序列对每个位置的元素进行顺序计算[7],它根据之前的隐藏状态h_{t-1}和当前时刻t的输入元素计算出当前的隐藏状态h_t。顺序计算的模式捕获了句子结构中的绝对位置信息和相对位置信息,但却阻碍了并行训练,限制了训练样本之间的并行处理能力。

目前,自注意力(Self-Attention,SA)机制已经成为机器翻译任务模型中的重要组成部分[8-11],这种机制忽略了输入与输出之间的距离与顺序关系,显式地捕捉到了当前单词与所有单词之间的语义关系,因此可以使用并行化的计算方式能够显著地减少训练时间,且在依赖性建模方面具有非常高的灵活性。SA 机制能够捕捉到非常全面的语义信息,却无法把握句子的结构信息[12-14]。因此,一些特定的任务会将这种注意力机制与循环神经网络或卷积神经网络结合使用以便在计算效率和获取结构信息方面达到一定的平衡。Hao[15]将 SA 机制与有序神经元 LSTM(Ordered Neurons LSTM[16])相结合,引入了面向语法的归纳偏置来生成语法树,实验表明该模型在机器翻译和逻辑推理任务中展现出优异的性能;此外,Hao[17]还对句子进行短语划分并对其进行合成,而后利用短

本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家自然科学基金(61672298,61873326,61802155);江苏高校哲学社会科学研究重点项目(2018SJZDI142)
This work was supported by the National Natural Science Foundation of China (61672298, 61873326, 61802155) and Key Research Projects of Philosophy and Social Sciences in Jiangsu Universities (2018SJZDI142).
通信作者:宋玉蓉(songyr@njupt.edu.cn)

语标签监督以及短语交互作用，采用多粒度表示方式增强 SA 机制对句子结构建模的能力，实验结果证明该方法效果良好。Yang[18] 提出利用卷积来限制 SA 机制的关注范围，增强相邻元素之间的依赖性，在对多种任务的实验结果中表明该模型在增强捕获局部信息能力方面效果显著。Fan[19] 提出一种基于句法信息和自注意力机制的方法，使用双向长短期记忆网络对隐式篇章关系建模，结果表明该方法获得了良好的效果。Wang[20] 提出“Tree Transformer”的概念使注意力头遵循树形结构，它在相邻单词之间增加“组成注意力”模块以便从源文本自动导出树结构，实验证明该模型在语言建模和学习更多可解释性的注意力得分方面有一定优势。

而在一些没有使用 RNN 结构的模型中通常会使用位置编码[19]，并将其与输入信息相结合，以这种方式将输入序列的位置信息注入到模型中。位置编码形式主要分为绝对位置编码方法、学习位置表示方法、相对位置表示方法。Wang[22] 使用依赖树表示句子的结构，对句子的潜在结构进行建模以获取句子中单词的位置信息，提出了一种结构位置表示方法，并将其与 SA 机制相结合，实验结果表明该模型在中译英任务中展现出了较好的性能，但是，这种位置编码方式无法准确地表示出长距离单词之间的位置关系。

多头注意力机制一直以来都备受关注，Paul[23] 采用启发式迭代剪枝的方法在不同层之间移除部分自注意力头，实验结果表明，绝大部分自注意力头被去除后并没有明显影响实验效果。Voita[24] 使用随机门和微分松弛的方法对注意力头进行剪枝，这种剪枝方法可以去除绝大多数自注意力头，而不会严重影响模型性能。

尽管 SA 机制在机器翻译领域得到了广泛的研究与应用，但基于该机制的模型结构建模能力仍有待提升。基于这种现状，文中将对数位置表示与 SA 机制相结合，提出一种对数位置表示（Logarithmic Position Representation，LPR）方法，应用到机器翻译模型中。

2 相关工作

2.1 自注意力

SA 机制由于其并行计算能力和建模的灵活性引起了广泛的关注，而 SA 机制中的多头注意力（Multi Head Attention，MHA）机制使模型能够从不同子空间关注到相应的信息。由于 SA 机制忽略句子中单词的位置因素，它可以显式地捕捉到当前单词和句子中所有单词之间的语义关系，而 MHA 机制则是把输入序列映射到不同的子空间中，这些子空间分别采用 SA 机制，进一步增强了机器翻译模型的性能。相较于传统的 RNN 模型，SA 机制具有参数更少、速度更快、效果更好等优点。

如图 1 所示，当使用 SA 机制处理每个单词（即输入序列中每个元素）时，比如对 x_i 进行计算时，SA 机制能够使其与序列中的所有单词进行关联，计算出它们之间的语义相似度，这样的好处在于能够帮助挖掘序列中所有单词之间的语义关系，从而更准确地对单词进行编码。

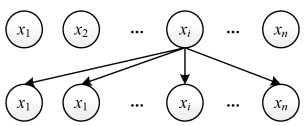


图 1 自注意力机制

Fig. 1 Self-attention mechanism

每一个注意力头对一组 n 元组的输入序列 $X=(x_1,x_2,\cdots,x_n)$，$x_i\in R^{d_x}$ 进行操作，然后通过计算得到一组 n 元组的输出序列 $Z=(z_1,z_2,\cdots,z_n)$，$z_i\in \boldsymbol{R}^{d_z}$。

输出序列 Z 中的元素 z_i 是由输入元素 x_i，x_j 经过线性变换并计算其加权和得到：

$$z_i=\sum_{j=1}^{n}\mathrm{softmax}\left(\frac{Q_iK_j^T}{\sqrt{d_k}}\right)V_j \tag{1}$$

softmax 函数中，对输入元素进行线性变换增强了表达能力。softmax 函数的计算方式如下：

$$\mathrm{softmax}(a_{ij})=\frac{\exp a_{ij}}{\sum_{k=1}^{n}\exp a_{ik}} \tag{2}$$

Q,K,V 分别表示 query，key，value，它们是对计算注意力分数有用的抽象表示，d_k 是 key 的维度，除以 $\sqrt{d_k}$ 是缩放点积，这样可以使渐变更加稳定，Q,K,V 计算方式分别如下：

$$Q_i=x_iW^Q \tag{3}$$

$$K_j=x_jW^K \tag{4}$$

$$V_j=x_jW^V \tag{5}$$

如图 2 所示，$W^Q,W^K,W^V\in \boldsymbol{R}^{d_x\times d_z}$ 是在训练过程中学习得到的矩阵，它们分别是 Q,K,V 的权重矩阵。每一个注意力头有属于自己特有的权重矩阵。

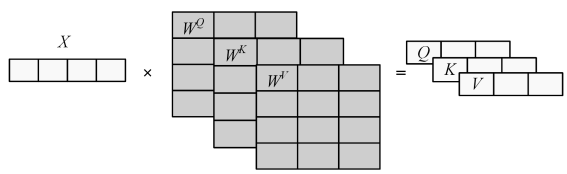


图 2 Q,K,V 计算过程

Fig. 2 Calculation of Q,K,V

SA 机制使用 l 个注意力头，所有注意力头的输出 z_h 都被合并，然后进行线性变换得到每个子层的输出。多头注意力机制扩展了模型专注于不同位置的能力。多头注意力机制的输出结果计算公式如下：

$$z^O=Concat(z_{head_1},\cdots,z_{head_l})W^O \tag{6}$$

z_{head_i} 表示第 i 个注意力头的输出向量，$Concat(\cdot)$ 的功能是将所有注意力头的输出向量合并，W^O 是模型训练过程中产生的权重矩阵。如图 3 所示，多头注意力机制合并各个注意力头的输出，然后进行线性变换得出最终输出。

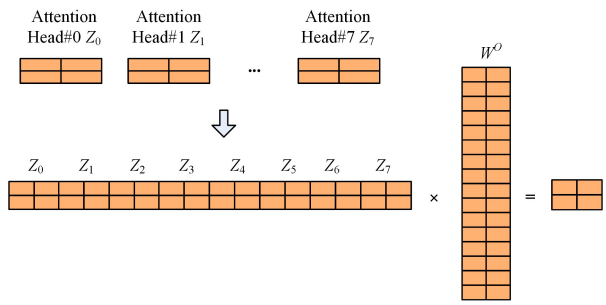


图 3 多头注意力机制

Fig. 3 Multi-head self-attention

2.2 Transformer

如图 4 所示，Transformer 采用编码器-解码器结构。编码环节通常由 6 个相同的编码器堆叠构成，每个编码器有二个子层：第一个是 MHA 机制层，第二个是全连接前馈神经（Feed Forward Neural，FFN）网络层。每个子层周围采用残差连接并进行归一化，即图中“Add&Norm”层作用，编码器

各子层输出为 $Norm(x+sublayer(x))$，这里 Norm 表示归一化，$sublayer(\cdot)$表示子层函数功能。解码环节同样由 6 个相同的解码器堆叠构成，每个解码器有 3 个子层：MHA 机制层、编码器-解码器注意力层以及全连接 FFN 网络层，每个子层周围采用残差连接并进行归一化，因此，解码器各子层输出仍为 $Norm(x+subLayer(x))$。

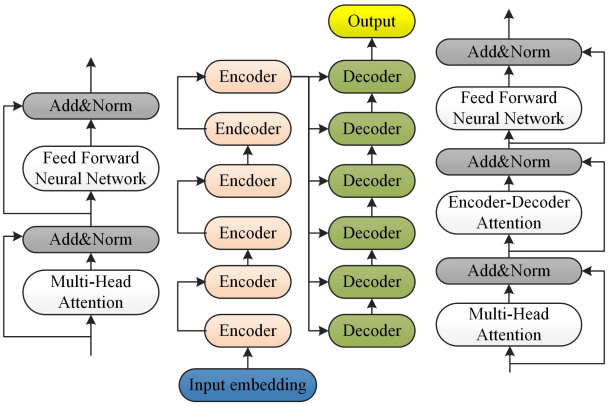


图 4 Transformer 模型结构

Fig. 4 Structure of Transformer model

2.3 正余弦位置编码

显而易见，Transformer 模型中没有递归层与卷积层，因此，为了使模型能够利用输入序列中的位置信息，Vaswani[25]提出将正余弦位置编码方式与 Transformer 中 SA 机制相结合进行应用，这种位置编码方式使用 sin 函数与 cos 函数进行位置编码，它的优点在于可以使模型的序列长度得到扩展，其本质上是一种绝对位置信息编码方式。正余弦位置编码计算方式如下所示：

$$PE_{(pos,2i)}=\sin\left(\frac{pos}{10000^{2i/d}}\right) \tag{7}$$

$$PE_{(pos,2i+1)}=\cos\left(\frac{pos}{10000^{2i+1/d}}\right) \tag{8}$$

上述公式中 pos 表示输入的位置，而 i 表示维度，即位置编码的每个维度都有相对应的正余弦函数，其中，式(7)代表偶数维度的正弦位置编码表示，式(8)代表奇数维度的余弦位置编码表示。

这种方式获得的位置编码表示虽然能够反应出单词之间的相对距离，但却缺乏方向性，并且这种位置信息会被 Transformer 中的注意力机制破坏。因此，文中提出一种新的位置表示方法——对数位置表示，并将其与 SA 机制相结合，使模型既能够有效利用 SA 机制并行化计算的优点，又能够准确地捕捉到单词之间的距离和方向信息。此外，不同于传统的相对位置表示方法在表示远距离单词间关系时进行“一刀切”的粗暴方式，由于对数函数收敛较慢，以渐进方式模糊化远距离的概念，对数位置表示可以更精确地捕捉到长距离单词之间的位置关系差异。

3 基于对数位置表示和自注意力的机器翻译新模型

文中提出一种基于对数位置表示和自注意力机制的机器翻译新模型，如图 5 所示，该模型共有 6 个编码器和 6 个解码器以及 1 个输出层，在编码器中有自注意力结合对数位置表示层、全连接 FFN 网络层；解码器中有自注意力结合对数位置表示层、编码器-解码器注意力层、全连接 FFN 网络层。输出层中包含了线性变换(linear)层和 softmax 全连接层。

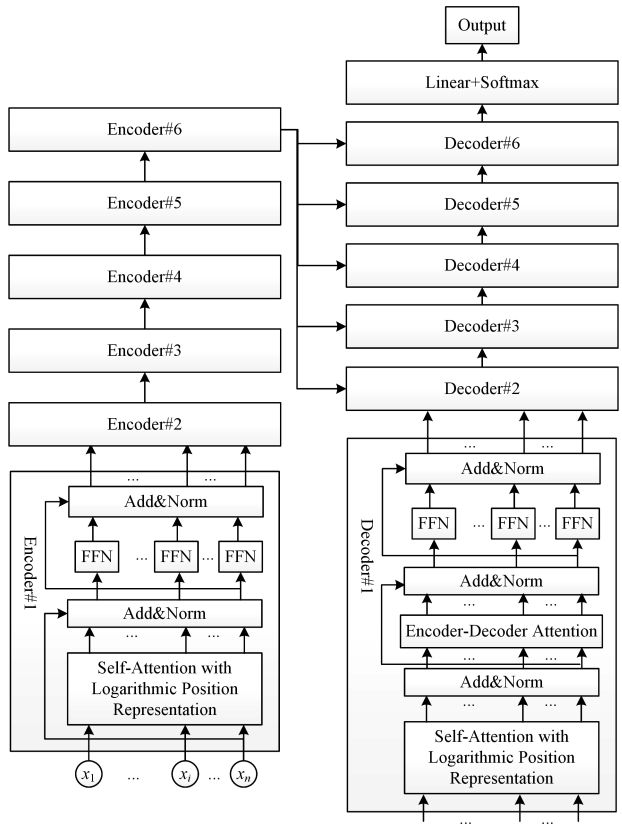


图 5 基于对数位置表示和自注意力的机器翻译模型

Fig. 5 Machine translation model based on logarithmic position representation and self-attention

该模型工作流程如下：

1)编码器：经过预处理的输入元素 $x_i\in X=(x_1,\cdots,x_n)$ 从底层编码器加载到模型中，经过编码器自注意力结合对数位置表示层作用获得新的文本表示，传入全连接 FFN 网络层。每层编码器的输出自下而上地传递到下一层编码器中。编码器每个子层周围都采用残差连接并进行归一化。

2)解码器：将最后一层编码器的输出传入到 6 个解码器中的编码器-解码器注意力层，底层解码器以预翻译结果作为输入，依次经过自注意力结合对数位置表示层、编码器-解码器注意力层、全连接 FFN 层。每层解码器的输出自下而上地传递到下一层解码器中。解码器每个子层周围都采用残差连接并进行归一化。

3)输出层：将解码器计算得到的文本表示传入线性变换层和 softmax 全连接层，通过计算获得输出结果。

3.1 自注意力结合对数位置表示

SA 机制没有考虑 RNN 和 CNN，所以会忽略文本中的序列信息，为了充分利用文本序列信息，文中提出了一种方法——将输入元素 $x_i\in X=(x_1,\cdots,x_n)$ 的位置信息提取出来并进行计算，文中提出的对数位置表示本质上是表示输入元素 x_i 与 x_j 之间的相对位置关系。将这些输入元素构建成以 $x_i(i=1,2,\cdots,n)$ 为节点，$e_{ij}\in\{(x_i,x_j)\mid x_i,x_j\in X\}$ 为边的有向完全图，e_{ij} 则包含了 x_i 与 x_j 之间的对数位置关系。

文中用向量 LP_{ij}^K，LP_{ij}^V 表示输入元素 x_i 与 x_j 之间的对数位置关系。将对数位置关系加入到模型中，对式(1)做出修改，得到如下公式：

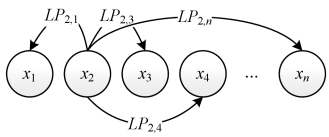
$$z_i=\sum_{j=1}^{n}\mathrm{softmax}\left(\frac{Q_i(K_j+LP_{ij}^K)^{\mathrm{T}}}{\sqrt{d_k}}\right)(V_j+LP_{ij}^V) \tag{9}$$

其中，LP_{ij}^{K}和LP_{ij}^{V}是在训练过程中通过学习得到的对数位置表示，与Q,K,V的维度是相同的，下一小节中会详细讲解LP_{ij}^{K}和LP_{ij}^{V}的推导过程。

位置信息的注入能够很大程度上改善SA机制中的编码器忽略输入序列层次结构的情况。在机器翻译、自然语言推理、智能问答系统等特定的任务中，位置信息发挥着极其重要的作用。接下来将详细介绍对数位置表示的概念。

3.2 对数位置表示

上一节中提到，以输入$x_i \in X=(x_1,\cdots,x_n)$为节点，$e_{ij} \in \{(x_i,x_j)\mid x_i,x_j \in X\}$为边构建完全有向图，$e_{ij}$保存了位置信息$LP_{ij}^{K}$，$LP_{ij}^{V}$，它们构建的关系网如图6所示。



$$LP_{2,1}^{K}=P_{-1}^{K} \quad LP_{2,4}^{K}=P_{2}^{K}$$
$$LP_{2,1}^{V}=P_{-1}^{V} \quad LP_{2,4}^{V}=P_{2}^{V}$$
$$LP_{2,3}^{K}=P_{1}^{K} \quad LP_{2,n}^{K}=P_{\log_k^{\lfloor n-2\rfloor}+1}^{K}$$
$$LP_{2,3}^{V}=P_{1}^{V} \quad LP_{2,n}^{V}=P_{\log_k^{\lfloor n-2\rfloor}+1}^{K}$$

图6 单词位置关系网络

Fig. 6 Network of positional relationships between words

图6中，假设log底数$k=2$，文本长度$n>4$。e_{ij}中保存的信息由LP_{ij}^{K}，LP_{ij}^{V}表示，LP_{ij}^{K}，LP_{ij}^{V}的对数位置关系表示计算方法如下。

$$LP_{ij}^{K}=P_{-\log_k^{\lfloor|j-i|\rfloor}-1}^{K},j<i \tag{10}$$
$$LP_{ij}^{K}=P_{\log_k^{\lfloor|j-i|\rfloor}+1}^{K},j>i \tag{11}$$
$$LP_{ij}^{V}=P_{-\log_k^{\lfloor|j-i|\rfloor}-1}^{V},j<i \tag{12}$$
$$LP_{ij}^{V}=P_{-\log_k^{\lfloor|j-i|\rfloor}-1}^{V},j>i \tag{13}$$

P_s^k，P_s^V分别是在训练过程中训练产生的对数位置表示。根据单词间相对位置的差异，对数位置表示序列$P^K=(P_{-\log_k^{\lfloor|1-n|\rfloor}-1}^{K},\cdots,P_{\log_k^{n-1}+1}^{K})(n>1)$，$P^V=(P_{-\log_k^{\lfloor|1-n|\rfloor}-1}^{V},\cdots,P_{\log_k^{n-1}+1}^{V})(n>1)$。上述表示，$P_s(s\in[-\log_k^{\lfloor|1-n|\rfloor}-1,\log_k^{n-1}+1]$，$s$为整数且$n>1$)是对数位置关系的编码表示，其本质上是顺序结构中单词之间的相对位置关系。而索引s则表示输入元素x_i与x_j之间的距离，当$s<0$，代表x_j在x_i的左边；当$s>0$，代表x_j在x_i的右边。对数位置关系编码参数的数量与log底数k以及文本长度n有关。

传统的相对位置表示方法最多只能表示窗口大小内的相对位置信息，而当x_i，x_j之间的距离超越了窗口的尺寸时，要想表示它们之间的位置关系通常会采用窗口边界的相对位置表示参数。所以，文中提出使用对数计算位置下标，无形中消除了窗口的概念，以渐进的方式模糊这种远距离的位置关系表示，这样可以更好地捕获长距离单词之间的位置结构关系。

对于有l个注意力头的SA模型而言，当输入长度为n时，存储对数位置表示参数的空间复杂度为$O(\ln^2 d_{lp})$，而文中在各注意力头中共享对数位置表示参数，如此可使空间复杂度便降低至$O(n^2 d_{lp})$。

4 实验

4.1 实验准备

文中使用tensorflow/tensor2tensor库来验证本文提出的模型，选用2014年机器翻译研讨会(Workshop on Machine Translation-2014，WMT-2014)中的Englisth-to-German(EN-DE)与English-to-French(EN-FR)机器翻译任务数据集。其中，EN-DE数据集中包含4.5×10^6组句对，而EN-FR数据集则包含36×10^6组句对。文中使用双语评估替换(Bilingual Evaluation Understudy，BLEU)作为翻译性能评价指标。

为验证文中提出的模型，实验中使用8个注意力头，每个注意力头都共享相同的对数位置表示参数。部分参数取值如表1所列。

表1 新模型中部分参数

Table 1 Some parameters settings in the new model

参数	值
输入层维度d_x	256
K维度d_x	64
V维度d_k	64
FFN层维度d_{ffn}	128
Dropout	0.1
激活函数	Sigmoid
Train steps	1 000
优化器	Adam

4.2 参数调整

在基于SA机制的模型中，解码器-编码器大多是层层堆叠，高层编码器往往学习到的是语义信息，底层编码器学习到的是表面纹理或者文字本身的信息。因此，本文讨论了SA机制的编码器层数对于机器翻译性能的影响。

如图7、图8所示，当$Layers\leqslant6$时，EN-DE和EN-FR任务的性能随着SA模型中Layers的增加而得到显著的提高；当$Layers>6$时，两种任务的性能曲线都趋于平缓。由数据曲线可以得出结论，文中所提模型的性能随编码器层数的增加而得到提升，这是因为编码器层数越多，模型所提取的语义信息和结构信息也就更加丰富。然而，无限制地增加编码器层数会降低模型的效率，因此，经过综合分析，文中提出的模型使用6个编码器层是最合适的。

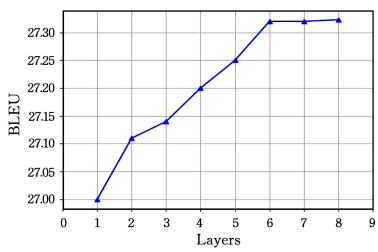


图7 EN-DE上的实验结果

Fig. 7 Experimental results on EN-DE

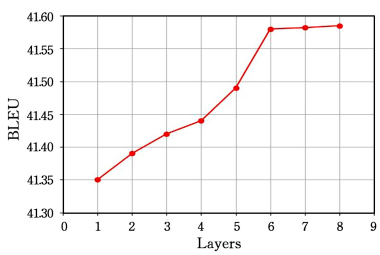


图8 EN-FR上的实验结果

Fig. 8 Experimental results on EN-FR

为了验证对数位置表示中的底数k对模型性能的影响，文中也做了许多实验进行调参。实验中分别针对EN-DE和

EN-FR 数据集对对数位置表示中的底数 k 进行调整。

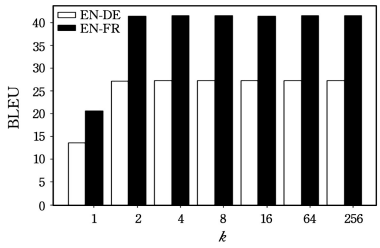


图 9 底数 k 值对实验结果影响

Fig. 9 Influence of the base k on experimental results

如图 9 所示，当 $k=1$ 时，EN-DE 和 EN-FR 的 BLEU 分数分别只有 13.61 与 20.59，这是因为当底数 $k=1$ 时，所有相对位置取值都为 0，这并不能区分出位置的差异。而当 $k\geqslant 2$ 时，继续提高 k 值，BLEU 并没有取得显著的变化，推测可知，当 $k\geqslant 2$ 时，无论 k 取何值，对数位置表示仍能体现单词间位置信息的差异。但是由于文中模型使用多个自注意力层，精准的对数位置表示信息能够帮助提升模型的翻译性能，因此，实验中一般情况下取 $k=4$。

本文还讨论了式(6)中 LP_{ij}^{K}，LP_{ij}^{V} 的影响，分别对是否使用这两个位置表达进行了实验，实验结果表 2 所列。

表 2 LP_{ij}^{K}，LP_{ij}^{V} 因子实验

Table 2 Experiment of LP_{ij}^{K}，LP_{ij}^{V}

LP_{ij}^{K}	LP_{ij}^{V}	EN-DE-BLEU	EN-FR-BLEU
0	0	13.59	20.66
1	0	27.12	41.57
0	1	27.21	41.57
1	1	**27.32**	**41.58**

如表 2 所列，如果 LP_{ij}^{K}，LP_{ij}^{V} 都被剔除，EN-DE-BLEU 和 EN-FR-BLEU 分别只有 13.59 和 20.66。而将 LP_{ij}^{K}，LP_{ij}^{V} 其中一种加入时，实验分数得到显著提升。两个因子都加入后，实验分数达到最高的 27.32 和 41.58，这说明文中提出的对数位置表示信息帮助模型捕获了文本的结构信息，有助于提升机器翻译模型的性能。

4.3 实验对比结果

文中将提出的新模型与 Transformer 以及其它模型进行比较，结果如表 3 所列。

表 3 各模型实验结果对比

Table 3 Experimental results comparison of different models

Model	EN-DE-BLEU	EN-FR-BLEU
GNMT+RL[3]	24.61	39.92
MoE[6]	26.03	40.56
ConvS2S[9]	25.16	40.46
Transformer[25]	26.51	40.23
SA+LPR	**27.32**	**41.58**

如表 3 所列，在两种任务下，文中提出的新模型性能都优于其它模型。对于 EN-DE，新模型相较于 Transformer 提高了 0.81BLEU。对于 EN-FR，本文方法相较于 Transformer 提高了 1.35BLEU。文中所提新模型能够取得如此大的性能提升主要得益于对文本结构信息的精准捕获与语义信息深层分析的有效结合。表中结果显示 EN-DE 的 BLEU 分数明显低于 EN-FR 的 BLEU 分数，这是由于 EN-FR 的数据集规模要远远大于 EN-DE 数据集，具有更加丰富的语料信息。

结束语 文中将对数位置表示与 SA 机制相结合，提高了机器翻译的性能。实验结果表明，文中提出的模型可以更精确地表示长距离单词之间的位置关系，在长句子较多的数据集任务中可以取得较好的分数，这是由于 log 函数收敛较慢，以渐进方式模糊化了远距离的概念，可以更精确地捕捉到长距离单词之间的位置关系差异。然而，文中提出的模型在短句子较多的数据集中的实验效果不尽如人意，这是由于使用对数取单词间相对位置下标时，对于表示短距离单词之间的位置关系精度相对于实际情况有偏差。

下一步将继续研究如何有效地结合句法分析与 SA 机制，以增强模型对句法结构进行建模的能力。将卷积网络与 SA 机制进行有效的结合以增强模型获取局部信息的能力也是值得继续挖掘的技术。

参 考 文 献

[1] SAK H, SENIOR A, BEAUFAVS F. Long short-term memory recurrent neural network architectures for large scale acoustic modeling[C]// Fifteenth Annual Conference of the International Speech Communication Association. 2014.

[2] CHUNG J, GULCEHRE C, CHO K H, et al. Empirical evaluation of gated recurrent neural networks on sequence modeling [J]. arXiv: 1412.3555, 2014.

[3] WU Y, SCHUSTER M, CHEN Z, et al. Google's neural machine translation system: Bridging the gap between human and machine translation[J]. arXiv: 1609.08144, 2016.

[4] JOZEFOWICZ R, VINYALS O, MIKE S M, et al. Exploring the limits of language modeling[J]. arXiv: 1602.02410, 2016.

[5] NIE Y, HAN Y, HUANG J, et al. Attention-based encoder-decoder model for answer selection in question answering [J]. Frontiers of Information Technology & Electronic Engineering, 2017, 18(4): 535-544.

[6] SHAZEER N, MIRHOSEINI A, MAZIARZ K, et al. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer[J]. arXiv: 1701.06538, 2017.

[7] CAO J, LI R. Fixed-time synchronization of delayed memristor-based recurrent neural networks[J]. Science China Information Sciences, 2017, 60(3): 108-122.

[8] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate[J]. arXiv: 1409.0473, 2014.

[9] GEHRING J, AULI M, GRANGIER D, et al. Convolutional sequence to sequence learning[C]// Proceedings of the 34th International Conference on Machine Learning-Volume 70. JMLR. org, 2017: 1243-1252.

[10] CHO K, VAN M B, GULCEHRE C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[J]. arXiv: 1406.1078, 2014.

[11] BAEVSKI A, EDUNOV S, LIU Y, et al. Cloze-driven pretraining of self-attention networks[J]. arXiv: 1903.07785, 2019.

[12] SHEN T, ZHOU T, LONG G, et al. Disan: directional self-attention network for rnn/cnn-free language understanding[C]// Thirty-Second AAAI Conference on Artificial Intelligence. 2018.

[13] XU D, RUAN C, KORPEOGLU E, et al. Self-attention with functional time representation learning[C]// Advances in Neural Information Processing Systems. 2019: 15889-15899.

[14] LIANG X B,REN F L,LIU Y K,et al. N-reader:machine reading comprehension model based on double layers of self-attention[J]. Journal of Chinese Information Processing, 2018, 32(10):134-141.
[15] HAO J,WANG X,SHI S,et al. Towards better modeling hierarchical structure for self-attention with ordered neurons[J]. arXiv:1909.01562,2019
[16] SHENY,TAN S,SORDONI A,et al. Ordered neurons:integrating tree structures into recurrent neural networks[J]. arXiv:1810.09536,2018.
[17] HAO J,WANG X,SHI S,et al. Multi-granularity self-attention for neural machine translation[J]. arXiv:1909.02222,2019.
[18] YANG B,WANG L,WONG D F,et al. Convolutional self-attention networks[J]. arXiv:1904.03107,2019.
[19] FAN Z W,ZHANG M,LI Z H. BiLSTM-based implicit discourse relation classification combining self-attention mechanism and syntactic information [J]. Computer Science, 2019, 46(5):221-227.
[20] WANG Y S,LEE H Y,CHEN Y N. Tree Transformer:Integrating Tree Structures into Self-Attention[J]. arXiv:1909.06639, 2019.
[21] ZHAO H,ZHANG Y,LIU S,et al. PSAnet:point-wise spatial attention network for scene parsing[C]// Proceedings of the European Conference on Computer Vision (ECCV). 2018:267-283.
[22] WANG X,TU Z,WANG L,et al. Self-attention with structural position representation[J]. arXiv:1909.00383,2019.
[23] MICHEL P,LEVY O,NEUBIG G. Are sixteen heads really better than one? [C]// Advances in Neural Information Processing Systems. 2019:14014-14024.
[24] VOITA E,TALBOT D,MOISEEV F,et al. Analyzing multi-head self-attention:specialized heads do the heavy lifting, the rest can be pruned[J]. arXiv:1905.09418,2019.
[25] VASWANI A,SHAZEER N,PARMAR N,et al. Attention is all you need[C]// Advances in Neural Information Processing Systems. 2017:5998-6008.

JI Ming-xuan, born in 1995, postgraduate. His main research interests include machine translation and emotion analysis.

SONG Yu-rong, born in 1971, Ph.D, professor, is a member of China Computer Federation. Her main research interests include network information dissemination and its control.

(上接第 82 页)

[9] RADOVANO M,IVANOVI M. Interactions between document representation and feature selection in text categorization[C]// International Conference on Database and Expert Systems Applications. Springer,Berlin,Heidelberg,2006:489-498.
[10] JHA V,SAVITHA R,SHENOY P D,et al. A novel sentiment aware dictionary for multi-domain sentiment classification[J]. Computers & Electrical Engineering,2018,69:585-597.
[11] PANG B,LEE L,VAITHYANATHAN S. Sentiment classification using machine learning techniques[C]// Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing(EMNLP). 2002:79-86.
[12] PALTOGLOU G,THEWALL M. A study of information retrieval weighting schemes for sentiment analysis[C]// Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2010:1386-1395.
[13] TRIPATHY A,AGRAWAL A,RATH S K. Classification of sentiment reviews using n-gram machine learning approach[J]. Expert Systems with Applications,2016,57:117-126.
[14] MIKOLOV T,SUTSKEVER I,CHEN K,et al. Distributed representations of words and phrases and their compositionality [C]// Advances in neural information processing systems. 2013: 3111-3119.
[15] WEI G S,WU K C. Sentiment analysis based on word vector model[J]. Computer Systems & Applications, 2017(3):184-188.
[16] WANG M Y,WU H,JIA X T. Research on Multi-Emotion Classification of Weibo Based on Word2vec and Extended Emotion Dictionary[J]. Journal of Northeast Normal University (Natural Science Edition),2019,51(1):55-62.
[17] TANG X B,WANG H Y. Research on Weibo Product Reviews Mining Model[J]. Journal of Intelligence,2013,32(2):107-111, 127.
[18] TAN S B. Hotel review corpus [EB/OL]. [2020-03-17]. https://www.aitechclub.com/data-detail?data_id=29.

JING Li, born in 1971, Ph.D, professor, is a member of China Computer Federation. Her main research interests include artificial intelligence and information security.

LI Man-man, born in 1992, postgraduate. Her main research interests include data analysis, data mining and natural language processing.