Journal of Computer Applications
计算机应用, 2022, 42(12): 3679 - 3685
ISSN 1001-9081
CODEN JYIIDU
2022-12-10
http://www.joca.cn

文章编号:1001-9081(2022)12-3679-07
DOI:10.11772/j.issn.1001-9081.2021101805

融合双向依存自注意力机制的神经机器翻译

李治瑾[1,2],赖　华[1,2*],文永华[1,2],高盛祥[1,2]
(1. 昆明理工大学 信息工程与自动化学院,昆明 650504; 2. 云南省人工智能重点实验室(昆明理工大学),昆明 650504)
(*通信作者电子邮箱405904235@qq.com)

摘　要:针对神经机器翻译中资源稀缺的问题,提出了一种基于双向依存自注意力机制(Bi-Dependency)的依存句法知识融合方法。首先,利用外部解析器对源句子解析得到依存解析数据;然后,将依存解析数据转化为父词位置向量和子词权重矩阵;最后,将依存知识融合到Transformer编码器的多头注意力机制上。利用Bi-Dependency,翻译模型可以同时对父词到子词、子词到父词两个方向的依存信息进行关注。双向翻译的实验结果表明,与Transformer模型相比,在富资源情况下,所提方法在汉-泰翻译上的BLEU值分别提升了1.07和0.86,在汉-英翻译上的BLEU值分别提升了0.79和0.68;在低资源情况下,所提方法在汉-泰翻译上的BLEU值分别提升了0.51和1.06,在汉-英翻译上的BLEU值分别提升了1.04和0.40。可见Bi-Dependency为模型提供了更丰富的依存信息,能够有效提升翻译性能。

关键词:神经机器翻译;双向依存注意力;多头注意力;父词;子词
中图分类号:TP391.2　　**文献标志码**:A

Neural machine translation integrating bidirectional-dependency self-attention mechanism

LI Zhijin[1,2], LAI Hua[1,2*], WEN Yonghua[1,2], GAO Shengxiang[1,2]
(1. *Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming Yunnan* 650504, *China*;
2. *Yunnan Key Laboratory of Artificial Intelligence (Kunming University of Science and Technology), Kunming Yunnan* 650504, *China*)

Abstract: Aiming at the problem of resource scarcity in neural machine translation, a method for fusion of dependency syntactic knowledge based on a Bidirectional-Dependency self-attention mechanism (Bi-Dependency) was proposed. Firstly, an external parser was used to parse the source sentence to obtain dependency parsing data. Then, the dependency parsing data was transformed into the position vector of the parent word and the weight matrix of the child word. Finally, the dependency knowledge was integrated into the multi-head attention mechanism of the Transformer encoder. By using Bi-Dependency, the translation model was able to simultaneously pay attention to the dependency information in both directions: the parent word to the child word and the child word to the parent word. Experimental results of bi-directional translation show that compared with the Transformer model, in the case of rich resources, the proposed method has the BLEU (BiLingual Evaluation Understudy) value on Chinese-Thai translation improved by 1.07 and 0.86 respectively, and the BLEU value on Chinese-English translation improved by 0.79 and 0.68 respectively; in the case of low resources, the proposed model has the BLEU value increased by 0.51 and 1.06 respectively on Chinese-Thai translation, and the BLEU value increased by 1.04 and 0.40 respectively on Chinese-English translation. It can be seen that Bi-Dependency provides the model with richer dependence information, which can effectively improve the translation performance.

Key words: neural machine translation; bidirectional-dependency attention; multi-head attention; parent word; child word

0　引言

近年来神经机器翻译方法取得了重要的进展,在主要语种间的翻译质量已经接近人工翻译水平。神经机器翻译主要依赖大规模的语料,在低资源情况下,神经机器翻译模型的翻译质量会出现明显下降。为解决资源稀缺的问题,神经机器翻译一般采用融合句法信息的方法。Eriguchi等[1]率先提出了一种端到端的语法神经机器翻译模型,该模型是基于

收稿日期:2021-10-22;**修回日期**:2022-01-06;**录用日期**:2022-01-24。　**基金项目**:国家自然科学基金资助项目(61732005, 61972186, 61762056, 61761026);云南省重大科技专项(202002AD080001-5, 202103AA080015);云南省高新技术产业专项(201606)。
作者简介:李治瑾(1997—),男,辽宁大连人,硕士研究生,主要研究方向:机器翻译、自然语言处理; 赖华(1966—),男,广西钦州人,副教授,硕士,主要研究方向:智能信息处理; 文永华(1979—),男(白族),云南大理人,博士研究生,主要研究方向:机器翻译; 高盛祥(1977—),女,云南大理人,副教授,博士,主要研究方向:机器翻译、自然语言处理、信息检索。

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

注意力机制的编解码模型并融合了成分句法信息，使解码器生成的单词与短语以及源句子的单词进行软对齐。Aharoni等[2]提出了一种融合目标语言成分句法的方法，通过翻译得到线性化或词汇化的成分句法结构，将目标语言的句法信息融入神经机器翻译模型；但该方法不针对低资源神经机器翻译。Gū等[3]提出在神经机器翻译模型中增加具有语法识别解码的成分句法结构，利用具有语法感知的注意力模型和对句子结构敏感的语言模型，提升翻译质量，该方法在句子语义流畅度上取得了较好的效果。以上研究主要基于循环神经网络（Recurrent Neural Network，RNN）[4]和长短期记忆（Long Short-Term Memory，LSTM）[5]模型框架。目前Transformer模型已经成为了基线模型，因此本文的研究基于Transformer模型。目前在依存句法融合方面一般只融合依存句法中的父词信息，利用子词到父词的方向能够确定句子中父词的位置信息，得到父词位置向量，增强句子中父词对机器翻译的影响；而父词到子词的方向较少被融合到机器翻译模型中。本文认为父词到子词方向能够提供句子中子词的位置信息，通过遍历得到句子中的全部子词位置信息，构建子词权重矩阵，增强子词对机器翻译的影响。这种明确的父词到子词和子词到父词的双向关联关系对机器翻译可能更加有效。由此本文提出了将这两种双向信息融合到翻译模型中的方法，通过更全面的结构信息融合提升机器翻译的性能。参照Bugliarello等[6]的方法，本文提出了双向依存自注意力机制（Bidirectional-Dependency self-attention mechanism，Bi-Dependency），将双向依存知识融合到Transformer编码器的多头注意力机制中，不仅利用了依存句法中子词到父词的信息也利用了父词到子词的信息，利用双向依存知识指导神经机器翻译。

本文的主要工作包括以下两个方面：

1）提出了基于双向依存自注意力机制的神经机器翻译模型，通过融合子词到父词和父词到子词的双向依存信息，提升了神经机器翻译的翻译效果。

2）提出了双向依存自注意力机制，将双向依存信息融合到Transformer模型编码器的多头注意力机制中；将句法结构信息有效地融入到了Transformer模型中。

1 相关工作

1.1 依存信息融合方法

针对神经机器翻译任务中资源稀缺的问题，目前的解决方法主要分为融合成分句法和融合依存句法两种方式，本文主要讨论融合依存句法的方式。融合依存句法知识利用句子中的词生成依存句法树，得到句子中词与词间的关系，这种明确的句法信息的引入有助于翻译模型更好地学习句子中的句法结构，缓解资源稀缺的问题。

Wu等[7]率先在基于RNN的翻译模型中引入了依存句法知识并提出了一种具有语法知识融合的方法，该方法有3个编码器和两个解码器同时需要提供目标语言的依存句法信息。该方法在解码端融合目标语言的依存句法信息，通过依存句法知识的指导，得到解码端的输出，但该方法不针对低资源条件下的神经机器翻译。Zhang等[8]通过将依存解析器的中间表示与单词嵌入进行级联，从而集成源语言端的语法，该方法由解析模型和神经机器翻译模型构成，将解析模型编码器生成的隐状态作为翻译模型的输入，在翻译的同时可以得到源语言句子的依存解析结果；但该方法不允许在源语言端学习字词单元。Saunders等[9]利用语法表示法对单词进行交织，提出了一种基于派生的表示形式，可以从序列中直接复制原始树，从而保持结构信息；但这样会导致更长的序列出现，且需要利用梯度累计的方式才能进行有效的训练。Choshen等[10]提出了一种基于生成转换序列的基于Transformer的树和图解码的通用方法，实验表明该方法的性能优于标准Transformer解码器。安静[11]利用依存句法将英文长句分割并证明了基于长句分割机器翻译的有效性。王振晗等[12]将源语言句法解析树融合到卷积神经网络中，在汉-越翻译中取得了很好的效果。

以上融合依存句法知识的研究主要基于RNN和LSTM模型框架进行研究，只有少量的研究是在Transformer模型框架下进行的。目前Transformer模型框架在许多双语的翻译上都取得了最佳的翻译效果，因此，本文将双向依存知识融合到Transformer模型中，以提升翻译质量。

1.2 基于Transformer模型的句法信息融合方法

Wang等[13]提出了一种隐式的集成源端语法的方法，使用端到端依存解析器的中间隐藏表示，将其隐藏为具有语法感知的单词表示。之后，将具有语法感知的单词表示形式与普通的词嵌入连接起来，以增强基本的神经机器翻译模型。该方法无需外部解析工具，但该方法并不针对低资源情况。Nguyen等[14]提出了一种具有层级累积的树结构注意力机制，将源语言句子序列解析为成分树结构后，先利用自下向上的检索累积，再进行自左向右的权重累积得到4个向量，输入到Transformer模型中，将叶子节点和非终端节点分别编码并输入到解码端。Zhang等[15]提出了通过基于互信息最大化的自监督神经深度建模的源-目标双语对齐的方法，基于神经机器翻译的词对齐，对齐源句和目标句的句法结构，通过互信息最大化源句和目标句的相互依赖性，结果显示了句法对齐的有效性和通用性。Slobodkin等[16]利用通用概念认知注解（Universal Conceptual Cognitive Annotation，UCCA）解析的方式获取源语言的解析数据，分别融入编码器或解码器并取得了较好的结果，证明了融合语义知识的有效性。张海玲等[17]提出利用句法层次化分析识别短语及句子框架并在中-英翻译上取得了较好的效果。Bugliarello等[6]提出了父母规模自注意力（Parent-scaled self-attention，Pascal）机制和一种将语法知识融入Transformer模型的方法，将依存信息中子词到父词的信息融合到多头注意力机制中，该方法是一种新颖的、无需参数的、具有依赖性的自注意力机制，可提高翻译质量。

以上基于Transformer模型的句法信息融合方法大多只融合子词到父词方向的信息，并未融合父词到子词方向的信息。本文提出了融合双向信息的方法，通过更全面的结构信息融合提升神经机器翻译的性能。

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

2　融合双向依存知识的神经机器翻译

本文模型基于Transformer框架,利用双向依存自注意力机制对Transformer编码器的多头注意力机制进行改进。双向依存自注意力机制的输入由源语言句子的嵌入矩阵$\boldsymbol{X}\in\boldsymbol{R}^{L\times d_{\text{model}}}$、源语言父词位置向量$\boldsymbol{P}\in\boldsymbol{R}^{L}$和源语言子词权重矩阵$\boldsymbol{C}\in\boldsymbol{R}^{L\times L}$构成,输出为双向依存自注意力机制的最终表示$\boldsymbol{M}^h$。图1展示了双向依存自注意力机制的结构。

2.1　双向依存信息

对于源语言中的父词信息,本模型利用外部解析工具得到父词位置序列,对于子词,本模型无需提供额外的依存解析工具,仅使用父词位置向量即可构建子词权重矩阵。对于根词,本文将其父词和子词定义为根词本身。

本文首先利用外部依存解析工具得到依存解析中的父词位置标记序列,从而得到句子序列中的父词位置向量$\boldsymbol{P}\in\boldsymbol{R}^{L}$,根据图1中的句子依存关系图可知句子中词与词间关系,箭头指向的词为子词,箭尾指向的词为父词,由此可知每个子词所属的父词在句子中的位置,从而得到父词位置向量$\boldsymbol{P}\in\boldsymbol{R}^{L}$。如图1中的句子,“两者”的父词为“出现”,“出现”在句子中的第3个位置,因此父词位置向量中第1个位置为3。以此类推,可得到图中的父词位置向量$\boldsymbol{P}\in\boldsymbol{R}^{L}$。

根据父词位置向量$\boldsymbol{P}\in\boldsymbol{R}^{L}$可得到源语言句子中的子词权重矩阵$\boldsymbol{C}\in\boldsymbol{R}^{L\times L}$。式(1)给出了子词权重矩阵$\boldsymbol{C}$的定义,假设$x_i$是可能的父词,则当$x_j$是$x_i$的子词时,元素$\boldsymbol{C}_{ij}$为1;否则为0。对于每个句子,使句子中的每个词与其本身对应。由于每个句子中的父词可能存在多个子词,因此本文将这些子词进行权重平均。对于根词,将其子词作为它本身并记录权重。通过这种方式,每个单词都会被告知其修饰语。

$$\boldsymbol{C}_{ij}=\begin{cases}1/n_i, & x_i\text{有多个子词}\\ 1, & i=j\text{且}x_i\text{是一个父词}\\ 0, & \text{其他}\end{cases}\tag{1}$$

其中:n_i是x_i的子词个数。同样根据图1中的依存关系图可知,每个父词拥有几个子词,例如,句子中的“出现”拥有包括其本身在内的4个子词,在子词权重平均后,子词权重矩阵第3行中的每个子词所在的位置均为1/4,其余没有子词的词语所在的行均为0,即可得到图1中的子词权重矩阵$\boldsymbol{C}\in\boldsymbol{R}^{L\times L}$。

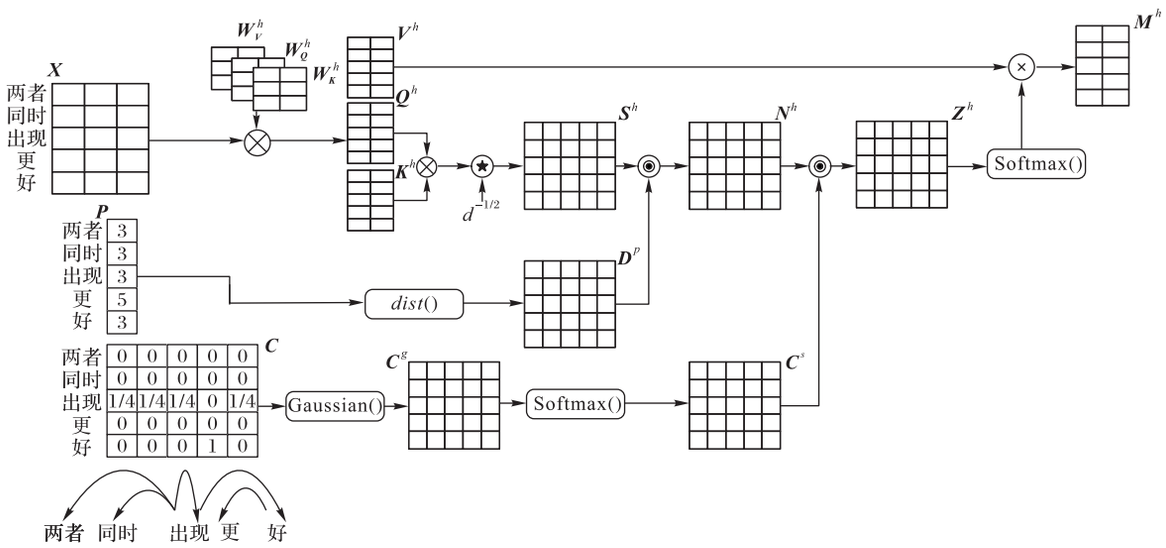

图1　双向依存自注意力机制的结构

Fig. 1　Structure of bidirectional-dependency self-attention mechanism

2.2　双向依存自注意力机制

在图1中,对于长度为L的源语言句子序列,双向依存自注意力机制中每个头的输入分别是嵌入矩阵$\boldsymbol{X}\in\boldsymbol{R}^{L\times d_{\text{model}}}$、源语言句子的父词位置向量$\boldsymbol{P}\in\boldsymbol{R}^{L}$和源语言句子的子词权重矩阵$\boldsymbol{C}\in\boldsymbol{R}^{L\times L}$。根据Vaswani等[18]的研究,在每一个注意力机制的头中,为每个标记进行计算可得到3个向量,分别是查询、键和值,从而得到3个矩阵$\boldsymbol{K}^h\in\boldsymbol{R}^{L\times d}$、$\boldsymbol{Q}^h\in\boldsymbol{R}^{L\times d}$和$\boldsymbol{V}^h\in\boldsymbol{R}^{L\times d}$,其中$d=d_{\text{model}}/H$,$H$为注意力机制中头的数量。之后计算每个查询、键和值,给出在给定位置编码时,要在输入的其他位置上设定的焦点分数,再将分数除以$\sqrt{d}$可以缓解点积较大时出现的梯度消失问题,如式(2)所示:

$$\boldsymbol{S}^h=\boldsymbol{Q}^h\boldsymbol{K}^{h\,\mathrm{T}}\big/\sqrt{d}\tag{2}$$

其中:T表示矩阵的转置。根据每个标记与位置t的依存父词位置p_t之间的距离,得到在位置t处的标记得分s_t:

$$n_{ij}^h=s_{ij}^h d_{ij}^p;\ j=1,2,\cdots,T\tag{3}$$

$$d_{ij}^p=dist(p_t,j)\tag{4}$$

其中:n_{ij}^h是父词融合矩阵$\boldsymbol{N}^h\in\boldsymbol{R}^{L\times L}$的第$t$行,代表与第$t$个父词接近度的归一化分数;$d_{ij}^p$是父词距离矩阵$\boldsymbol{D}^p\in\boldsymbol{R}^{L\times L}$的第$(t,j)$个位置,其中$d_{ij}$包含每个标记$j$与依存知识中每个父词位置间的距离关系,此距离计算定义为以p_t为中心且方差为σ^2、正态分布为$\mathrm{N}(p_t,\sigma^2)$的概率密度值:

$$dist\left(p_t,j\right)=f_{\mathrm{N}}\left(j\,|\,p_t,\sigma^2\right)=\frac{1}{\sqrt{2\pi\sigma^2}}\mathrm{e}^{-\frac{(j-p_t)^2}{2\sigma^2}}\tag{5}$$

根据分数矩阵$\boldsymbol{S}^h\in\boldsymbol{R}^{L\times L}$和父词距离矩阵$\boldsymbol{D}^p\in\boldsymbol{R}^{L\times L}$可得到父词融合矩阵$\boldsymbol{N}^h\in\boldsymbol{R}^{L\times L}$:

$$\boldsymbol{N}^h=\boldsymbol{S}^h\odot\boldsymbol{D}^p\tag{6}$$

利用解析完成的父词信息,构建子词权重矩阵$\boldsymbol{C}\in\boldsymbol{R}^{L\times L}$,此权重矩阵根据输入序列中每个词拥有子词的数

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

量构建，式(1)给出了子词权重矩阵 $\boldsymbol{C}$ 的定义。

由图1可知，子词权重矩阵 $\boldsymbol{C} \in \boldsymbol{R}^{L\times L}$ 中存在过多的零元素，本文利用加入高斯噪声的方式处理这些零元素。为子词权重矩阵 $\boldsymbol{C} \in \boldsymbol{R}^{L\times L}$ 中的元素添加高斯噪声，此高斯噪声是以 c_t 为中心且方差为 σ^2、正态分布为 $\mathrm{N}(c_t,\sigma^2)$ 的概率密度的公式，(t,j) 表示子词权重矩阵中的每个元素，利用这种方式可得到子词高斯权重矩阵 $\boldsymbol{C}^g \in \boldsymbol{R}^{L\times L}$：

$$\boldsymbol{C}_{ij}^{g} = f_N\left(j\,|\,c_t,\sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}}\mathrm{e}^{-\frac{(j-c_t)^2}{2\sigma^2}} \tag{7}$$

得到子词高斯权重矩阵 $\boldsymbol{C}^g \in \boldsymbol{R}^{L\times L}$ 后，为提高模型的收敛速度，利用Softmax函数对矩阵进行归一化处理，得到子词依存矩阵 $\boldsymbol{C}^s \in \boldsymbol{R}^{L\times L}$：

$$\boldsymbol{C}^s = \mathrm{Softmax}\left(\boldsymbol{C}^g\right) \tag{8}$$

根据父词融合矩阵 $\boldsymbol{N}^h \in \boldsymbol{R}^{L\times L}$ 和子词依存矩阵 $\boldsymbol{C}^s \in \boldsymbol{R}^{L\times L}$，可得到子词融合矩阵 $\boldsymbol{Z}^h \in \boldsymbol{R}^{L\times L}$：

$$\boldsymbol{Z}^h = \boldsymbol{N}^h \odot \boldsymbol{C}^s \tag{9}$$

最后，应用Softmax函数为句子中的每个标记生成权重分配，再将得到的子词融合矩阵 $\boldsymbol{Z}^h \in \boldsymbol{R}^{L\times L}$ 与值矩阵 $\boldsymbol{V}^h \in \boldsymbol{R}^{L\times d}$ 相乘，获得双向依存自注意力机制头的最终表示 $\boldsymbol{M}^h$。

由于本文模型将依存标记融合到了翻译模型中，因此在计算损失 L 时，使用交叉熵损失函数，如式(10)所示：

$$L = -[l_i\,\ln(p_i) + (1-l_i)\ln(1-p_i)] \tag{10}$$

其中：l_i 表示第 i 个父词的标记，正类为1，负类为0；p_i 表示第 i 个样本为正的概率。

本文同时利用子词到父词和父词到子词方向的信息且无需配置其他额外的训练参数。$\boldsymbol{D}^p \in \boldsymbol{R}^{L\times L}$ 的距离仅取决于父词的位置向量，子词权重矩阵 $\boldsymbol{C}^s \in \boldsymbol{R}^{L\times L}$ 只取决于每个词在句子中拥有的子词数量。本文模型在构建子词权重矩阵时，无需使用外部解析器，可以在训练模型前节省大量对源语言进行解析的时间。双向依存自注意力机制模型是Bugliarello等[6]的父母规模自注意力机制的扩展，本文加入了父词到子词方向的依存信息，增强神经机器翻译中子词对机器翻译的影响。

2.3 多头注意力机制

双向依存自注意力机制是对多头注意力机制中点积注意力机制的扩展，图2展示了双向依存自注意力机制在多头注意力机制中的融合方法。本文在多头注意力机制中的融合方法进行了设计上的选择，实验结果将在3.5.2节介绍。本文的双向依存自注意力机制仅在多头注意力机制的第1层的8个头中进行融合，如图2所示，编码器的整体结构并未更改，输入句子 x 进行词嵌入和位置编码后输入到多头注意力机制中，在多头注意力机制中，经过线性化的查询 $\boldsymbol{Q}$、键 $\boldsymbol{K}$、值 $\boldsymbol{V}$，父词位置向量 $\boldsymbol{P} \in \boldsymbol{R}^{L}$ 和子词权重矩阵 $\boldsymbol{C} \in \boldsymbol{R}^{L\times L}$ 作为输入，输入到双向依存自注意力机制中，得到每个头的最终表示 $\boldsymbol{M}^h$。其余处理过程与基本的Transformer模型相同，未对编码器其他位置进行改变。双向依存自注意力机制融合到了多头注意力机制的8个头中，且只在第1层融合双向依存自注意力机制取得最好的效果，这在本文随后的实验中得以验证，证明了双向依存自注意力机制的有效性。

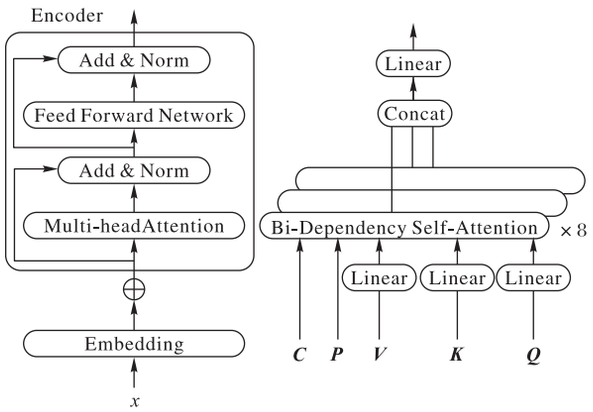

图2 多头注意力机制

Fig. 2 Multi-head attention mechanism

2.4 双向依存信息忽略

根据Bugliarello等[6]的方法，由于缺乏与标准解析工具平行的语料库，因此本模型的父词依存知识依赖于外部依存解析工具的结果；但根据Dredze等[19]的研究，对域外数据进行评估时，依存解析工具的性能会下降。为防止本文模型过度拟合到嘈杂的依赖性，本文为双向依存自注意力机制引入了两种正则化的技术，分别是父词信息忽略和子词信息忽略的方法。这种方法与Srivastava等[20]的dropout方法类似，会在模型训练阶段忽略一定的父词和子词信息。通过以一定的概率 q，将父词距离矩阵 $\boldsymbol{D}^p \in \boldsymbol{R}^{L\times L}$ 和子词依存矩阵 $\boldsymbol{C}^s \in \boldsymbol{R}^{L\times L}$ 中的每一行随机设置为 $\boldsymbol{1} \in \boldsymbol{R}^{L}$ 来忽略父词的位置信息和子词的依存信息。这两种正则化技术的引入可以生成不同父词距离矩阵和子词依存矩阵，使模型学习不同的矩阵形式，最终通过取平均的策略，防止过拟合的问题。

3 实验与结果分析

3.1 实验数据

为验证本文提出的基于双向依存自注意力机制的神经机器翻译融合方法，本文分别在汉-泰、汉-英，英-德上进行了双向翻译实验，并压缩了汉-泰和汉-英的训练数据进行了双向翻译实验，其中：汉-泰语料是通过互联网爬取的106万的平行语料，分成训练集、验证集和测试集；汉-英语料为CWMT语料库，该语料库由801万个句子对组成，作为训练集，使用newsdev2017作为验证集，使用newstest2017作为测试集；英-德语料为IWSLT14的英德语料库，该语料库由17.4万个句子对组成，作为训练集和验证集，使用newstest2015作为测试集。压缩后的汉-泰和汉-英训练数据为20万的平行语料，用汉-泰$_{小}$和汉-英$_{小}$表示，如表1所示。

表1 数据集详情

Tab. 1 Details of datasets

语料	语料总数	训练集	验证集	测试集
汉-泰	1 066 004	1 055 002	5 001	6 001
汉-英	8 021 474	8 018 471	2 002	1 001
英-德	174 272	160 239	7 283	6 750
汉-泰$_{小}$	211 002	200 000	5 001	6 001
汉-英$_{小}$	203 003	200 000	2 002	1 001

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

3.2　数据预处理

本文模型采用和Vaswani等[18]相同的预处理步骤,使用Koehn等[21]提出的Moses模型对数据进行标记,并从源语言端和目标语言端删除了超过80个标记的句子,步骤如下:

1)数据筛选。首先删除超过80个标记的句子,之后删除存在乱码的数据,最后通过人工筛选删除存在错误的句子。

2)分词。对于汉语本文使用jieba分词,对于泰语使用JointCut进行分词,对于英语和德语,本文不将其分为字符级。

3)依存句法解析。汉语使用LTP语言云平台进行分词和依存解析,泰语利用spaCy-Thai进行分词和依存解析,英语和德语使用Stanford CoreNLP进行依存解析。为保证分词结果与依存解析的结果可以一一对应,本文在进行依存解析前,不对源语言句子进行分词,直接利用依存解析工具的分词结果来保证模型的输入不会发生错误。

4)字节对编码(Byte Pair Encoding, BPE)。本文采用BPE大小均为1.6万的词表。

3.3　模型参数设置

本文模型是基于Transformer模型的扩展,使用Fairseq工具包中的PyTorch0.4.1实现本文的模型。根据Papineni等[22]提出的通过小格网络搜索的方法,利用BLEU值作为本文的评价指标。本文选择Transformer和Pascal作为对比实验的基准模型,实验分别基于Transformer架构。所有实验均在单个NVIDIA RTX 2070 SUPER GPU上进行。本文使用Vaswani等[18]最新的Tensor2Tensor中的超参数设置,按照Vaswani等[18]的学习时间表进行了4 000个warm-up优化。类似于Szegedy等[23]的研究,在训练过程中使用的标签平滑率为0.1。在验证时使用和Wu等[24]类似的波束大小为4且长度罚分为0.6的波束搜索。本文采用的学习率为0.000 7,批次大小max-tokens为4 096,dropout为0.3。在压缩数据的实验中,本文采用了8 000个warm-up优化。

3.4　实验结果

本文分别在汉泰数据集、CWMT汉英数据集、IWSLT14英德数据集和压缩后的汉泰、汉英数据集上进行了实验,实验结果如表2所示。

表2　不同模型双向翻译的BLEU结果

Tab. 2　BLEU results of bidirectional translation among different models

模型	汉-泰	泰-汉	汉-英	英-汉	英-德	德-英	汉-泰$_{小}$	泰-汉$_{小}$	汉-英$_{小}$	英-汉$_{小}$
Transformer	9.16	7.37	21.29	19.14	28.30	34.30	3.15	2.61	10.92	9.37
Pascal	9.49	7.98	21.67	19.53	28.64	34.60	3.59	3.14	11.31	9.65
Bi-Dependency	**10.23**	**8.23**	**22.08**	**19.82**	**28.76**	**34.73**	**3.66**	**3.67**	**11.96**	**9.77**

由表2可见,Bi-Dependency在汉-泰双向翻译中,BLEU相较于Transformer的翻译结果提升了1.07和0.86;在汉-英翻译任务上,Bi-Dependency的BLEU也显著提升了0.79和0.68;在英-德上,Bi-Dependency的翻译结果与Transformer相比虽有提升但不显著。在压缩数据集后,Bi-Dependency在汉-泰双向翻译中,BLEU与Transformer模型相比分别有0.51和1.06的提升。在汉-英翻译任务中,BLEU分别提升了1.04和0.40。从表2可进一步分析出,汉-泰的翻译整体效果较差,这可能是由于在泰语和汉语进行分词时,汉语和泰语的词无法较好地对应,使得模型在学习时存在较大的偏差;而英-德上,Bi-Dependency的翻译结果提升并不明显,这可能由于目前的Transformer模型在英德上的翻译比较成熟,Transformer模型可以较好地学习到英语和德语的句法结构,从而使得融合双向依存知识的方式翻译效果不显著。本文通过融合双向依存知识,在富资源和低资源情况下的翻译质量均有一定的提升,说明本文提出的双向依存自注意力机制是有效的。

3.5　实验分析

为验证本文提出的基于双向依存自注意力机制的汉泰神经机器翻译方法的合理性,分别设计了双向依存信息、多头注意力机制不同的层中融合双向依存信息和高斯权重函数对模型翻译效果的影响实验。

3.5.1　双向依存信息对翻译结果的影响分析

为验证融合源语言双向依存信息的作用,本文在汉-英数据集上进行了融合双向依存信息的有效性实验。定义"Transformer+CWord(Child Word)"表示只融合依存知识中父词到子词方向的信息;定义"Pascal"表示只融合依存知识中子词到父词方向的信息;定义"Bi-Dependency"表示本文模型框架,实验结果如表3所示。

表3　融合单/双向依存信息的BLEU值对比

Tab. 3　BLEU values comparison of fusing with unidirectional-/bidirectional-dependency information

模型	汉-英	英-汉	汉-英$_{小}$	英-汉$_{小}$
Transformer	21.29	19.14	10.92	9.37
Transformer+CWord	21.64	19.01	10.99	9.67
Pascal	21.67	19.53	11.31	9.65
Bi-Dependency	**22.08**	**19.82**	**11.96**	**9.77**

由表3的实验结果可以看出,Bi-Dependency取得了最好的结果:在汉-英的翻译任务上,与Transformer+CWord模型相比,BLEU分别提升了0.44和0.81,与Pascal模型相比,BLEU分别提升了0.41和0.29。在压缩数据集上,Bi-Dependency的BLEU值同样有较大的提升。根据以上的结果可以看出,在源语言端融合依存知识与基本的Transformer模型相比均有显著的提升,说明在源语言端融合依存句法知识对翻译任务是有帮助的。Transformer+CWord和Pascal的BLEU值差距很小,说明在源语言端融合依存知识中的父词或子词的翻译效果无明显差距。这可能是由于在机器翻译任务中,融合子词到父词的单向信息和父词到子词的单向信息属于相同类型的融合方式;因此,在源语言端,只融合子词到父词信息与只融合父词到子词信息的效果大致相同。本文提出的Bi-Dependency翻译模型在汉-英的双向翻译任务上取得了最高BLEU值,获得了最好的翻译效果,说明在源

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

语言端融合双向依存知识对神经机器翻译任务具有较大的帮助。

3.5.2 多头注意力机制不同的层中融合双向依存信息对翻译结果的影响分析

根据Bugliarello等[6]的研究，本文也在汉-英数据集上，在多头注意力机制不同的层上进行了双向依存自注意力机制层实验，以验证在第几层融合双向依存知识是更加有效的，实验结果如表4所示。

表4 不同注意力层中融合双向依存信息的BLEU值对比

Tab. 4 BLEU values comparison of fusing bidirectional-dependency information in different attention layers

注意力层	汉-英	英-汉	汉-英$_{小}$	英-汉$_{小}$
1	**22.08**	**19.82**	**11.96**	**9.77**
2	21.43	18.94	11.12	9.67
3	21.74	19.40	10.48	9.67
4	21.49	18.81	10.97	9.37
5	21.69	18.94	11.18	9.72
6	21.45	18.99	10.99	9.72

表4展示了双向依存自注意力机制在多头注意力机制不同层上的实验结果。通过表4可知，Bi-Denpendency模型在多头注意力机制的第一层融合双向依存句法知识取得了最好的效果。与最低的结果相比，在汉-英的翻译任务上，分别提升了0.65和1.01个BLEU值。压缩数据后，分别提升了1.48和0.40个BLEU值。当Bi-Dependency放置在较低层时，模型在测试集上的性能会明显降低。这样的结果证实了Raganato等[25]的发现：在第一层中更多的注意力仅集中在需要翻译的单词本身上，而不是其上下文。由此可以推断出，在第一层融合句法相关性可以有效地学习单词表示，从而进一步提高Transformer模型的翻译准确性。

3.5.3 高斯权重矩阵对翻译结果的影响分析

为了验证在子词权重矩阵中添加高斯噪声的作用，本文在汉-英数据集上进行了高斯权重矩阵实验，定义“Bi-Dependency-GWF(Gaussian Weight Function)”表示子词权重矩阵中不添加高斯权重函数，实验结果如表5所示。

表5 添加高斯噪声前后的BLEU值对比

Tab. 5 Comparison of BLEU values before and after adding Gaussian noise

模型	汉-英	英-汉	汉-英$_{小}$	英-汉$_{小}$
Transformer	21.29	19.14	10.92	9.37
Bi-Dependency-GWF	20.82	18.95	9.95	8.40
Bi-Dependency	**22.08**	**19.82**	**11.96**	**9.77**

根据表5可知，Bi-Dependency取得了最好的效果。在汉-英的翻译任务上，与Bi-Dependency-GWF相比，分别提高了1.26和0.87个BLEU值。在压缩数据集上，分别提高了2.01和1.37个BLEU值。而Bi-Dependency-GWF模型与基本的Transformer模型相比同样存在较大的差距，分别下降了0.47和0.19个BLEU值，压缩数据集后同样出现了明显的下降。因此在双向依存自注意力机制中不添加高斯噪声的翻译结果有明显的下降，结果低于Bi-Dependency和Transformer模型。由此可以证明高斯噪声的添加是必要且有效的。不添加高斯噪声时，翻译结果出现下降，本文认为这可能是由于子词权重矩阵中过多的0元素在与父词融合矩阵$\boldsymbol{N}^h \in \mathbf{R}^{L\times L}$进行点乘时，使得生成的子词融合矩阵$\boldsymbol{Z}^h \in \mathbf{R}^{L\times L}$中出现了过多的0元素，从而对原本的父词融合矩阵产生了大量的噪声，使得翻译效果出现了明显的下降。

4 结语

本文针对神经机器翻译任务，提出了双向依存自注意力机制(Bi-Dependency)。实验结果表明，通过在多头注意力机制中融合双向依存知识的方式，对神经机器翻译任务的质量有一定的提升；通过对比实验证明，利用双向依存知识可以给翻译模型提供更丰富的依存信息，同时这种方式对低资源翻译任务同样是有效的。通过实验结果也可看出，目前汉泰神经机器翻译的总体效果较差，这可能是由于泰语分词效果较差和实验设备限制导致的，因此，如何更好地针对汉语和泰语的特性将会是未来的研究重点。

参考文献(References)

[1] ERIGUCHI A, HASHIMOTO K, TSURUOKA Y. Tree-to-sequence attentional neural machine translation[C]// Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Stroudsburg, PA: Association for Computational Linguistics, 2016: 823-833.

[2] AHARONI R, GOLDBERG Y. Towards string-to-tree neural machine translation[C]// Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Stroudsburg, PA: Association for Computational Linguistics, 2017: 132-140.

[3] GŪ J, SHAVARANI H S, SARKAR A. Top-down tree structured decoding with syntactic connections for neural machine translation and parsing[C]// Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA: Association for Computational Linguistics, 2018: 401-413.

[4] ELMAN J L. Finding structure in time[J]. Cognitive Science, 1990, 14(2): 179-211.

[5] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

[6] BUGLIARELLO E, OKAZAKI N. Enhancing machine translation with dependency-aware self-attention[C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2020: 1618-1627.

[7] WU S Z, ZHANG D D, ZHANG Z R, et al. Dependency-to-dependency neural machine translation [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2018, 26(11): 2132-2141.

[8] ZHANG M S, LI Z H, FU G H, et al. Syntax-enhanced neural machine translation with syntax-aware word representations[C]// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Stroudsburg, PA: Association for Computational Linguistics, 2019: 1151-1161.

[9] SAUNDERS D, STAHLBERG F, DE GISPERT A, et al. Multi-representation ensembles and delayed SGD updates improve syntax-

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

based NMT[C]// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Stroudsburg, PA: Association for Computational Linguistics, 2018: 319-325.

[10] CHOSHEN L, ABEND O. Transition based graph decoder for neural machine translation[EB/OL]. (2021-01-29)[2021-05-10]. https://arxiv. org/pdf/2101. 12640v1. pdf.

[11] 安静. 基于依存句法分析与序列标注的英文长句机器翻译[J]. 兰州理工大学学报, 2018, 44(1): 100-103. (AN J. Machine translation of long English sentence based on dependency parsing and sequence labeling[J]. Journal of Lanzhou University of Technology, 2018, 44(1): 100-103.)

[12] 王振晗,何建雅琳,余正涛,等. 融合句法解析树的汉-越卷积神经机器翻译[J]. 软件学报, 2020, 31(12): 3797-3807. (WANG Z H, HE J Y L, YU Z T, et al. Chinese-Vietnamese convolutional neural machine translation with incorporating syntactic parsing tree[J]. Journal of Software, 2020, 31(12): 3797-3807.)

[13] WANG C Y, WU S Z, LIU S J. Source dependency-aware transformer with supervised self-attention[EB/OL]. (2019-09-05)[2021-05-10]. https://arxiv. org/pdf/1909. 02273.

[14] NGUYEN X P, JOTY S, HOI S, et al. Tree-structured attention with hierarchical accumulation[EB/OL]. (2021-02-19)[2021-05-10]. https://arxiv. org/pdf/2002. 08046. pdf.

[15] ZHANG T F, HUANG H Y, FENG C, et al. Self-supervised bilingual syntactic alignment for neural machine translation[C]// Proceedings of the 35th AAAI Conference on Artificial Intelligence. Palo Alto, CA: AAAI Press, 2021: 14454-14462.

[16] SLOBODKIN A, CHOSHEN L, ABEND O. Semantics-aware attention improves neural machine translation[C]// Proceedings of the 11th Joint Conference on Lexical and Computational Semantics. Stroudsburg, PA: Association for Computational Linguistics, 2022: 28-43.

[17] 张海玲,邵玉斌,杨丹,等. 基于句法规则层次化分析的神经机器翻译[J]. 小型微型计算机系统, 2021, 42(11): 2300-2306. (ZHANG H L, SHAO Y B, YANG D, et al. Neural machine translation based on hierarchical analysis of syntactic rules[J]. Journal of Chinese Computer Systems, 2021, 42(11): 2300-2306.)

[18] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook, NY: Curran Associates Inc., 2017: 6000-6010.

[19] DREDZE M, BLITZER J, TALUKDAR P P, et al. Frustratingly hard domain adaptation for dependency parsing[C]// Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Stroudsburg, PA: Association for Computational Linguistics, 2007: 1051-1055.

[20] SRIVASTAVA N, HINTON G, KRIZHEVSKY A, et al. Dropout: a simple way to prevent neural networks from overfitting[J]. Journal of Machine Learning Research, 2014, 15: 1929-1958.

[21] KOEHN P, HOANG H, BIRCH A, et al. Moses: open source toolkit for statistical machine translation[C]// Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions. Stroudsburg, PA: Association for Computational Linguistics, 2007: 177-180.

[22] PAPINENI K, ROUKOS S, WARD T, et al. BLEU: a method for automatic evaluation of machine translation[C]// Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2002: 311-318.

[23] SZEGEDY C, VANHOUCKE V, IOFFE S, et al. Rethinking the inception architecture for computer vision[C]// Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2016: 2818-2826.

[24] WU Y H, SCHUSTER M, CHEN Z F, et al. Google's neural machine translation system: bridging the gap between human and machine translation[EB/OL]. (2016-10-08)[2021-05-10]. https://arxiv. org/pdf/1609. 08144. pdf.

[25] RAGANATO A, TIEDEMANN J. An analysis of encoder representations in transformer-based machine translation[C]// Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP. Stroudsburg, PA: Association for Computational Linguistics, 2018: 287-297.

This work is partially supported by National Natural Science Foundation of China (61732005, 61972186, 61762056, 61761026), Yunnan Province Major Science and Technology Special Project (202002AD080001-5, 202103AA080015), Yunnan Province High-tech Industry Special Project (201606).

LI Zhijin, born in 1997, M. S. candidate. His research interests include machine translation, natural language processing.

LAI Hua, born in 1966, M. S., associate professor. His research interests include intelligent information processing.

WEN Yonghua, born in 1979, Ph. D. candidate. His research interests include machine translation.

GAO Shengxiang, born in 1977, Ph. D., associate professor. Her research interests include machine translation, natural language processing, information retrieval.

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net