北京大学学报(自然科学版) 第 58 卷 第 1 期 2022 年 1 月
Acta Scientiarum Naturalium Universitatis Pekinensis, Vol. 58, No. 1 (Jan. 2022)
doi: 10.13209/j.0479-8023.2021.111

融合小句对齐知识的汉英神经机器翻译

苗国义[1] 刘明童[2] 陈钰枫[1] 徐金安[1,†] 张玉洁[1] 冯文贺[3]

1. 北京交通大学计算机与信息技术学院, 北京 100044; 2. 创新工场人工智能工程院, 北京 100080;
3. 广东外语外贸大学语言工程与计算实验室, 广州 510420; † 通信作者, E-mail: jaxu@bjtu.edu.cn

摘要 针对当前神经机器翻译在捕捉复杂句内小句间的语义和结构关系方面存在不足, 导致复杂句长文本翻译的篇章连贯性不佳的问题, 提出一种融合小句对齐知识的汉英神经机器翻译方法。首先提出手工和自动相结合的标注方案, 构建大规模小句对齐的汉英平行语料库, 为模型训练提供丰富的小句级别的汉英双语对齐知识; 然后设计一种基于小句对齐学习的神经机器翻译模型, 通过融合小句对齐知识, 增强模型学习复杂句内小句间语义结构关系的能力。在 WMT17, WMT18 和 WMT19 汉英翻译任务中的实验表明, 所提出的方法可以有效地提升神经机器翻译的性能。进一步的评测分析显示, 所提方法能有效地提高汉英神经机器翻译在复杂句翻译上的篇章连贯性。

关键词 神经机器翻译; 小句对齐; 结构关系; 篇章连贯性

Incorporating Clause Alignment Knowledge into Chinese-English Neural Machine Translation

MIAO Guoyi[1], LIU Mingtong[2], CHEN Yufeng[1], XU Jin'an[1,†], ZHANG Yujie[1], FENG Wenhe[3]

1. School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044; 2. Sinovation Ventures AI Institute, Beijing, 100080; 3. Laboratory of Language Engineering and Computing, Guangdong University of Foreign Studies, Guangzhou 510420; † Corresponding author, E-mail: jaxu@bjtu.edu.cn

Abstract Currently, neural machine translation (NMT) is insufficient in capturing the semantic and structural relationships between clauses in complex sentences, which often results in poor discourse coherence of long and complex sentence translation. To address this problem, the paper proposes a Chinese-English NMT approach by integrating the clause alignment knowledge into NMT. Firstly, a labeling scheme combining manual and automatic annotation is introduced to annotate a large-scale clause aligned Chinese-English parallel corpus that provides rich clause-level Chinese-English bilingual alignment knowledge for model training. Then, a NMT model is designed based on clause alignment learning for enhancing the ability of the model to learn the semantic structure relationships between clauses within complex sentences. Experimental results on WMT17, WMT18 and WMT19 Chinese-English translation tasks demonstrate that proposed method can significantly improve the NMT performance. Evaluation and analysis show that proposed method can effectively improve the discourse coherence of complex sentence in Chinese-English machine translation.

Key words neural machine translation; clause alignment; structural relationship; discourse coherence

当前, 机器翻译模型一般基于平行的对齐语料建模[1−5], 模型依赖学习单语词与词之间的语义关联以及双语间词语语义的对齐信息, 将一种语言翻译为另一种语言, 特别地, 神经机器翻译通过注意力机制自动学习对齐信息, 展示出优越的性能。然而, 由于现有平行语料缺少小句(clause)级别的对齐信息, 使得模型难以自动学习和获取篇章结构信息, 以致在翻译复杂句时往往性能较低。

国家重点研发计划(2020AAA0108001)、国家自然科学基金(61976015, 61976016, 61876198, 61370130)和广东省基础与应用基础研究基金(2020A1515011056)资助

收稿日期: 2021–06–09; 修回日期: 2021–08–13

近年来，神经机器翻译在上下文信息表示和学习方面取得很大的进展。Jean 等[6]和 Zhang 等[7]引入额外编码器模块，对更大的上下文进行编码，并分别应用在基于 RNN 和 Transformer 的神经翻译模型中。Miculicich 等[8]利用层次注意力结构模型，通过词级和句子级分层注意力表示，融合多个上下文，并提高句子的语义表示能力。Shi 等[9]利用对抗学习方法来提高句子表示以及双语对齐学习能力。最近，Bao 等[10]提出 G-Transformer 模型，把整个篇章信息融入句子的表示中来提高对长文本语义的理解和翻译。然而，只通过增加上下文信息不能有效地解决篇章翻译连贯性等问题。从理论上看，篇章一般以小句而非大句(sentence)为基础单位。从双语差异来看，双语的篇章差异集中在复杂句层面。从汉英翻译来看，双语的主从句差异、连接词差异和指代差异等集中体现在复杂句层面[11−12]。

图 1 给出一个汉英复杂句错译的例子。一个由多个小句构成的复杂中文长句被当前性能世界一流的谷歌神经翻译系统翻译成多个孤立小句，小句间的逻辑语义关系严重偏离源语言句子的表达。例如，人工译文中由“although”引导的主从结构关系被机器错误地翻译成由“and”和“but”引导的并列结构关系。图 1 的例子清楚地表明，目前神经机器翻译无法有效地捕捉复杂句语境下小句间的篇章结构关系以及源语言与目标语言之间的篇章结构对齐知识。小句是语篇中基本的篇章结构单位[13]，基于小句的学习对机器翻译有重要的意义[11−12]，但当前的神经机器翻译研究并没有关注这一点。

针对以上问题，本文提出一种融合小句对齐知识的汉英神经机器翻译方法。在数据层面，针对训练数据稀缺的问题，我们标注了 4M 句对基于小句对齐的汉英复杂句平行语料，将汉英双语小句对齐知识显式地标注于平行语料库中，为模型训练提供丰富的小句级别的结构对齐知识。在模型层面，我们设计一种基于小句对齐学习的神经机器翻译模型，通过增强源端基于小句成分的句子语义表示，以及增强源端和目标端小句对齐学习来有效地融合小句对齐知识，鼓励模型学习复杂句内小句间的语义结构信息，提高模型对复杂句长文本翻译的篇章连贯性和衔接性。

1 基于小句对齐的汉英平行语料库构建方法

本文采用标注式建模方式，从 WMT 公开数据集中抽取 4M 对复杂句对。首先采用人工方式标注小规模语料，然后训练模型自动对复杂句对进行大规模标注，形成大规模基于小句对齐的汉英平行语料，为神经机器翻译模型提供显式汉英小句对齐知识。我们参考冯文贺[11]的小句切分与对齐方案，采用“源语优先”的对齐策略，首先按既定的汉语基本篇章单位进行切分，然后参考汉语切分结果切分英语小句，并进行汉英小句对齐。为获得自动标注的大规模语料，先进行小规模的人工标注，手工标注 10 万对复杂句的小句切分和对齐信息，在其上进行模型训练和方法验证。然后，用本文方法进行其余所有数据的自动标注。

1.1 基于序列标注的汉英小句识别方法

小句识别任务也称为基本语篇单位(elementary discourse unit, EDU)识别。受 Li 等[14]的启发，本文采用基于 Bi-LSTM-CRF 的序列标注模型来识别和切分汉英小句。我们把小句识别视为序列标注任务，从而实现小句边界的自动识别。如果一个词在小句的结束位置，则定义该词标签为“Y”；如果一个词在小句内部，但不在小句结束位置，则定义该词标签为“N”。针对模型设计，我们充分考虑词的词性特征和句法特征对小句边界的影响。首先，利用斯坦福句法分析器 Stanford CoreNLP[15]获取输入句子中每个词的词性(part of speech, POS)特征和句法特征，其中句法特征由父结点短语标记表示；然后，把预训练所得的词向量和词性以及句法特征向

源语句: 1少年姓孙，/2属马，/3比小水小着一岁，/4个头也没小水高，/5人却本分实诚。 (选自贾平凹《浮躁》)
人工译文: 1 This boy, a member of the Sun family, /2 had been born in the year of the horse. /3 Although he was a year younger /4 and a head shorter than Water Girl, /5 he was honest and sincere. (by professional translator Goldblatt, 1991)
机器译文: 1 The young man, surnamed Sun, /2 is a horse, /3 and one year younger than Xiaoshui, /4 not taller than Xiaoshui, /5 but he is honest. (by Google NMT system, 2021.5.1)

图 1 汉英神经机器翻译复杂句错译的示例

Fig. 1 An example of mistranslations in Chinese-English complex sentence machine translation

量相加，送入双向 LSTM (Bi-LSTM)[16]层去学习词的上下文特征表示；最后，Bi-LSTM 输出结果被送入 CRF[17]层，做二分类来预测当前词是否属于小句的边界。将此模型用在本文手工标注的 10 万句对数据上，为测试算法的准确性，将数据集分成 10 份，轮流将其中 9 份作为训练数据，1 份作为测试数据。对 10 万句对数据进行 10 次 10 折交叉验证，经过对每个可能切分的位置进行判断，中文小句识别效果达到 P=92.0，R=93.6，F1=92.8，英文小句的识别效果达到 P=94.6，R=93.0，F1=93.8。

1.2 基于词对齐学习的汉英小句对齐方法

汉英小句识别完成后，需要做汉英小句对齐，并为每个小句打上对齐标签和序号。传统的句对齐方法包括基于长度特征、词汇特征和位置特征等方法。本文采用 Ding 等[18]提出的基于词汇特征的句对齐方法，把双语词对齐知识融入汉英小句对齐模型。我们先使用基于统计的词对齐工具 Giza++[19]，在大规模汉英平行语料上学习到一个双语对齐词典。然后设计一个由两个双向 RNN (Bi-directional RNN)[20]构成的编码器。对汉英句对上每个词 x_i 在双语词典中查找其对齐词 y_i，这样源语小句和目标语小句都会产生一个对应的对齐词汇序列。把源语小句和目标语小句每个词与其对齐词的词向量拼接后，送入编码器的两个双向 RNN 进行训练。利用余弦距离，计算源语与目标语小句间的语义相关度矩阵。语义相关度矩阵经过最大池化，转换成一个向量，并被送入多层感知机，最终预测两个小句是否对齐。为提高汉英小句对齐精度，在对齐模型预测的基础上，本文也加入基于小句长度特征和位置特征的辅助判断机制。通过对本文手工标注的 10 万句对数据进行 10 次 10 折交叉验证测试，汉英小句对齐效果达到 P=91.4，R=89.8，F1=90.6。

图 2 给出一个汉英小句对齐的标注示例。源和目标句子都是由多个小句构成的小句复合体(复杂句)，复杂句内不同小句由标号切分开，汉英小句通过相同的标号对齐。由图 2 可见，标点并不是小句切分的唯一依据，通常是依据词之间的语义关联切分小句。

本文通过以上标注方法，采用手工和自动相结合的方式，将小句对齐知识显式地标注在 4M 句对复杂句平行语料中，为汉英神经机器翻译提供丰富的蕴含小句结构对齐知识的训练数据。另外，平行语料中所选择的每条复杂句都是多个小句的复合体，可以视为具有完整小句关联结构的篇章单位，对模型学习篇章层面的语义结构知识是有意义的。

2 融合小句对齐知识的神经机器翻译模型

为使模型有效地学习到小句对齐知识，我们设计一种基于小句对齐学习的神经机器翻译模型。一方面，增强源端基于小句成分的句子语义表示；另一方面，增强源端与目标端小句对齐学习。两方面结合起来，可以更好地提高翻译模型对复杂句内小句间结构信息的感知和学习能力。图 3 给出融合小句对齐知识的神经机器翻译模型架构。

2.1 增强源端基于小句成分的句子语义表示

本文在 Transformer[4]架构的基础上，提出一种多路协同自注意力机制(Multi-way Coordination Self-Attention，MC-SefAtt)来增强编码器源语言句子基于小句成分的语义表示，具体方法如下。

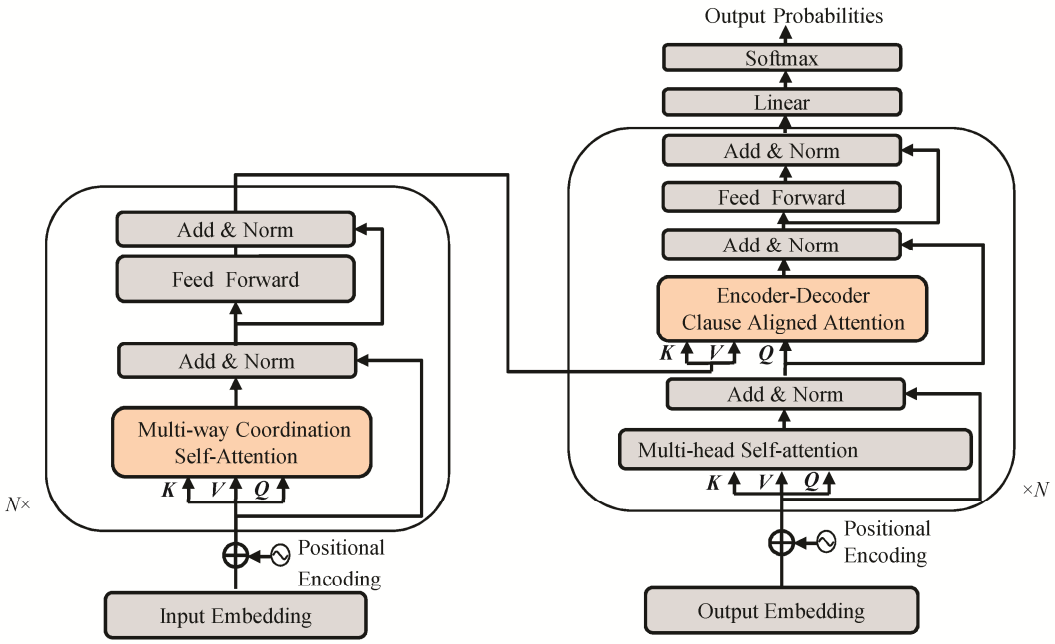
编码器由相同的 N 层堆叠构成。在编码器输入层，把输入句子的词序列每个词的词嵌入融合位置编码作为输入。由于标注数据含有大量的小句对齐标签，考虑到标签蕴含丰富的小句层面的语义结构信息，我们把每个标签视为标签词(如结构连接词)，随其他词按正常方式输入。

在编码器自注意力子层，我们提出一种多路协同自注意力机制，把输入的完整的复杂句词序列 $W=(w_1, \ldots, w_L)$ 按照每个词 w_i 的位置划分为整句序列 W 和包含 w_i 的小句序列 W_j ($1\le j\le M$, M 为输入序列的小句个数)。然后，在整句序列 W 上计算自注意力，得到句表示矩阵 $\overline{\boldsymbol{H}}$，同时对每个小句序列 W_j

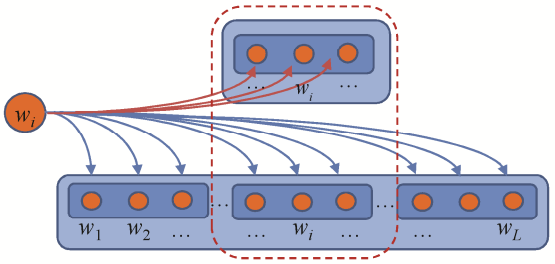
源语句: <1>少年姓孙，</1> <2>属马，</2> <3>比小水小着一岁，</3> <4>个头也没小水高，</4> <5>人却本分实诚。</5>
目标语句: <1>This boy, a member of the Sun family, </1> <2>had been born in the year of the horse. </2> <3>Although he was a year younger</3> <4>and a head shorter than Water Girl, </4> <5>he was honest and sincere. </5>

图 2　基于复杂句的汉英小句对齐标注示例

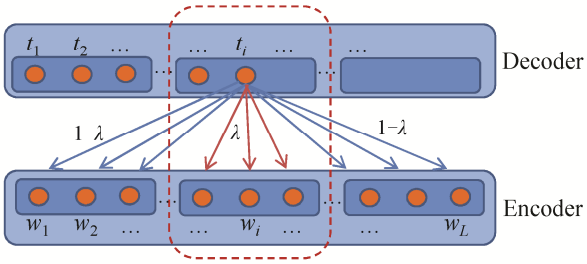
Fig. 2　An annotation example of Chinese-English clause alignment based on complex sentences



(a) 融合小句对齐知识的神经机器翻译模型



(b) 编码器端多路协同自注意力机制示意图



(c) 编码器–解码器小句对齐注意力机制示意图

图 3 融合小句对齐知识的神经机器翻译模型架构图及提出的两种注意力机制示意图

Fig. 3 Architecture of proposed NMT model by integrating clause alignment knowledge and schematic diagrams of two proposed attention mechanisms

单独计算自注意力，把计算所得的每个小句的句表示矩阵 $\boldsymbol{H}_j$ 按位置顺序累加到整句表示 $\overline{\boldsymbol{H}}$ 上，达到增强小句语义表示的目的。在整句序列 W 内的点乘自注意力计算公式如下:

$$\text{SelfAtt}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{Softmax}(\frac{\boldsymbol{Q}\boldsymbol{K}^{\text{T}}}{\sqrt{d_{\text{model}}}})\boldsymbol{V}, \quad (1)$$

SelfAtt(·)为基于放缩点乘运算的自注意力函数，$\boldsymbol{Q}$, $\boldsymbol{K}$ 和 $\boldsymbol{V}$ 分别表示从输入序列 W 转换得到的查询(query)、键(key)和值(value)的矩阵表示。d_{model} 表示模型的维度，用于放缩操作来减少矩阵的方差。最后，点乘结果经 Softmax(·)归一化操作后，与 $\boldsymbol{V}$ 相乘，得到与 $\boldsymbol{V}$ 大小相同的句表示矩阵 $\overline{\boldsymbol{H}}$。

同时，与式(1)并行计算每个小句序列 W_j 内的点乘自注意力。计算公式如下:

$$\text{SelfAtt}(\boldsymbol{Q}_j, \boldsymbol{K}_j, \boldsymbol{V}_j) = \text{Softmax}\left(\frac{\boldsymbol{Q}_j \boldsymbol{K}_j^{\text{T}}}{\sqrt{d_{\text{model}}}} + \textbf{Mask}\right)\boldsymbol{V}_j, \quad (2)$$

$\boldsymbol{Q}_j$, $\boldsymbol{K}_j$ 和 $\boldsymbol{V}_j$ 分别表示从每个小句序列 W_j 转换得到的 query, key 和 value 的矩阵表示，$1 \le j \le M$; **Mask** 为掩码矩阵，其作用是掩码掉小句以外其他的词表示，使得当前词只与对应小句内部的词做相关性计算。

经过式(2)的计算，按位置顺序得到每个小句的序列表示矩阵集合: $\{\boldsymbol{H}_1, \ldots, \boldsymbol{H}_j, \ldots, \boldsymbol{H}_M\}(1 \le j \le M)$，再把这个矩阵集合按照位置顺序拼接形成与 $\overline{\boldsymbol{H}}$ 大小相同的矩阵 $\boldsymbol{H}^*$，$\boldsymbol{H}^*$ 是蕴含丰富的小句结构信息的整句语义表示。最后，把多路径自注意力计算所得的句表示 $\overline{\boldsymbol{H}}$ 和 $\boldsymbol{H}^*$ 相加，得到编码器端 MC-SefAtt 注意力机制的句表示 $\boldsymbol{H}$。此处提出的 MC-SefAtt 注意力机制没有新增任何额外的参数。$\boldsymbol{H}$ 经过全连接前馈

神经网络 FFN(·)和层标准化操作 LayerNorm(·)的进一步抽象后，得到其所在编码器第 n 层源语言句子的语义表示 $\boldsymbol{H}^n(1\leq n\leq N)$:

$$\boldsymbol{H}=\overline{\boldsymbol{H}}+\boldsymbol{H}^*, \tag{3}$$

$$\boldsymbol{H}^n=\text{LayerNorm}(\text{FFN}(\boldsymbol{H})+\boldsymbol{H})。 \tag{4}$$

2.2 增强源端和目标端小句对齐学习

与编码器类似，解码器也由相同的 N 层堆叠构成。本文在解码器每层的自注意力机制子层和全连接前馈神经网络子层之间设计一个编码器–解码器小句对齐注意力子层(clause aligned cross attention, CA-CrossAtt)来对双语之间小句对齐信息进行建模，借助第 1 节在平行数据中标注的小句对齐标签，通过正则化的方法提高双语小句间的注意力对齐权重，鼓励模型更好地从大规模标注数据中学习基于小句的结构对齐知识，从而提高神经机器翻译对复杂句的翻译能力。

与编码器输入层处理方法类似，解码器输入层把目标语言词序列每个词的词嵌入融合位置编码作为输入，将每个标签视为标签词，随其他词按正常方式输入。

我们把解码器当前第 n 层($1\leq n\leq N$)的输入表示(即第 $n-1$ 层输出的目标语言句子表示)记为 $\boldsymbol{S}_i^{n-1}$，其中 i 表示当前时间步。$\boldsymbol{S}_i^{n-1}$ 首先经过自注意力子层和层标准化操作后得到句子表示矩阵 $\overline{\boldsymbol{S}_i^n}$，其中位置 i 的词表示为向量 $\boldsymbol{S}_i$:

$$\overline{\boldsymbol{S}_i^n}=\text{LayerNorm}(\text{SelfAtt}(\boldsymbol{S}_i^{n-1},\boldsymbol{S}_i^{n-1},\boldsymbol{S}_i^{n-1})+\boldsymbol{S}_i^{n-1}), \tag{5}$$

接下来，对解码器目标语言句表示 $\overline{\boldsymbol{S}_i^n}$ 和编码器最顶层(第 N 层)源语言句表示 $\boldsymbol{H}^N$ 之间的小句对齐信息进行建模，增强源端和目标端的小句对齐学习。本文设计一种正则化的编码器–解码器小句对齐注意力机制(CA-CrossAtt)。首先，依据解码器端目标语言句子当前词 t_i 所在小句的序号，找到源端句子对应的小句编号，例如 t_i 处于第 2 个小句，则可确定对应的源端第 2 个小句内所有词的位置范围(如 $w_m\ldots w_n$)。其次，把矩阵 $\boldsymbol{H}^N$ 预处理为两个矩阵，一个是把从第 m 至 n 行之外的部分遮蔽住(**Mask** 为 0)的矩阵 $\boldsymbol{H}_1^N$ (第 2 个小句的序列表示)，另一个是把第 m 至 n 行遮蔽住的矩阵 $\boldsymbol{H}_2^N$ (第 2 个小句之外的序列表示)，然后对这两个句表示矩阵分配不同的权重进行注意力运算:

$$\boldsymbol{C}_i^n=\text{CA-CrossAtt}(\boldsymbol{S}_i,\boldsymbol{H}^N,\boldsymbol{H}^N), \tag{6}$$

$$\begin{aligned}&\text{CA-CrossAtt}(\boldsymbol{S}_i,\boldsymbol{H}^N,\boldsymbol{H}^N)\\&=\text{Softmax}\left(\frac{\lambda\boldsymbol{S}_i(\boldsymbol{H}_1^N)^{\mathrm{T}}+(1-\lambda)\boldsymbol{S}_i(\boldsymbol{H}_2^N)^{\mathrm{T}}}{\sqrt{d_{\text{model}}}}+\textbf{Mask}\right)\boldsymbol{H}^N,\end{aligned} \tag{7}$$

其中，$\boldsymbol{S}_i$ 表示 query，$\boldsymbol{H}_1^N$ 和 $\boldsymbol{H}_2^N$ 分别作为 key 参与点乘运算，$\boldsymbol{H}^N$ 作为 value 进行计算，参数 λ 用于调节分配注意力权重，经 CA-CrossAtt(·)运算得到解码器端当前词基于源端句子的上下文向量表示 $\boldsymbol{C}_i^n$，$\boldsymbol{C}_i^n$ 蕴含编码器端当前词 t_i 对源端对齐小句的更多关注信息，对预测译文的下一个词有重要作用。并且，CA-CrossAtt 注意力子层没有新增任何额外的参数。

最后，解码器顶层的上下文向量表示 $\boldsymbol{C}_i^N$ 经过前馈神经网络和层标准化后，由 Softmax 层进行归一化操作，并预测下一个词的翻译概率:

$$\boldsymbol{S}_i^N=\text{LayerNorm}(\text{FFN}(\boldsymbol{C}_i^N)+\boldsymbol{C}_i^N), \tag{8}$$

$$P(y_i\mid y_{<i},x)\propto\text{Soft max}(\boldsymbol{W}_w\boldsymbol{S}_i^N)。 \tag{9}$$

3 实验与结果分析

3.1 实验数据

当前常用的篇章级机器翻译训练数据包括 TED 演讲数据集(TED Talks)、中英字幕数据集(TVSUB)、WMT 公开评测任务提供的 News-Commentary 数据集以及 Europarl 数据集等，但这些都是规模受限数据集，并且用于汉英翻译任务的数据非常稀缺。针对这种情况，我们从 WMT 大规模公开数据集 United Nations Parallel Corpus v1.0 中筛选 4M 句对汉英复杂句平行句对，并在上面标注小句对齐标签(见 1.1 节和 1.2 节)。本文用该标注数据集作为训练数据，使用中到英翻译方向的 WMT newsdev2017-ZHEN 作为验证集，使用 WMT newstest2017-ZHEN, newstest2018-ZHEN 和 newstest 2019-ZHEN 这 3 个测试集验证模型性能。为验证本文方法的有效性，基线系统都采用 4M 标注数据去掉小句对齐标签后的数据集进行训练。数据集统计信息见表 1。

3.2 实验设置

采用 BPE[21]子词切分方法，源端和目标端词表均设为 40K；编码器和解码器都设为 6 层，多头注意力头数设为 8，隐层维度和前馈神经网络维度分别设为 512 和 2048；训练集的 Batch Size 设为 64，采用 Adam Optimizer[22]优化器，优化器初始学习率设为 0.00005，Dropout[23]比率设为 0.1；其他设置采

表 1 数据集统计信息
Table 1 Statistical information of data sets

数据集	句子	词	
		中文	英文
Training data	4.0M	134.8M	151.3M
newsdev2017-ZHEN	2.0K	52.3K	61.8K
newstest2017-ZHEN	2.0K	49.1K	56.9K
newstest2018-ZHEN	3.9K	94.1K	117.5K
newstest2019-ZHEN	2.0K	60.6K	84.2K

用 Vaswani[4]系统的默认设置。本文模型的基线系统 Transformer 采用开源框架 OpenNMT[24]。

3.3 主要实验结果

本文选择对字母大小写不敏感的 BLEU-4[25]评价指标对译文进行质量评估，使用 multi-bleu.pl 脚本进行计算。与已公开发表的神经机器翻译方面的工作进行性能比较(表 2)。可以看出，与 Bahdanau 等[2]的基于 RNN 的神经机器翻译模型 RNNSearch 相比，我们的模型在 BLEU 值上平均取得 2.99 个点的提升。与 Gehring 等[3]提出的基于卷积神经结构的翻译模型 ConvS2S 相比，我们的模型平均提高 2.19 个点。与 Vaswani 等[4]提出的完全基于自注意力机制的 Transformer (base)模型相比，我们的模型平均获得 1.57 个点的提升。与 Shi 等[9]提出的基于对抗学习的句对齐学习方法相比，我们的模型平均取得 0.59 个点的提升。由于本文模型中两种注意力机制都没有新增任何参数，仅在标签输入时引入极少量参数，因此本文方法比基线系统的得分明显提高，可以排除单纯因参数量增加导致效果提升这一因素，验证了本文方法的有效性。与已有方法相比，本文方法关注小句间语义结构在整个篇章层次结构中的重要作用，通过小句的增强表示和小句对齐学习，有效地提升了机器翻译的性能。

表 2 WMT 汉–英翻译任务上的主要评测结果
Table 2 Main experimental results on WMT Chinese-English translation tasks

模型	newstest2017	newstest2018	newstest2019	Avg
RNNSearch[2]	17.26	17.43	19.87	18.19
ConvS2S[3]	17.94	18.45	20.58	18.99
Transformer (base)[4]	18.66	18.83	21.34	19.61
SentAlign[26]	19.37	19.78	22.62	20.59
本文模型	**19.98**	**20.13**	**23.44**	**21.18**

说明：粗体数字为最优结果。

3.4 模型各个部分有效性分析

我们分析了模型中各个部分对最终神经机器翻译性能的影响，实验结果如表 3 所示。

从表 3 容易看出，通过增强源端基于小句成分的句子表示和学习源端与目标端小句对齐知识，本文模型有效地改进了机器翻译性能。模型(1)在基线系统(Transformer)基础上使用多路协同自注意力机制(MC-SefAtt)，在句级语义表示的基础上融入小句级语义表示，BLEU 值比基线系统平均提升 0.48 个点，表明引入小句语义表示可以增强源语言句子表示能力，并改进神经机器翻译模型的性能。模型(2)在基线系统的基础上使用编码器–解码器小句对齐注意力机制(CA-CrossAtt)，增强了编码器和解码器在小句层面的对齐学习能力，捕获更多源端和目标端小句级语义关联特征，BLEU 值比基线系统平均提升 1.05 个点。模型(3)在基线系统的基础上同时采用 MC-SefAtt 和 CA-CrossAtt 两种注意力机制，编码器编码能力和解码器预测能力进一步提升，BLEU 值比基线系统平均提升 1.57 个点。实验结果表明，本文提出的融合小句对齐知识的方法，可以使模型有效地学到双语小句层面的语义结构对齐特征，从而提高神经机器翻译的精度。

3.5 编码器–解码器小句对齐注意力分析

本文在模型中使用基于小句对齐的编码器–解码器注意力机制，并采用正则化方法，使用参数 λ

表 3 模型各个部分有效性分析结果
Table 3 Effectiveness analysis results of each component of the model

模型	newstest2017	newstest2018	newstest2019	Avg
基线系统[4]	18.66	18.83	21.34	19.61
(1) + MC-SefAtt	19.14	19.26	21.87	20.09
(2) + CA-CrossAtt	19.41	19.68	22.89	20.66
(3) + MC-SefAtt + CA-CrossAtt	19.98	20.13	23.44	21.18

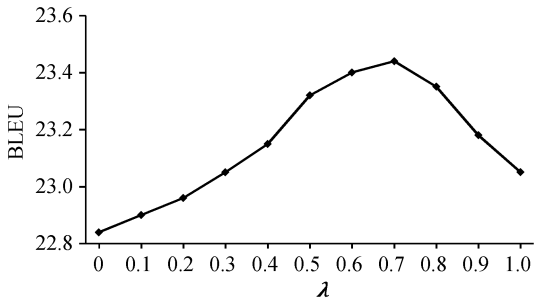


图 4 不同 λ 取值对模型性能的影响
Fig. 4 Effect of different λ values on model performance

源文：1 为规范建筑行为，2 新区管委会根据国家和上海市的有关规定，3 及时出台了一系列规范建设市场的文件，4 其中包括工程施工招投标管理办法、5 拆迁工作若干规定，6 基本做到了每个环节都有明确而又具体的规定。
Transformer: 1 In order to standardize construction behavior, 2 the New District Management Committee issued a series of documents to regulate the construction market in time 3 according to the relevant provisions of the state and Shanghai, 4 including the construction bidding management measures, 5 and several regulations on demolition work 6 have basically achieved every link.
本文模型：1 In order to standardize the construction behavior, 2 the New District Management Committee issued a series of documents to regulate the construction market in time 3 according to the relevant provisions of the state and Shanghai, 4 including the construction bidding management measures 5 and several regulations on demolition work, 6 which have basically achieved clear and specific regulations for every link.

图 5　翻译实例对比

Fig. 5　Comparison of translation examples

调节和分配注意力权重。图 4 展示在 newstest2019 测试集上不同λ取值对模型性能的影响。当λ从 0 增至 0.7 时，模型获得 0.6 个 BLEU 点的提升，表明当更多注意力分布在小句对齐信息上时，模型性能得到提升；但当λ取值超过 0.7 时，模型性能开始下降。我们认为过多的注意力分布在小句对齐上会损害模型的性能，因此把λ值设为 0.7 来优化编码器–解码器注意力机制，以便提升模型的翻译性能。

3.6 实例分析

为了进一步验证模型在复杂句上的翻译能力，我们进行翻译实例对比和分析。图 5 给出一个复杂句翻译实例，容易看出，Transformer (基线系统)的译文中，子句 5 与 6 之间出现严重的语义结构关系错误(红色标记)，并且子句 6 中出现漏译，这些翻译错误被本文模型纠正过来(蓝色标记)。该实例进一步验证了本文模型通过小句对齐知识的学习，能更好地感知和学到复杂句内小句间的结构关系，从而提高对复杂句的翻译性能，提升复杂句长文本翻译的篇章连贯性。同时也验证了本文模型通过细粒度的小句对齐学习，进一步提升源语言和目标语言句子间的对齐建模能力，使翻译充分性[26]得到提高，在一定程度上缓解了机器翻译的漏译问题，也提高了简单句的翻译效果。

4 结语

针对当前汉英复杂句机器翻译中存在的篇章连贯性问题，本文提出一种融合小句对齐知识的神经机器翻译解决方法。在数据层面，采用小规模手工和大规模自动的方式标注 4M 句对基于小句对齐的汉英复杂句平行语料，将汉英双语小句结构对齐知识显式地标注于平行语料库中，为汉英机器翻译贡献了小句对齐的平行双语数据资源。在模型层面，提出一种基于小句对齐学习的神经机器翻译模型，充分利用标注语料库提供的小句对齐信息，通过增强源端基于小句成分的句子语义表示和源端与目标端小句对齐学习来有效融合小句对齐知识，训练模型学习更多复杂句内小句层面的语义结构特征。在 WMT17，WMT18 和 WMT19 翻译任务公开测试集上的实验结果表明，本文方法能够有效地提升汉英神经机器翻译的性能。分析结果表明，本文方法在增强复杂句长文本翻译的篇章连贯性方面有明显的改进，对提高篇章翻译的效果有很大的帮助。本文提出的模型通过细粒度的小句对齐学习，增强了源端和目标端句子间的语义对齐建模能力，使机器翻译漏译问题得到改善，也提升了简单句的翻译精度。

今后的工作中，我们将考虑在小句对齐的基础上，显式地建模基于小句的语义结构信息，进一步提高神经机器翻译对复杂句长文本的翻译性能。

参考文献

[1] Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks // NIPS. Montreal, 2014: 3104–3112

[2] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate // ICLR. San Diego, 2015: 1–15

[3] Gehring J, Auli M, Grangier D, et al. Convolutional sequence to sequence learning // Proceedings of the 34th International Conference on Machine Learning. Sydney, 2017: 1243–1252

[4] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need // NIPS. Los Angeles, 2017: 5998–6008

[5] Zhang W, Feng Y, Meng F, et al. Bridging the gap between training and inference for neural machine translation // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, 2019: 4334–4343

[6] Jean S, Lauly S, Firat O, et al. Does neural machine translation benefit from larger context? [EB/OL]. (2017–04–17)[2021–03–05]. https://arxiv.org/abs/1704.05135

[7] Zhang Jiacheng, Luan Huanbo, Sun Maosong, et al. Improving the transformer translation model with document-level context // EMNLP. Brussels, 2018: 533–542

[8] Miculicich L, Ram D, Pappas N, et al. Document-level neural machine translation with hierarchical attention networks // EMNLP. Brussels, 2018: 2947–2954

[9] Shi X, Huang H, Jian P, et al. Improving neural machine translation with sentence alignment learning. Neurocomputing, 2021, 420: 15–26

[10] Bao Guangsheng, Zhang Yue, Teng Zhiyang, et al. G-transformer for document-level machine translation [EB/OL]. (2021–05–31)[2021–06–01]. https://arxiv.org/abs/2105.14761

[11] 冯文贺. 汉英篇章结构平行语料库构建与应用研究. 北京: 科学出版社, 2019

[12] 葛诗利, 宋柔. 基于成分共享的英汉小句对齐语料库标注体系研究. 中文信息学报, 2020, 34(6): 27–35

[13] Mann W, Thompson S A. Rhetorical structure theory: toward a functional theory of text organization. Text, 1988, 8(3): 243–281

[14] Li Y, Lai C, Feng J, et al. Chinese and English elementary discourse units segmentation based on Bi-LSTM-CRF model // Proceedings of the 19th Chinese National Conference on Computational Linguistics. Haikou, 2020: 1068–1078

[15] Manning C D, Mihai S, John B, et al. The Stanford CoreNLP natural language processing toolkit // Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Baltimore, 2014: 55–60

[16] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation, 1997, 9(8): 1735–1780

[17] Lafferty J, Mccallum A, Pereira F. Probabilistic models for segmenting and labeling sequence data // Proceedings of the Eighteenth International Conference on Machine Learning. Williamstown, 2001: 282–289

[18] Ding Y, Li J, Gong Z, et al. Improving neural sentence alignment with word translation. Frontiers of Computer Science, 2020, 15(1): 1–10

[19] Och F J, Ney H. A systematic comparison of various statistical alignment models. Computational Linguistics, 2003, 29(1): 19–51

[20] Cho K, van Merriënboer B, Gulcehre C, et al. Learning phrase representations using RNN encoderdecoder for statistical machine translation // Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Doha, 2014: 1724–1734

[21] Sennrich R, Haddow B, Birch A. Neural machine translation of rare words with subword units // Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin, 2016: 1715–1725

[22] Kingma D P, Ba J. Adam: a method for stochastic optimization [EB/OL]. (2014–12–22)[2021–03–06]. https://arxiv.org/abs/1412.6980

[23] Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: a simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research, 2014, 15(1): 1929–1958

[24] Klein G, Kim Y, Deng Y, et al. OpenNMT: open-source toolkit for neural machine translation // Proceedings of ACL 2017: System Demonstrations. Vancouver, 2017: 67–72

[25] Papineni K, Roukos S, Ward T, et al. BLEU: a method for automatic evaluation of machine translation // Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Philadelphia, 2002: 311–318

[26] Tu Z, Liu Y, Shang L, et al. Neural machine translation with reconstruction // Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence. San Francisco, 2017: 3097–3103