融合篇章结构位置编码的神经机器翻译

亢晓勉[1,2]，宗成庆[1,2]

（1. 中国科学院自动化研究所模式识别国家重点实验室，北京 100190；2. 中国科学院大学，北京 100049）

摘　要：现有的文档级神经机器翻译方法在翻译一个句子时大多只利用文档的上下文词汇信息，而忽视了跨句子的篇章语义单元之间的结构关系。针对此问题，提出了多种篇章结构位置编码策略，利用基于修辞结构理论的篇章树结构，对篇章树上位于不同篇章单元的单词之间的位置关系进行了表示。实验表明，通过位置编码的方式，在基于 Transformer 框架的神经机器翻译模型中有效地融合了源端的篇章结构信息，译文质量得到了显著提升。

关键词：神经机器翻译；篇章结构；位置编码；篇章分析；修辞结构理论

中图分类号：TP391

文献标识码：A

doi: 10.11959/j.issn.2096−6652.202016

Fusion of discourse structural position encoding for neural machine translation

KANG Xiaomian[1,2], ZONG Chengqing[1,2]

1. National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
2. University of Chinese Academy of Sciences, Beijing 100049, China

Abstract: Most of existing document-level neural machine translation (DocNMT) methods focus on exploring the utilization of the lexical information of context, which ignore the structural relationships among the cross-sentence discourse semantic units. Therefore, multiple discourse structural position encoding strategies were proposed to represent the positional relationships among the words in discourse units over the discourse tree based on rhetorical structure theory (RST). Experimental results show that the source-side discourse structural position information is effectively fused into the DocNMT models underlying the Transformer architecture by the position encoding, and the translation quality is improved significantly.

Key words: neural machine translation, discourse structure, position encoding, discourse analysis, rhetorical structure theory

1　引言

近年来，随着人工智能技术在自然语言处理任务中的广泛应用[1-4]，机器翻译（machine translation，MT）得到了快速发展。但是，无论是基于规则的翻译方法，还是统计机器翻译（statistical machine translation，SMT）方法和神经机器翻译（neural machine translation，NMT）方法，通常是以句子为单位进行翻译的[3]。在实际场景中，常常需要翻译一个完整的段落或者文档，此时句子级的翻译系统只能孤立地翻译文档中的每个句子。但事实上，文档具有衔接性和连贯性[5]，文档中的句子之间存在指代、省略、重复等衔接现象[6]和语义的连贯关系。因此，在翻译时应当考虑文档上下文的影响，确保生成更加准确、连贯的译文。

尽管近年来研究人员不断提出文档级别的机

收稿日期：2020−03−02；**修回日期**：2020−04−01

通信作者：宗成庆，cqzong@nlpr.ia.ac.cn

基金项目：国家自然科学基金资助项目（No.U1836221）

Foundation Item: The National Natural Science Foundation of China (No. U1836221)

器翻译方法[7-11]，但很少有工作关注篇章语义单元之间的结构关系。特别是在 NMT 系统中，目前的文档级神经机器翻译（document-level neural machine translation，DocNMT）方法主要着力于网络结构的设计，以更有效地利用上下文句子[12-18]。一部分研究者也开始针对文档中的衔接现象提出了相应的评价方法和模型[19-20]。但这些工作在利用上下文时大多直接使用注意力（attention）机制自动学习单词之间的关系，并未对篇章语言学理论中研究的篇章单元之间的结构化信息进行建模。针对这一问题，本文首次探索了在 DocNMT 系统中融合篇章结构信息。

文档的结构化表示早已引起篇章语言学者的关注[21]。他们提出了主位推进理论[22]、分段式语篇表示理论[23]等篇章理论，对文档中语义单元之间的关系进行了形式化表示。其中，修辞结构理论[24]（rhetorical structure theory，RST）得到了广泛研究和应用。RST 认为，文档可以用树形结构来表示。树的叶节点被称为基本篇章单元（elementary discourse unit，EDU），是最小的篇章语义单位。非终端节点由 2 个或多个相邻的篇章单元向上合并构成。在合并时，语义上更加重要的单元被称为“核心（nucleus）”，修饰“核心”的其他单元则被称为“卫星（satellite）”。“核心–卫星”关系又被细化为转折、递进等多种修辞关系。在图 1 所示的例子中，文档包含 3 个句子（S_1~S_3），被切分为 4 个 EDU（e_1~e_4）。图中的树结构中标注了 3 种修辞关系（证明、连接、阐述），箭头所指为“核心”单元。RST 风格的篇章自动分析器的构建任务一直是篇章分析的重要研究方向[25-27]，RST 结构也被成功应用于情感分析[28]、自动文摘[29]等自然处理任务中。在机器翻译中，参考文献[30]基于目标端 RST 结构设计了评价方法。参考文献[31]在 SMT 系统中针对复句的翻译提出了根据 RST 结构对 EDU 的翻译进行调序的方法。但在 NMT 系统中，尚未有工作探索如何利用篇章的结构信息。

本文针对 NMT 中 Transformer 结构的特点，首次提出在文档翻译中采用位置编码的方式来融合基于 RST 的篇章结构信息。本文以段落为单位进行翻译。首先，笔者通过已有的篇章分析工具对源端待翻译的段落进行解析，得到对应的篇章树。之后，本文提出了 5 种简单而有效的策略，对每个单词在篇章树中所属的 EDU 范围、EDU 之间的层次位置、“核心–卫星”关系等结构信息进行编码表示，通过位置编码的方式增强编码器对源语言单词的编码能力。本文在 DocNMT 模型上对提出的篇章结构位置编码策略进行了验证。在英译中和英译德任务的多个数据集上的实验结果表明，本文的方法可以有效地编码篇章中的结构信息，从而改善文档翻译的质量。在英译中任务上，翻译评价指标（bilingual evaluation understudy，BLEU）值获得了最高 0.78 个百分点的提升。

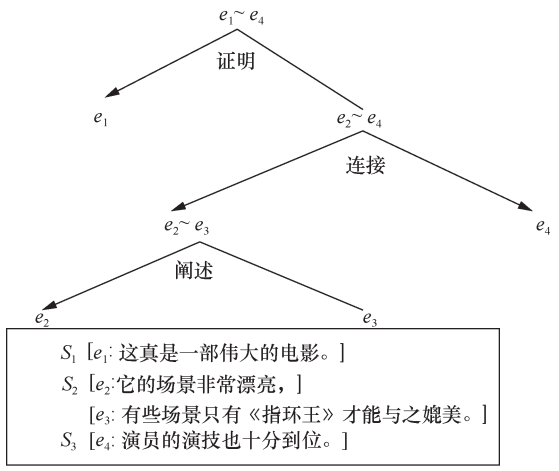

图 1　RST 篇章结构树的例子

2　研究背景和现状

2.1　文档级机器翻译

DocNMT 模型所利用的上下文既可以是源语言端的其他句子[12-15]，也可以是目标语言端翻译过的历史句子[16-17,20]。同时，根据上下文句子所在的范围，DocNMT 方法还可以被分为在线（online）方法和离线（offline）方法[15]：前者仅利用当前待翻译句子之前的句子作为上下文，而后者则使用文档中除当前翻译句子之外的所有句子作为上下文。由于篇章树结构的构建需要全局的上下文，因此在本文中，设定待翻译句子的上下文为源语言端的所有其他句子。

已有 DocNMT 方法对上下文的使用方式主要包含 2 类：级联和层次化。参考文献[13, 17, 20]将所有上下文句子级联成一个更长的单词序列，进而通过注意力机制进行编码。参考文献[14, 18]则先对每个上下文句子分别进行 attention 操作，生成各自的句子向量，再对句子向量进行 attention，生成最终的上下文语义表示。

无论设定何种上下文来源和使用方式，现有的 DocNMT 模型都没有利用篇章结构信息，且没有对

篇章结构信息进行建模。

2.2 Transformer

NMT 是目前主流的机器翻译方法。它采用端到端的序列生成框架，包括编码器和解码器 2 个部分。在翻译时，NMT 先通过编码器将源语言句子中的单词编码为语义表征向量，再由解码器根据源端的语义表征向量和已经生成的目标端历史序列，逐词地生成目标端的翻译结果。Vaswani 等人[30]于 2017 年提出了 Transformer 结构，在多个翻译任务上的性能都明显地超越了基于循环神经网络[33]和卷积神经网络[34]的 NMT 方法。本文提出的方法和基准模型是基于 Transformer 结构实现的。

Transformer 结构通过多头自注意力（multi-head self-attention）机制直接捕捉句子中任意 2 个单词之间的关系。具体地，设词向量维度为 d，源语言句子为 $\boldsymbol{X}=\{x_1,x_2,\cdots,x_J\}$，$\boldsymbol{X}\in\boldsymbol{R}^{J\times d}$。经过线性变换，可以得到 3 个不同的向量：

$$\boldsymbol{Q}=\boldsymbol{X}\boldsymbol{W}_Q,\boldsymbol{K}=\boldsymbol{X}\boldsymbol{W}_K,\boldsymbol{V}=\boldsymbol{X}\boldsymbol{W}_V \tag{1}$$

其中，$\boldsymbol{W}_Q$、$\boldsymbol{W}_K$、$\boldsymbol{W}_V$ 为线性变换矩阵。

则自注意力机制的输出 $\boldsymbol{H}$ 通过式（2）得到：

$$\boldsymbol{H}=\mathrm{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\top}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{2}$$

其中，d_k 表示 $\boldsymbol{K}$ 的维度。

通过 $\boldsymbol{Q}$ 与 $\boldsymbol{K}$ 的点积操作，自注意力机制可以建立任意 2 个单词之间的直接关联，更利于并行计算。

然而，点积造成了序列中位置信息的缺失。因此，为记录单词在句子序列中的位置 pos，Transformer 在编码和解码词向量时引入了重要的位置编码（position encoding，PE）向量。该向量由位置编码函数 TransPE(·)得到，计算过程如下：

$$\mathrm{TransPE}(\mathrm{pos})=f\left(\mathrm{pos}\Big/10\,000^{2i/d}\right) \tag{3}$$

其中，d 为向量的总维度，i 为某一维度对应的索引。当 i 为奇数时，$f(\cdot)=\sin(\cdot)$；当 i 为偶数时，$f(\cdot)=\cos(\cdot)$。

原始的位置编码采用的是单词在句子中的绝对位置。在此基础上，参考文献[35]提出了相对位置编码。参考文献[36]采用基于依存句法结构的绝对位置编码和相对位置编码，进一步提升了翻译性能。受这些工作的启发，本文探索基于 RST 树结构的位置编码，从而有效地利用篇章分析得到的结构信息来帮助提升翻译质量。

3 篇章结构位置编码

RST 表示的篇章结构树具有以下特点。

- EDU 是树的叶节点，通常由小句或短语构成。EDU 之间不存在交叉或覆盖，因此文档中的一个单词只能位于一个 EDU 中。
- 一个非终端节点由它的子节点依据修辞关系合并构成，它包含的文本不要求以句子为单位。
- 篇章树具有多层级的结构，不同 EDU 在树上的深度不同。
- 合并 2 个节点时，在语义上，“核心”比“卫星”更加重要。

针对上述 RST 篇章结构的特点，本文充分利用篇章树中的 EDU 边界、层级结构和“核心–卫星”关系等结构信息，在第 3.1~3.3 节分别设计了 5 种位置编码策略：绝对 EDU 位置编码（Abs EDU-PE）、相对 EDU 位置编码（Rel EDU-PE）、绝对深度位置编码（Abs Depth-PE）、相对深度位置编码（Rel Depth PE）、路径位置编码（Path-PE）。图 2 给出了这些编码的示例。需要注意的是，这些位置编码都是以 EDU 为单位的，因此同一个 EDU 中的单词拥有相同的篇章结构位置编码。在第 3.4 节中，笔者将这些位置编码与 DocNMT 系统进行融合。

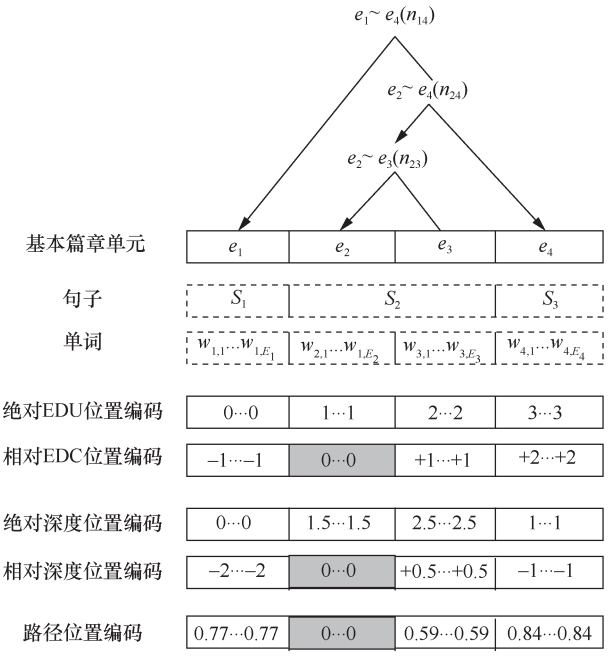

图 2 篇章结构位置编码示例

3.1 EDU 位置编码

根据单词所处 EDU 在文档中的位置，本文首先提出了 EDU 位置编码（EDU-PE）。它能够使模型在编码过程中更加清晰地区分由 EDU 分割的语义边界。本文考虑了绝对 EDU 位置编码（Abs EDU-PE）和相对 EDU 位置编码（Rel EDU-PE）2 种策略。相对 EDU 位置编码是根据上下文单词所处 EDU 相对于当前编码单词所处 EDU 的位置进行编码的，当前 EDU 中的单词的位置编码为 0，位于它前面的 EDU 编码为负值，位于它后面的 EDU 编码为正值。

3.2 深度位置编码

为了利用单词所处 EDU 在篇章树上的深度信息，本文提出了绝对深度位置编码和相对深度位置编码 2 种策略。

（1）EDU 节点的绝对深度 abs_depth 的计算

步骤 1　计算各 EDU 节点的原始深度 ori_depth。本文定义最上层 EDU 节点的原始深度为 0，其他 EDU 节点的原始深度自顶向下逐层递增。在图 2 中，e_1、e_2、e_3、e_4 的原始深度分别为 0、2、2、1。

步骤 2　若 2 个 EDU 节点互为兄弟节点且构成“核心–卫星”关系，则对它们的深度进行修正（具有“多核心”关系的 EDU 的绝对深度和相对深度不修正）。虽然这 2 个 EDU 在篇章树上的原始深度相同，但核心 EDU 比卫星 EDU 更重要，因此核心 EDU 的绝对深度 $\text{abs\_depth} = \text{ori\_depth} - 0.5$，卫星 EDU 的绝对深度 $\text{abs\_depth} = \text{ori\_depth} + 0.5$。例如图 2 中，$e_2$、$e_3$ 的绝对深度分别被修正为 2–0.5=1.5 和 2+0.5=2.5。

（2）EDU 节点的相对深度 rel_depth 的计算

步骤 1　计算各 EDU 节点的原始深度 ori_depth。其计算过程与计算绝对深度的步骤 1 相同。

步骤 2　计算 EDU 的相对原始深度 $\text{ori\_depth}_{\text{rel}}$。当前 EDU 节点 e 的相对原始深度为固定值 0。其他 EDU 节点 e' 的相对原始深度为 $\text{ori\_depth}_{\text{rel}} = \text{ori\_depth}(e') - \text{ori\_depth}(e)$。在图 2 的例子中，若 e_2 为当前 EDU，则 e_1、e_2、e_3、e_4 的相对原始深度分别为–2、0、0、–1。

步骤 3　若 2 个 EDU 节点互为兄弟节点，并且构成“核心–卫星”关系，那么需要基于相对原始深度 $\text{ori\_depth}_{\text{rel}}$ 对它们进行深度修正。修正方式与计算绝对深度的步骤 2 相同，当前 EDU 节点的相对深度不做修正。因此，e_2、e_3 的相对深度分别为 0 和 0+0.5=0.5。

3.3 路径位置编码

本节根据篇章树上 EDU 之间的路径和“核心–卫星”关系计算路径位置编码。首先，本文根据“核心–卫星”关系对篇章树上所有的边进行赋值。“核心”边的权重为常数 $w_N (w_N \geqslant 0.5)$，“卫星”边的权重为 $w_S = 1 - w_N$。其次，固定当前 EDU 节点 e 中单词的路径位置编码为 0。对任意的其他 EDU 节点 e'，通过以下 3 个步骤计算它的路径位置编码。

步骤 1　在树上寻找 e 与 e' 的共同父节点 n_{father}。

步骤 2　分别得到 e' 到 n_{father} 的路径 $\text{Path}(e' \to n_{\text{father}})$ 和 e 到 n_{father} 的路径 $\text{Path}(e \to n_{\text{father}})$。找到位于 $\text{Path}(e \to n_{\text{father}})$ 上的 n_{father} 的子节点，记作 $\tilde{n}$。

步骤 3　节点 e' 相对于当前 EDU 节点 e 的路径位置编码 $\text{PathPE}(e')$ 的计算式如下：

$$\text{PathPE}(e') = \frac{1}{1 - \sum_{w \in \text{P}(e,e')} \log w} \tag{4}$$

$$\text{P}(e,e') = \text{Path}(e' \to n_{\text{father}}) \cap \text{Path}(e \to \tilde{n})$$

在图 2 的示例中，假设当前的 EDU 节点 $e = e_2$，核心边权重 $w_N = 0.8$。在计算节点 $e' = e_1$ 相对于 e_2 的路径位置编码时，依据上述步骤可以得到 $n_{\text{father}} = n_{14}$，$\text{Path}(e' \to n_{\text{father}}) = e_1 \to n_{14}$，$\text{Path}(e \to n_{\text{father}}) = e_2 \to n_{23} \to n_{24} \to n_{14}$，$\tilde{n} = n_{24}$。因此，$\text{P}(e,e')$ 中包含的边有 3 条：$e_1 \to n_{14}(w_N)$、$e_2 \to n_{23}(w_N)$、$n_{23} \to n_{24}(w_N)$。则 e_1 相对于 e_2 的路径位置编码表示为 $1/(1 - \log 0.8 \times 3) \approx 0.77$。

3.4 与机器翻译的融合

本文将上述方法得到的各种位置表示统一称为篇章结构位置（discourse structural position，DSP）。本文将篇章结构位置编码与 Transformer 结构下的文档翻译模型进行融合。本文在实验中对比了以下 2 种融合方式。

（1）加法方式

与原始 Transformer 中的单词绝对位置编码一样，本文将经过 TransPE(·)得到的篇章结构位置编码 TransPE(DSP)直接与词向量相加。

（2）非线性方式

受参考文献[33-34]的启发，本文尝试将篇章结构位置编码 TransPE(DSP)与原始的单词绝对位置编码 TransPE(pos)通过非线性函数进行融合，得到

最终的位置编码，再与词向量相加，如式（5）所示：

$$\mathrm{PE}=\tanh\left(W\cdot\left[\mathrm{TransPE}(\mathrm{pos});\mathrm{TransPE}(\mathrm{DSP})\right]+b\right) \tag{5}$$

其中，W 和 b 是可学习的参数。多种篇章结构位置编码可以混合使用，此时非线性融合方式中的 $\mathrm{TransPE}(\mathrm{DSP})$ 为多种位置编码的级联。

4 实验设置

4.1 实验数据

本文的实验使用英译中、英译德的 TED 演讲数据和英译德 Europarl 数据。其中，TED 演讲数据来自 IWSLT17 评测，英译中和英译德的 TED 演讲数据分别包含 1 906 和 1 698 篇演讲，平均每篇演讲包含 121 个句子。在 2 个语言对上均选取 dev-2010 作为开发集，tst-2013~2015 作为测试集。考虑到 TED 数据集规模较小，本文也在大规模的 Europarl 数据上进行了实验。该数据由 Maruf 等人[18]整理提供。本文中训练集、开发集、测试集的设置与参考文献[18]一致。

在实验时，考虑到内存大小的限制，笔者对原始的文本进行段落划分，将一个段落视作一个篇章来验证本文的方法。本文采用与参考文献[17]相同的设置，以每 16 个句子作为一个段落进行划分。划分后的数据规模的统计见表 1。表中数据分别表示训练集、开发集和测试集的规模。

表 1 数据规模（训练集/开发集/测试集）的统计

数据来源	句子数目/个	段落数目/段
英译中 TED	0.23 M/0.9 K/3.9 K	15.3 K/59/261
英译德 TED	0.21 M/0.9 K/3.4 K	13.7 K/60/230
英译德 Europarl	1.67 M/3.6 K/5.1 K	156.5 K/330/477

4.2 基准模型

本文在基于 Transformer 结构的 DocNMT 模型上进行实验。为了公平起见，本文选择在编码器端对上下文信息进行融合。由于篇章树的构建要求分析篇章中的所有句子，因此本文的翻译模型使用离线的上下文，即文档中除当前句子之外的所有其他句子。因此，本文在参考文献[18]提出的 2 种使用离线上下文的文档翻译方法（FlatAtt、HierAtt）中加入篇章结构位置编码。本文将与以下 3 个基准模型进行比较。

- Base：标准句子级 Transformer 翻译模型。该模型使用参考文献[3]中的“base”模型进行参数设置。
- FlatAtt：参考文献[18]中的“Attention word”策略。即分别对每个上下文句子进行编码，再将编码后的所有上下文单词的状态向量进行拼接得到新的序列，计算当前单词与该序列中单词的 attention。
- HierAtt：该模型分别计算当前单词与每个上下文句子中单词的 attention 以及整个句子的 attention。本文采用参考文献[18]中的“H-Attention sparse-soft”策略。

本文使用开源工具 THUNMT 复现了上述 3 个基准模型。所有模型均使用 6 层编码器和 6 层解码器，多头注意力机制的头数为 8，隐变量和前馈层的维度大小分别为 512 和 2 048。在英译中 TED 任务中，英文和中文词表大小分别为 25 K 和 30 K。在英译德翻译任务中，源语言和目标语言共享同一个词表，在 TED 语料和 Europarl 语料上的词表规模分别为 15 K 和 30 K。所有语料在翻译前都要通过双字节编码（byte pair encoding，BPE）处理[37]切分为子词。由于本文提出的篇章结构位置编码得到的是词的位置表示，因此属于同一个单词的子词具有相同的篇章结构位置编码。

现有的 DocNMT 模型大多是通过两阶段法训练得到的：第一阶段训练一个句子级的翻译系统，在此基础上再在第二阶段训练文档级翻译的相关模块。本文只在 DocNMT 模型训练的第二阶段引入篇章结构位置编码。在训练时，本文以段落为单位随机打乱语料，但不改变段落内部的句子顺序。训练的最小批次设置为 3 000 个字符。本文的模型参数通过 Adam 方法进行更新，该方法中的参数 $\beta_1=0.98$，$\beta_2=0.98$。

4.3 RST 篇章分析

本文提出的方法需要提前解析被翻译的文档。RST 风格的篇章自动分析器的构建一直是篇章分析中的重要研究方向。RST 风格的篇章分析主要包括 2 个步骤：EDU 的切分和树结构的建立。目前基于神经网络的英文篇章分析器已经取得了不错的效果。由于缺少标注语料等问题，其他语言上的 RST 篇章分析的研究成果较少，因此本文以英文作为翻译的源语言来验证本文提出的方法。本文使用开源的英文 RST 篇章分析工具 DPLP 对英文段落进行解析得到树结构。不考虑修辞关系识别的结果，DPLP 结构解析的核心性（nuclearity）F1 值[27]在公开的新

闻领域测试集上可以达到 71.13 %。由于训练该工具的 RST 语料是在新闻领域进行标注的，所以本文对 DPLP 在 TED 演讲数据上的表现做了简单分析。本文从英译德 TED 语料中随机抽取 50 个段落，人工标注了它们的篇章结构树。在 50 个段落中使用 DPLP 进行自动解析的核心性 F1 值为 58.3%。

可以看出，尽管 DPLP 在 TED 演讲数据上相比标准新闻领域测试集性能有明显下降，但仍然可以正确解析多数的篇章结构。因此，本文利用该篇章分析工具的结果在 DocNMT 中引入篇章结构信息。

5　实验结果与分析

本文在英译中 TED 演讲数据、英译德 TED 演讲数据和英译德 Europarl 数据集上测试提出的篇章结构位置编码方法，用 BLEU 值评价翻译的译文质量。在解码时，束搜索的大小设为 4。

5.1　路径位置编码中的权重

为了确定在路径位置编码计算（第 3.3 节）中最优的“核心”边权重 w_N，本文在英译德 TED 开发集数据上进行调参。本文在 HierAtt 模型上使用非线性融合方式加入路径位置编码。不同“核心”边权重的 BLEU 值如图 3 所示。当 w_N 为 0.8 时，融合 Path-PE 的文档级翻译模型能生成 BLEU 值最大的译文。在后续实验中，w_N 的取值为 0.8。

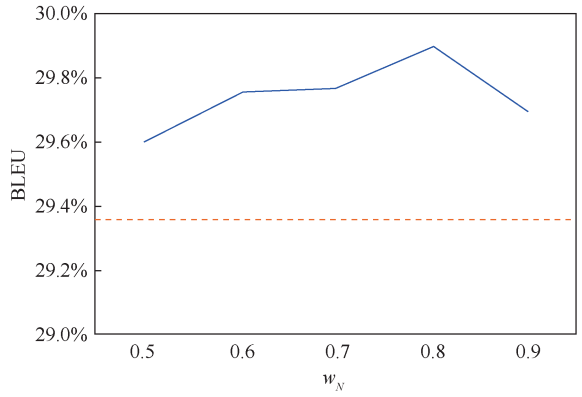

图 3　不同“核心”边权重的 BLEU 值

5.2　篇章结构位置编码策略的比较

本文首先在英译德 TED 开发集数据上讨论了不同的篇章结构位置编码策略和融合方式对 DocNMT 模型性能的影响。本节实验统一采用 HierAtt 模型。篇章结构位置编码策略的比较见表 2。

表 2 中模型 3~7 使用第 3.4 节中的加法融合方式，在 DocNMT 模型中引入篇章结构位置编码；模型 8~12 使用非线性融合方式。从表 2 可以看出以下信息。

- 在文档级翻译模型 HierAtt 中增加篇章结构位置编码后可以提升 BLEU 值，其中，通过非线性的方式融合路径位置编码（模型 12）带来的提升最大，提升了 0.51%。
- 对比 2 种融合方式可以看出，在对深度位置编码和路径位置编码进行融合时，非线性融合方式的效果优于加法融合方式。这 2 种编码策略与 RST 树的层次结构相关。而对于 EDU 位置编码的使用来说，2 种融合方式没有明显区别。
- 对比分别使用 EDU 信息（模型 3~4、8~9）、深度信息（模型 5~6、10~11）和路径信息（模型 7、12）的编码策略可以看出，路径位置编码对模型性能的改善最为显著，深度位置编码（Depth-PE）次之，EDU 位置编码（EDU-PE）带来的提升最小。
- 在加法融合方式中（模型 3 对比模型 4，模型 5 对比模型 6），绝对位置编码的翻译效果更好，而在非线性融合方式中（模型 8 对比模型 9，模型 10 对比模型 11），相对位置编码的翻译效果更好。但无论是绝对位置编码还是相对位置编码，同种融合方式下二者的差异并不显著。

表 2　篇章结构位置编码策略的比较

融合方式	编号	模型	BLEU
无	1	Base	28.63%
	2	HierAtt	29.38%
加法融合	3	+Abs EDU-PE	29.52%
	4	+Rel EDU-PE	29.49%
	5	+Abs Depth-PE	29.68%
	6	+Rel Depth-PE	29.65%
	7	+Path-PE	29.77%
非线性融合	8	+Abs EDU-PE	29.48%
	9	+Rel EDU-PE	29.53%
	10	+Abs Depth-PE	29.76%
	11	+Rel Depth-PE	29.80%
	12	+Path-PE	**29.89%**

基于上述分析，本文选择基于非线性融合方式的 3 种策略：相对 EDU 位置编码、相对深度位置编码和路径位置编码作为之后实验的篇章结构位置编码。

5.3　主要结果

本文分别在第 4.1 节所述的英译中 TED 演讲数

据、英译德 TED 演讲数据和英译德 Europarl 测试集上进行测试。表 3 展示了在 HierAtt 模型上运用非线性融合方式加入篇章结构位置编码后的 BLEU 值。表 3 中，“+”表示在 HierAtt 模型中加入篇章结构位置编码，“*”表示进行显著性检验后相较于 HierAtt 统计显著（显著性检验概率 p>0.5）。各测试集中 BLEU 值最高的结果用粗体标记。

与句子级的翻译模型（Base）相比，文档级翻译模型（HierAtt）可以借助全局的上下文提升翻译质量，在此基础上，加入本文提出的篇章结构位置编码可以进一步提升文档级翻译模型的性能。与 HierAtt 模型相比，本文的方法在英译中 TED 演讲数据、英译德 TED 演讲数据和英译德 Europarl 数据上的 BLEU 值分别取得了最高 0.78%、0.66%和 0.52%的提升。

表 3　在 HierAtt 模型上运用非线性融合方式加入篇章结构位置编码后的 BLEU 值

模型	英译中 TED	英译德 TED	英译德 Europarl
Base	21.54%	28.44%	28.87%
HierAtt	22.29%	29.31%	29.76%
+Rel EDU-PE	22.41%	29.52%	29.95%
+Rel Depth-PE	22.82%*	29.61%	30.12%*
+Path-PE	22.78%*	29.83%*	30.17%*
+Rel EDU-PE +Rel Depth-PE	**23.07**%*	29.70%	30.21%*
+Rel EDU-PE +Rel Depth-PE +Path-PE	22.98%*	**29.97**%*	**30.28**%*

同时，根据表 3 的实验结果可以得出如下结论。

- 相较于仅包含序列化 EDU 切分信息的 EDU 位置编码，基于篇章树的层级结构和“核心–卫星”关系的深度位置编码和路径位置编码对提升翻译质量有更大的帮助。
- 同时使用多种编码策略的效果优于单独使用一种编码策略。不同的编码策略可以从不同角度更全面地捕捉篇章中位于不同 EDU 之间的单词的结构关联。

5.4　篇章结构位置编码对模型的影响

本节讨论篇章结构位置编码在不同的文档级翻译模型上的影响大小。本文分别在 2 种文档级翻译模型 FlatAtt 和 HierAtt 中同时加入相对 EDU 位置编码、相对深度位置编码和路径位置编码，不同的文档级翻译模型在英译中 TED 测试集上的结果见表 4。可以看出，尽管使用层次化 attention 的 HierAtt 模型能够更好地利用上下文信息，但篇章结构位置编码对 FlatAtt 模型的提升更加显著。

表 4　不同文档级翻译模型在英译中 TED 测试集上的 BLEU 值

类型	FlatAtt	HierAtt
无篇章结构位置编码	22.06%	22.29%
融合篇章结构位置编码	23.05 %	22.98 %

6　结束语

篇章结构是语义的一种形式化表示，已经在篇章分析领域被研究多年。然而，对于文档级神经机器翻译而言，目前的方法大多只是从模型的角度出发去探索有效的网络结构，并未真正利用篇章分析的结论对模型进行指导。

本文首次尝试探索了修辞结构理论表示的篇章结构在基于 Transformer 的文档级神经机器翻译中的应用。本文提出了多种篇章结构位置编码策略，对 RST 篇章树中的 EDU 边界、深度、“核心–卫星”关系等结构信息进行了表示，并通过位置编码与文档级翻译模型进行融合，在一定程度上改善了文档级翻译模型的性能。

在未来工作中，笔者将进一步探索：如何在翻译模型中模拟对篇章结构的解析过程，减少篇章分析工具带来的误差传递；如何利用大规模单语文档数据自动地学习适合于翻译任务的篇章结构，缓解模型对篇章分析工具的依赖。

参考文献：

[1] 郑南宁. 人工智能新时代[J]. 智能科学与技术学报, 2019, 1(1): 1-3.
ZHENG N N. The new era of artificial intelligence[J]. Chinese Journal of Intelligent Science and Technology, 2019, 1(1): 1-3.

[2] 张钹. 人工智能进入后深度学习时代[J]. 智能科学与技术学报, 2019, 1(1): 4-6.
ZHANG B. Artificial intelligence is entering the post deep-learning era[J]. Chinese Journal of f Intelligent Science And Technology, 2019, 1(1): 4-6.

[3] 宗成庆. 统计自然语言处理[M]. 北京: 清华大学出版社, 2013.
ZONG C Q. Statistical natural language processing[M]. Beijing: Tsinghua University Press, 2013.

[4] 杜倩龙, 宗成庆, 苏克毅. 利用上下文相似度增强词对齐效果的自然语言推理方法[J]. 智能科学与技术学报, 2020, 2(1): 26-35.
DU Q L, ZONG C Q, SU K Y. Enhancing alignment with context similarity for natural language inference[J]. Chinese Journal of Intelligent Science and Technology, 2020, 2(1): 26-35.

[5] BROWN G, BROWN G D, BROWN G R, et al. Discourse analysis[M]. Cambridge: Cambridge University Press, 1983.

[6] HALLIDAY M A K, HASAN R. Cohesion in English[M]. [S.l.]: Routledge, 2014.

[7] GONG Z X, ZHANG M, ZHOU G D. Cache-based document-level statistical machine translation[C]//The Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2011: 909-919.

[8] XIONG D, DING Y, ZHANG M, et al. Lexical chain based cohesion models for document-level statistical machine translation[C]//The Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2013: 1563-1573.

[9] TU M, ZHOU Y, ZONG C Q. Enhancing grammatical cohesion: generating transitional expressions for SMT[C]//The Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2014: 850-860.

[10] HARDMEIER C. Discourse in statistical machine translation[D]. Uppsala: Uppsala University, 2014.

[11] WANG L Y, TU Z P, WAY A, et al. Exploiting cross-sentence context for neural machine translation[C]//The Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2017: 2826-2831.

[12] VOITA E, SERDYUKOV P, SENNRICH R, et al. Context-aware neural machine translation learns anaphora resolution[C]//The Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2019: 1264-1274.

[13] ZHANG J C, LUAN H B, SUN M S, et al. Improving the transformer translation model with document-level context[C]//The Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2018: 533-542.

[14] MICULICICH L, RAM D, PAPPAS N, et al. Document-level neural machine translation with hierarchical attention networks[C]//The Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2018: 2947-2954.

[15] YANG Z X, ZHANG J C, MENG F D, et al. Enhancing context modeling with a query-guided capsule network for document-level translation[C]//The Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2019: 1527-1537.

[16] TU Z P, LIU Y, SHI S M, et al. Learning to remember translation history with a continuous cache[J]. Transactions of the ACL, 2018, 6: 407-420.

[17] XIONG H, HE Z J, WU H, et al. Modeling coherence for discourse neural machine translation[C]//The AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2019: 7338-7345.

[18] MARUF S, MARTINS A F T, HAFFARI G. Selective attention for context-aware neural machine translation[C]//TheConference of the North American Chapter of the ACL. Stroudsburg: ACL, 2019: 3092-3102.

[19] BAWDEN R, SENNRICH R, BIRCH A, et al. Evaluating discourse phenomena in neural machine translation[C]//The Conference of the North American Chapter of the ACL. Stroudsburg: ACL, 2018: 1304-1313.

[20] VOITA E, SENNRICH R, TITOV I. When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion[C]//The Annual Meeting of the Associationfor Computational Linguistics. Stroudsburg: ACL, 2019: 1198-1212.

[21] KANG X M, ZONG C Q, XUE N W. A survey of discourse representations for Chinese discourse annotation[J]. ACM Transactions on Asian and Low-Resource Language Information Processing, 2019, 18(3): 1-25.

[22] ROTHWELL A D. Thematic progression as a functional resource in analysing texts[J]. Circulo de Linguistica Aplicada a la Communication, 2001 (5): 2.

[23] ASHER N, ALEX L. Logics of conversation[M]. Cambridge: Cambridge University Press, 2003.

[24] MANN W C, THOMPSON S A. Rhetorical structure theory: toward a functional theory of text organization[J]. Text & Talk, 1988, 8(3): 243-281.

[25] HERNAULT H, PRENDINGER H, DUVERLE D A. HILDA: adiscourse parser using support vector machine classification[J]. Dialogue & Discourse, 2010, 1(3): 1-33.

[26] FENG V W, HIRST G. Text-level discourse parsing with rich linguistic features[C]//The Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2012: 60-68.

[27] JI Y F, EISENSTEIN J. Representation learning for text-level discourse parsing[C]//The Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2014: 13-24.

[28] BHATIA P, JI Y F, EISENSTEIN J. Better document-level sentiment analysis from RST discourse parsing[C]//The Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2015: 2212-2218.

[29] GERANI S, MEHDAD Y, CARENINI G, et al. Abstractive summarization of product reviews using discourse structure[C]//The Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2014: 1602-1613.

[30] GUZMAN F, JOTY S, LIUIS A, et al. Using discourse structure improves machine translation evaluation[C]//The Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2014: 687-698.

[31] TU M, ZHOU Y, ZONG C Q. A novel translation framework based on rhetorical structure theory[C]//The Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2013: 370-374.

[32] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//The 31st Annual Conference on Advances in Neural Information Processing Systems. Boston: MIT Press, 2017: 5998-6008.

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

[34] GEHRING J, AULI M, GRAGIER D, et al. Convolutional sequence to sequence learning[C]//The 34th International Conference on Machine Learning. New York: ACM Press, 2017: 1243-1252.

[35] SHAW P, USZKOREIT J, VASWANI A. Self-attention with relative position representations[C]//The Conference of the North American Chapter of the Association for Computational Linguistics. Stroudsburg: ACL, 2018: 464-468

[36] WANG X, TU Z P, WANG L Y, et al. Self-attention with structural position representations[C]//The Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2019: 1403-1409.

[37] SENNRICH R, DADDOW B, BIRCH A. Neural machine translation of rare words with subword units[C]//The Annual Meeting of the Associationfor Computational Linguistics. Stroudsburg: ACL, 2016: 1715-1725.

[作者简介]

亢晓勉（1991− ），男，中国科学院自动化研究所模式识别国家重点实验室博士生，主要研究方向为机器翻译、篇章分析。

宗成庆（1963− ），男，博士，中国科学院自动化研究所模式识别国家重点实验室研究员、博士生导师，主要研究方向为机器翻译、自然语言处理和文本数据挖掘等。