融合全局和局部信息的汉语宏观篇章结构识别

范亚鑫[1]，蒋峰[1]，褚晓敏[1]，李培峰[1 2]，朱巧明[1 2]
[1]苏州大学计算机科学与技术学院，苏州，中国
[2]苏州大学人工智能研究院，苏州，中国
{20194227042, 20194027003}@stu.suda.edu.cn, {xmchu, pfli, qmzhu}@suda.edu.cn

摘要

作为宏观篇章分析中的基础任务，篇章结构识别任务的目的是识别相邻篇章单元之间的结构，并层次化构建篇章结构树。已有的工作只考虑局部的结构和语义信息或只考虑全局信息。因此，本文提出了一种融合全局和局部信息的指针网络模型，该模型在考虑全局的语义信息同时，又考虑局部段落间的语义关系密切程度，从而有效地提高宏观篇章结构识别的能力。在汉语宏观篇章树库（MCDTB）的实验结果表明，本文所提出的模型性能优于目前性能最好的模型。

关键词： 宏观篇章分析；结构识别；自顶向下；指针网络

Combining Global and Local Information to Recognize Chinese Macro Discourse Structure

Yaxin Fan[1], Feng Jiang[1], Xiaomin Chu[1], Peifeng Li[1 2], Qiaoming Zhu[1 2]
[1]School of Computer Science and Technology, Soochow University, Suzhou, China
[2]AI Research Institute, Soochow University, Suzhou, China
{20194227042, 20194027003}@qq.com, {xmchu, pfli, qmzhu }@suda.edu.cn

Abstract

As the fundamental task in macro discourse analysis, the discourse structure recognition task aims to identify the structure between adjacent discourse units and build a discourse structure tree hierarchically. Existing work only considers local structural and semantic information or only global information. Therefore, this paper proposes a pointer network model that integrates global and local information. It can effectively improve the ability of macro text structure recognition by considering the global semantic information and the closeness of the semantic relationship between paragraphs. The experimental results in the Chinese macro discourse treebank show that the proposed model outperforms the state-of-the-art model.

Keywords: Macro Discourse Analysis , Structure Recognition , Top-Down , Pointer Network

1 引言

©2020 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版
基金项目：国家自然科学基金(61772354,61836007,61773276);江苏高校优势学科建设工程资助项目

当前，自然语言处理的研究内容已经从词汇理解、句法分析等浅层语义分析领域延伸到深层语义理解的篇章分析领域。篇章分析是自然语言处理领域的重点和难点，其主要任务是从整体上分析一篇文章的逻辑结构和篇章单元之间的语义关系，进而从更深的层次挖掘自然语言文本的语义和结构信息。篇章分析有助于理解篇章的中心思想和主要内容，可以提升自然语言处理相关应用的性能，例如问答系统(Liakata et al., 2013)和自动文摘(Cohan and Goharian, 2015)等。

篇章分析的研究分析可分为微观和宏观两个层面。微观层面研究的是句子和句子、句群和句群之间的结构和关系；宏观层面研究的是段落和段落、章节和章节之间的结构和关系。当前篇章分析主要集中在微观层面，而宏观层面的研究较少。Chu et al. (2020)提出了一个宏观篇章结构表示体系，其中，以段落为基本篇章单元（Elementary Discourse Units,EDUs），相邻两个段落以篇章关系连接在一起，并构成更大的篇章单元（Discourse Units,DUs）,这些篇章单元层层向上，最终将一篇文章构成一棵完整的篇章结构树。

宏观汉语篇章树库（Macro Chinese Discourse Treebank,MCDTB）(Jiang et al., 2018b)对宏观篇章结构进行了标注。本文以MCDTB中的一篇文章（chtb_0282）来说明宏观篇章结构，如例1所示。其中，p_1介绍了推行公务员制度交流会的情况，p_2补充了会议时间以及参会人员；p_3讲述了李鹏总理肯定了推行公务员制度的成效，p_4讲述了李鹏总理提出推行公务员制度要依法办事；p_5补充其他参会人员。p_2补充了p_1描述的交流会的相关信息，因此p_1与p_2构成补充关系，p_2和p_4分别阐述了交流会的内容，因此构成了并列关系，其形成的篇章单元对上文（p_1和p_2构成的篇章单元）进行解说，形成解说关系，p_5是对全文的补充，即对p_1到p_4的信息进行补充。

p_1:国务院总理李鹏今天在中南海紫光阁会见中国推行公务员制度经验交流会全体代表时指出，推行公务员制度是中国政治体制改革的一项重要内容，是干部人事制度的重大改革，是建立社会主义市场经济体制的客观需要，要有领导、有步骤地加快推行步伐。
p2:这次推行公务员制度经验交流会是昨天开始召开的，各省、自治区、直辖市人事厅局长、国务院各部委、直属机构人事部门的负责人共一百二十多人出席了会议。
p3:李鹏肯定了一年来国家公务员制度推行工作取得的成效。他说，我们要认真总结和推广这些好的经验，建立起激励竞争和勤政廉政机制，建立一支以为人民服务为宗旨、密切联系群众、精干高效、廉洁奉公、忠于职守的国家公务员队伍，增强政府机关的生机和活力。
p4:李鹏提出，推行公务员制度，要按照《国家公务员暂行条例》依法办事，不能有随意性。要把这项工作作为政治体制改革的一件大事来抓，结合改革、精简机构来推行公务员制度；要形成公务员的新陈代谢机制，使青年人才不断地进入到公务员队伍当中。
p5:国务委员李贵鲜、罗干参加了会见。（完）

例1.李鹏强调要加快推行公务员制度

p_1-p_5构成的篇章结构树如图1所示。图中，叶子节点（p_1-p_5）为段落，即宏观篇章结构中的基本篇章单元（EDUs）；相邻叶子节点通过篇章关系联系起来，通过连接后构成的节点是篇章单元（DUs），表示两个基本篇章单元之间的关系；箭头指向的是核心，即重要的篇章单元。具体而言，篇章单元之间通过篇章关系相连接，最终形成一棵完整的篇章结构树。本文研究的主要内容就是识别相邻篇章单元之间的结构，并层次化构建篇章结构树。

在MCDTB语料库上，已有的篇章结构识别的研究(Jiang et al., 2018a; Zhou et al., 2019)都只考虑相邻两个篇章单元的语义关系，如果相邻两个篇章单元语义关系很接近，那么这两个篇章单元就会大概率以某种关系连接起来，形成一个更大的篇章单元，进而层次化的构建篇章结构树。但是这些研究都只考虑局部的上下文信息，而没有将整个文章的语义信息（全局信息）有效的运用到篇章结构识别任务中。

在RST-DT(Carlson et al., 2007)的篇章结构识别任务中,Lin et al. (2019)提到每次考虑相邻两个篇章单元容易受到局部信息的影响，而错误的相邻篇章结构判断会将错误的信息传播到上层，从而影响上层结构的识别。而Van (1980)的宏观篇章结构理论也指出宏观结构是更高层次的结构，表现为篇章整体的语义连贯，每一层的宏观结构都是由下层结构支撑起来的。篇章的

(c) 2020 中国中文信息学会计算语言学专业委员会

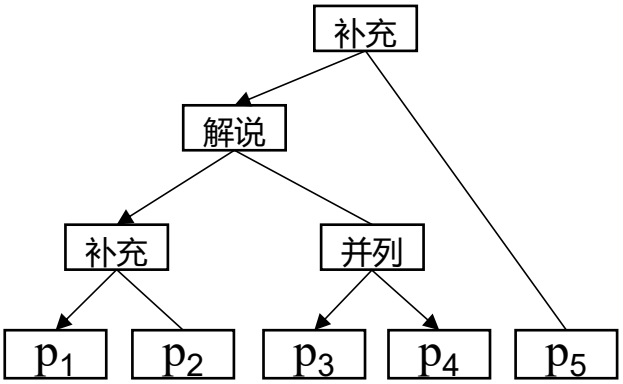


图 1. 宏观篇章结构树（chtb_0282）

宏观语义信息（即全局信息）往往能体现篇章的展开结构，可用于检验一个篇章是否连贯。因此本文认为在考虑局部信息的同时，全局信息也应该被考虑用来辅助篇章结构的识别。

基于以往的研究都只考虑局部的上下文信息，且受到宏观篇章结构理论的启发，本文提出一种融合全局和局部信息的指针网络模型，用于自顶向下的识别篇章结构，并构建篇章结构树。在该模型中，本文采用交互注意力机制捕获相邻两个段落之间的语义联系，即局部信息；指针网络的编码层用来捕获整个篇章的语义，即全局信息；而指针网络的解码层用来融合全局和局部信息，为两个段落之间的语义分配一个概率，概率越大，表明这两个段落之间的语义联系越弱，则需要进行篇章单元的切分。对切分形成的两个篇章单元，根据深度优先原则，递归地进行切分，从而自顶向下的构建完整的篇章结构树。在MCDTB上的实验结果表明，本文的模型优于目前性能最好的模型。

2 相关工作

在已有的研究工作中，无论中文还是英文都更注重微观篇章结构的分析，而对于宏观篇章结构的分析还处于起步阶段。涉及到宏观篇章结构的语料库主要有英文修辞结构篇章树库（RST Discourse Treebank,RST-DT）(Carlson et al., 2007)和中文的宏观汉语篇章树库（MCDTB）(Jiang et al., 2018b)。现将两个语料和相关模型介绍如下：

修辞结构篇章树库（RST-DT）以修辞结构理论（RST）为理论依据，标注了385篇《华尔街日报》文章。在该语料库的研究中，Hernault et al. (2010)提出了基于SVM的篇章分析器HILDA,该模型以贪婪的方式自底向上构建篇章结构树；Joty et al. (2013)等利用动态CRF模型分别构建了句子级别和篇章级别的分析器；Ji and Eisenstein (2014)参考深度学习的做法，采用线性变换将表面特征转换成隐空间通过移进规约进行篇章解析；Lin et al. (2019)采用指针网络，构建了一个句子级的篇章解析器，但上述研究都是在微观层面。在宏观层面，Sporleder and Lascarides (2004)对RST-DT修正和裁剪后采用最大熵模型进行了宏观篇章结构识别。

宏观汉语篇章树库（MCDTB）遵循RST修辞结构理论，对720篇文章进行了宏观篇章信息的标注，包括篇章结构、主次和语义关系等。在MCDTB上进行篇章结构识别，构建完整篇章结构树的研究不多。Jiang et al. (2018a)采用序列标注的思想，提出一个基于条件随机场的模型（LD-CM）。该模型对结构和主次进行联合学习，从而自底向上的构建篇章结构树；Zhou et al. (2019)提出了一个基于神经网络的模型（MVM）。该模型从多个角度匹配两个篇章单元之间的语义，从而识别篇章结构，并采用移进规约的方法构建篇章结构树。然而LD-CM是基于传统机器学习的方法，用到了较多的手工特征，考虑相邻两个篇章单元的语义联系；同样MVM也只考虑相邻两个篇章单元的语义联系。这两种方法都只考虑了局部的上下文信息，没有有效运用全局信息辅助篇章结构的识别。

3 PNGL模型

本文提出了一种融合全局和局部信息的指针网络模型（Pointer Network on Global and Local information）的模型自顶向下的识别汉语宏观篇章结构，其架构如图2所示。该架构包括三个部分:1)段落编码层（Paragraph Encoder Layer,PEL）,用来捕获段落的语义表示；2)段

(c) 2020 中国中文信息学会计算语言学专业委员会

落交互层（Paragraph Interactive Layer,PIL）,用来捕获相邻两个段落的语义联系，即局部信息；3)指针网络（Pointer Network），指针网络的编码层用来捕获整个篇章的语义表示，即全局信息，解码层融合局部和全局信息,用来识别篇章结构并自顶向下的构建篇章结构树。

对于一篇文章表示为$P = \{p_1, p_2, \cdots, p_m\}$,其中$p_i$是段落词语序列，m是文章的段落数。将$p_i$通过段落编码层（PEL），得到段落编码为$R = \{r_1, r_2, \cdots, r_m\}$。将相邻两个段落的编码通过段落交互层（PIL），得到表示相邻两个段落语义联系的表示$H = \{h_1, h_2, \cdots, h_{m-1}\}$,$h_i$表示段落$p_i$和$p_{i+1}$之间的语义联系的紧密程度，即得到了局部信息。同时将r_i平均池化之后通过指针网络编码层，编码层是双向GRU，最后一个时间步输出作为整个篇章的语义表示，即全局信息（例如e_5表示整个篇章的语义）。

指针网络解码层是单向GRU，本文根据深度优先的原则，使用栈来生成篇章结构树。在第t步，栈顶的篇章单元$DU_{(l,r)}$出栈。解码层的输入为篇章单元$DU_{(l,r)}$的语义表示e_r，即编码层第r步的输出；解码层的输出为d_t,d_t和局部信息H进行交互，通过计算注意力来融合全局信息和局部信息，从而为每一个h_i分配一个概率，其中$l \leq i \leq r-1$。概率越大，则表示段落p_i和p_{i+1}之间的语义联系越松散，则应该在p_i和p_{i+1}之间进行切分，形成新的篇章单元$DU_{(l,i)}$和$DU_{(i+1,r)}$。切分后段落数大于2的篇章单元入栈，递归地对栈顶篇章单元进行切分，直至栈空。根据切分得到的所有篇章单元构建篇章结构树。

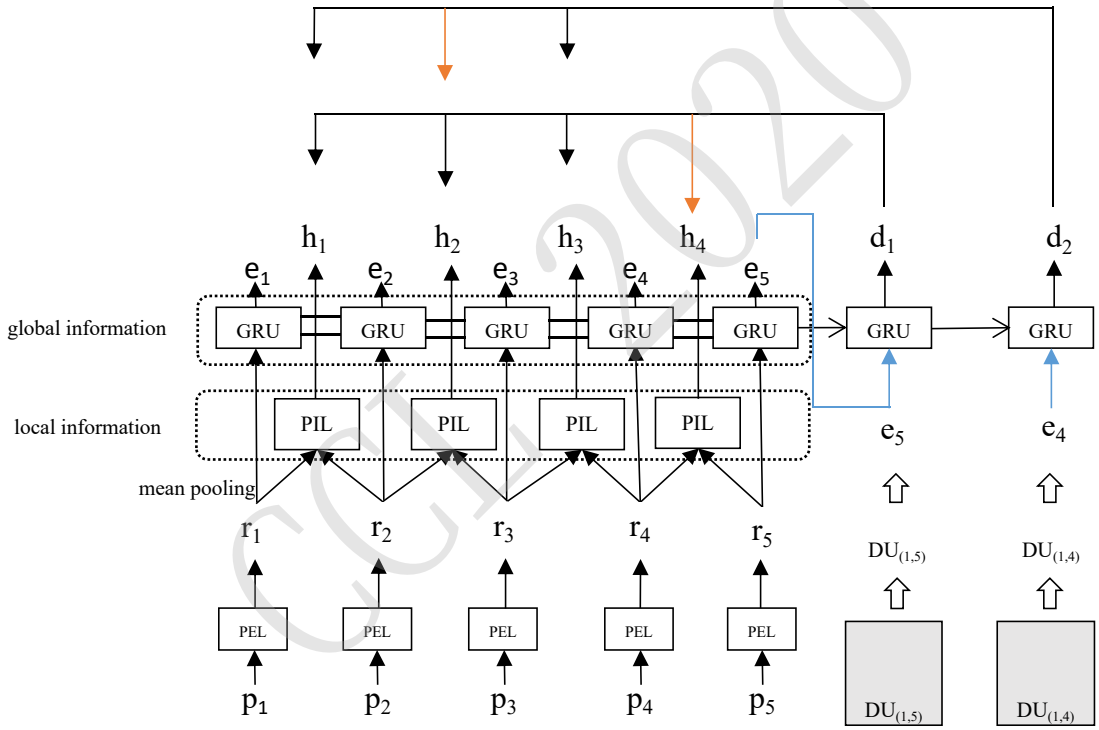


图 2. PNGL模型框架图

3.1 段落编码层

段落编码层（PEL）用来对段落进行编码，获得段落的语义信息。目前大多数的工作大多采用LSTM(Hochreiter and Schmidhuber, 1997)对输入序列进行编码。LSTM虽然具备一定长序列建模能力，但是在处理宏观篇章单元的时候，仍稍显不足。因为宏观篇章单元的最小颗粒度是段落，包含更多的词语，随着词数的增加使得篇章单元内出现更复杂的词间依赖，而LSTM按照时序来处理文本，当相距很远的词语存在依赖关系时，LSTM很难捕获到这种关系。最近，通过注意力机制直接对输入序列进行编码(Vaswani et al., 2017; Xu et al., 2019)可取得不错的效果，其计算公式如式（1）所示。

$$Attention\,(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

(c) 2020 中国中文信息学会计算语言学专业委员会

在其编码的过程中，序列中的每一个词语都与序列中的其他词语进行匹配计算，因而更容易捕获长距离词语之间的依赖关系，本质上注意力机制是对输入序列进行加权求和，因而比LSTM保留了更多的原始输入的信息。而多头注意力机制允许模型可以在不同的表示子空间中学习到相关的信息，可以使得模型更好的捕获长远距离依赖关系。因此在PNGL模型中，本文采用多头注意力机制进行段落层编码。如式（2）所示。

$$\begin{aligned} MultiHAtt\left(Q,K,V\right) &= Concat\left(head_1,\cdots,head_h\right)W^O \\ head_i &= Attention\left(QW_i^Q, KW_i^K, VW_i^V\right) \end{aligned} \tag{2}$$

参数矩阵W_i^Q $\in$R$^{d_m\times d_k}$,W_i^K $\in$R$^{d_m\times d_k}$,W_i^V $\in$R$^{d_m\times d_v}$,W^O $\in$R$^{hd_v\times d_m}$。段落编码层输入词语序列$p=\{x_1,x_2,\cdots,x_l\}$，l是段落中词语的个数，每一个词语$x_i\in R^e$使用其对应的词向量表示。得到段落编码结果$r_i\in R^{m\times d_m}$，如式（3）所示，其中$W_S^Q,W_S^K,W_S^V\in R^{d_m}$是共享的转换矩阵，从而在编码时将段落映射到相同的特征空间。

$$r_i = MultiHAttt(pW_S^Q, pW_S^K, pW_S^V) \tag{3}$$

3.2 段落交互层

段落交互层（PIL）用来捕获相邻两个段落之间的语义联系（局部信息）。一些研究人员通过注意力机制直接对序列之间的交互建模，并且提出了一些交互注意力机制。例如，Guo et al. (2018)提出一种模拟双向阅读的交互注意力机制，他从人类阅读的角度出发，发现人类在判断两个序列之间的关系时往往需要来回阅读这两个序列，尤其是考虑两个序列中联系比较紧密的词之间的语义联系。受交互注意力机制工作的影响，Xu et al. (2019)采用式（1）对序列之间的交互进行建模，并在篇章关系识别任务中取得了不错的效果，因此本文利用多头交互注意力机制获得段落之间交互的语义联系。

对于两个段落$p_1=\{x_1,x_2,\cdots,x_m\}$和$p_2=\{x_1,x_2,\cdots,x_n\}$，使用式（3）得到段落编码$r_1$和$r_2$,然后使用式（4）对段落之间的交互进行建模。

$$\begin{aligned} I_1 &= MultiHAtt\left(r_2W_{i1}^Q, r_1W_{i1}^K, r_1W_{i1}^V\right) \\ I_2 &= MultiHAtt\left(r_1W_{i2}^Q, r_2W_{i2}^K, r_2W_{i2}^V\right) \end{aligned} \tag{4}$$

式（4）首先通过转换矩阵$W_{i1}^Q,W_{i1}^K,W_{i1}^V\in R^{d_m\times d_i}$和$W_{i2}^Q,W_{i2}^K,W_{i2}^V\in R^{d_m\times d_i}$对输入序列做了映射。在多头注意力交互层，通过交换两个序列的query值，每个序列的词语都根据与另一个序列中所有词语的联系进行了重新编码，从而得到段落p_1和p_2彼此相关的向量表示$I_1\in R^{m\times d_i}$和$I_2\in R^{n\times d_i}$。最后通过平均池化操作获得包含彼此信息的段落表示$C_1,C_2\in R^{d_i}$。在包含彼此信息的段落表示C_1，C_2上，通过非线性变换进一步捕获段落之间的交互信息，将变换得到的向量h_1表示段落p_1和p_2之间语义联系的紧密程度,如式（5）所示,其中,$W_h\in R^{d_m\times 3d_i}$是参数矩阵。

$$h_1 = tanh(W_h[C_1, C_2, C_1 - C_2]) \tag{5}$$

3.3 指针网络

序列到序列的模型(Sutskever et al., 2014)提供了输入序列和输出序列长度可以不同的灵活性,但是由于该模型仍然需要固定输出词汇表的大小，而输出词表的大小取决于输入序列的长度，从而限制了需要指向输入序列某个位置的问题的适用性。而指针网络(Vinyals et al., 2015)通过使用注意力作为一个指向机制解决了这个问题。具体说来，对于输入序列$X=\{x_1,x_2,\cdots,x_n\}$，首先经过编码层得到输出$Y=\{y_1,y_2,\cdots,y_n\}$。在解码层的每一个时间步t，输出的状态d_t会和序列Y进行交互，计算注意力，然后通过softmax层获得关于输入序列的概率分布。因此，在PNGL模型中，本文运用指针网络获得关于文章相邻两个段落之间的语义联系（H）的概率分布，进而确定文章的切分位置。

(c) 2020 中国中文信息学会计算语言学专业委员会

3.3.1 编码层

Chung et al. (2014)的研究表明，GRU(Cho et al., 2014)和LSTM在很多任务上的性能不分伯仲，但是GRU拥有更少的参数，容易收敛，因此在编码层本文使用两层的双向GRU进行编码。以chtb_0282为例，本文将文章$P = \{p_1, p_2, p_3, p_4, p_5\}$通过段落编码层，得到段落编码$R = \{r_1, r_2, r_3, r_4, r_5\}$,然后采用平均池化操作输入到双向GRU中。双向GRU的输出为$E = \{e_1, e_2, e_3, e_4, e_5\}$,其中$e_i = [e_i^f; e_i^b]$。$e_i^f$和$e_i^b$分别是正向和反向的输出。此时$e_i$综合了前面i-1个段落的语义信息，即获得全局信息。而该全局信息隐含了篇章单元之间的结构信息和语义联系，对于最终篇章结构树的构建起着不可忽视的作用。

3.3.2 解码层

在解码层采用的也是一个两层的GRU。以chtb_0282为例，本文将编码层的输出$E = \{e_1, e_2, e_3, e_4, e_5\}$作为Decoder层的输入。在第t步解码时，篇章单元$DU_{(l,r)}$出栈，解码层会综合当前篇章的全局信息e_r和t步之前生成的结构语义信息生成当前状态d_t。d_t和段落交互层的输出$H = \{h_l, h_{l+1}, \cdots, h_{r-1}\}$进行交互，融合全局和局部信息，通过一个softmax层得到关于H的概率分布。如式（6）所，其中$\sigma(.,.)$是融合全局和局部信息的函数，具体为点积运算；α_t为关于H的概率分布。

$$
\begin{aligned}
s_{t,i} &= \sigma(d_t, h_i), i = l \cdots r-1 \\
\alpha_t &= softmax(s_t) = \frac{exp(s_{t,i})}{\sum_{i=l}^{r-1} exp(s_{t,i})}
\end{aligned}
\tag{6}
$$

如果通过softmax层后h_i被分配的概率值越大，表明段落p_i和p_{i+1}之间的语义联系越松散，因此更应该切分开，从而将整个篇章分为两个篇章单元$DU_{(l,i)}$和$DU_{(i+1,r)}$。根据深度优先的原则，每一步解码，段落数量大于2的篇章单元将继续入栈，递归地对篇章单元进行切分，直至栈空，过程如图3所示。

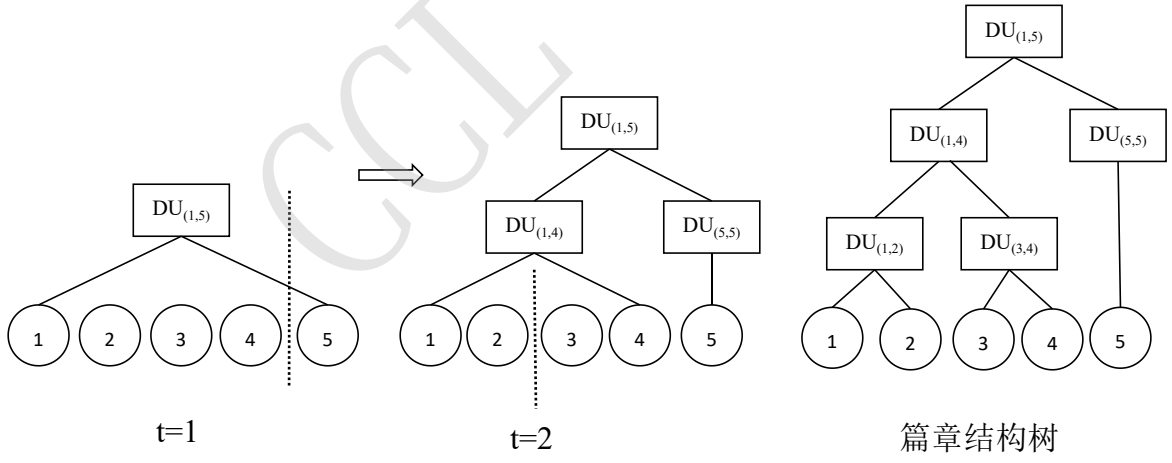


图 3. 解码过程

3.4 损失函数

在PNGL模型中，损失函数本文采用负对数似然函数进行计算，如式（7）所示。$y_{<t}$是在解码层第t步之前已经产生的篇章单元，T是入栈的篇章单元数。为了防止过拟合，本文在指针网络的编码和解码层进行了dropout操作。

$$
L(\theta_s) = -\sum_{i=1}^{batch} \sum_{t=1}^{T} logP_{\theta_s}(y_t | y_{<t}, X) \tag{7}
$$

(c) 2020 中国中文信息学会计算语言学专业委员会

4 实验

4.1 实验设置

本文在宏观汉语篇章树库（MCDTB）上对模型结构识别的性能进行了评估。MCDTB定义了三大类十五小类篇章关系，并标注了摘要，段落中心句、篇章结构等宏观篇章信息。MCDTB总计有720篇新闻报道的文章，每篇文章的段落数从2到22不等，其段落分布如表1所示。

段落	2	3	4	5	6	7	8	9	10	11	12	> 12
数量	29	122	159	144	91	58	37	33	15	13	14	15

表1.段落分布

本文使用Jiang et al. (2018a)遵循段落分布划分好的数据集进行试验，其中训练集576篇，测试集144篇。为了与Zhou et al. (2019)的实验设置一致，本文将所有的非二叉树都转换为右二叉树。另外，本文遵循Morey et al. (2017)对RST-DT上篇章结构分析模型的评价标准，同样采用内部节点正确率（等价于micro-F1）来衡量模型性能。本文将词向量维度设置为300，采用Word2Vec(Mikolov et al., 2013)进行预训练。在段落编码层和段落交互层转换矩阵映射的维度d_m和d_i都被设置为512;段落编码层多头注意力机制中头数h设置为8，其中$d_k = d_v = d_m/h = 64$;训练过程中batch大小设置为32，dropout率设置为0.5。

4.2 实验结果

本文将文中提出的模型PNGL和基准系统进行了对比，基准系统分为两种：1）只考虑局部信息2）只考虑全局信息，基准系统介绍如下：

LD-CM：性能最好的传统模型(Jiang et al., 2018a)，只考虑局部信息。该模型采用条件随机场，运用较多的手工特征，考虑相邻两个篇章单元能够合并，贪婪的自底向上识别篇章结构，从而构建篇章结构树。

MVM:性能最好的神经网络模型(Zhou et al., 2019)，只考虑局部信息。该模型从词、局部上下文以及话题这三个角度出发，提出了词对相似度机制来衡量相邻两个篇章单元的语义。并采用移进规约的方法每次考虑相邻两个篇章单元能否合并，从左到右识别篇章结构，从而构建篇章结构树。

PN:本文复现了在RST-DT上表现优异的结构识别模型PN(Lin et al., 2019)，只考虑全局信息。该模型是一个指针网络，在编码层使用双向GRU对整个文章进行编码，解码层使用单向GRU进行解码，自顶向下的识别篇章结构，构建篇章结构树。

模型	内部节点正确率（%）
LD-CM	54.71
MVM	56.11
PN	56.25
PNGL	**58.42**

表2.模型在MCDTB上的性能比较

实验结果如表2所示。PNGL模型比仅考虑局部信息的LD-CM模型性能提升了3.71，比仅考虑局部信息的MVM模型（目前在MCDTB上最好的结构识别的模型）性能提升了2.31，比仅考虑全局信息的PN模型性能提升了2.17。宏观篇章结构理论(Van, 1980)指出，文章会有一个总摄全篇的主题，并层层分解，由下层命题展开。这说明段落或篇章单元之间的关系并非很松散，都是在对主题进行分层面的展开叙述。

而LD-CM和MVM都是考虑相邻两个篇章单元联系的紧密程度，但是这两个篇章单元是围绕共同的主题展开的，如果仅仅考虑两个篇章单元之间的联系，模型往往会偏向于将这两个篇章单元合并成更大的篇章单元。而PN模型通过考虑整个篇章单元的语义信息，将篇章单元切分成两个较小的篇章单元。PN模型会对所有可能形成的两个较小篇章单元语义联系的紧密程度进行排序，取语义联系最松散的两个较小篇章单元作为切分结果。但是每个篇章单元往往包含较复杂的段落语义信息，仅仅考虑全局信息，模型很难对两个较小篇章单元之间的语义联系的紧密程度进行正确的排序。

(c) 2020 中国中文信息学会计算语言学专业委员会

本文的模型PNGL通过改进段落的语义编码，在指针网络编码层学习到更好的全局信息的同时，又考虑相邻两个段落之间语义联系的紧密程度，从而在性能上有所提升，这说明综合考虑全局和局部信息对于识别篇章结构并构建篇章结构树非常有效。

5 实验分析

5.1 全局和局部信息的影响

以往的研究表明(Lin et al., 2019)，采用基于转移的方法进行结构识别，往往对于底层的识别能力比较好，而上层的识别能力比较差。主要原因是每一步的识别都只考虑局部信息，这会将错误传播到后续步骤，导致上层的结构的识别能力较差。

为了研究局部信息和全局信息分别对底层和顶层结构识别的影响，本文在PNGL模型的基础之上去掉段落交互层，即只考虑全局信息，得到模型PNGL(-local)。本文对只考虑局部信息最好的模型MVM以及只考虑全局信息最好的模型PNGL(-local)在最底下两层内部节点正确率和最顶上三层内部节点正确率[0]进行了统计分析，如表3所示。

模型	最底下两层节部结点正确率%	最顶上三层内部节点正确率%
MVM	**46.95**	60.28
PNGL(-local)	42.68	**65.35**

表3.局部和全局信息分别对底层和顶层结构识别的影响

由表3的实验结果可知，相比于只考虑全局信息的模型，MVM在最底下两层节点正确率更高，这说明考虑局部信息的对于底层结构识别有帮助。PNGL（-local）在最上三层的节点正确率要高于MVM，说明相比于考虑局部信息的模型，只考虑全局信息对上层结构识别有帮助。因此本文认为在全局信息的基础上加入局部信息可以增强模型对于底层节点的识别能力。

为了研究在全局信息的基础之上融合局部信息对于结构识别的影响，本文在模型PN的基础之上，加入段落交互层，综合考虑全局和局部信息，得到模型PN(+local)；而PN和PNGL（-local）都是只考虑全局信息的指针网络模型，它们的区别在于PN采用双向GRU对段落进行编码，而PNGL（-local）采用多头注意力机制对段落进行编码。本文统计了内部结点正确率以及最底下两层内部结点正确率，如表4所示。

模型	内部节点正确率（%）	最底下两层内部节点正确率（%）
PN	56.25	46.65
PN(+local)	56.87	47.26
PNGL(-local)	56.57	42.68
PNGL	58.42	48.48

表4.加入局部信息后模型识别性能比较

表4实验结果表明，在加入局部信息之后PNGL和PN（+local）的最底下两层内部节点正确率分别提高了1.01和5.8。PNGL相较于PN（+local），性能有更多的提升，其原因在于PN（+local）是直接使用双向GRU对段落进行编码，而PNGL是使用多头注意力机制对段落进行编码，由于多头注意力机制相较于双向GRU更容易捕获长距离单词之间的依赖关系，能保留更多原始的信息，对段落的编码更有效。那么相邻两个段落的编码输入到段落交互层进行交互，段落交互层就能更好的捕获段落之间的语义联系。通过捕获到更好的局部信息，模型PNGL增强了对底层结构的识别能力，从而从整体上提高了模型的性能。

5.2 模型对长短文识别性能比较

为了比较模型对于长文和短文的识别能力，本文分别统计了长文和短文内部节点正确率，如表5所示。从表中数据可知，模型对短文结构的识别性能较好，而长文结构的识别性能较差。主要原因在于无论采用什么方法构建篇章结构树，都会产生级联错误，而对长文来说，则更加明显。但和只考虑局部信息的模型以及只考虑全局信息的模型相比，本文的模型PNGL综合考虑全局和局部信息，在短文和长文的结构识别的性能都有提升。

[0]由于最顶层的根节点所表示的结构总是固定，因此本文考虑最顶上三层和最底下两层内部结点正确率来表示模型对于顶层和底层结构识别的性能好坏。

(c) 2020 中国中文信息学会计算语言学专业委员会

模型	内部节点正确率%	
	≤ 6	> 6
LD-CM	65.24	42.23
MVM	65.81	44.59
PN	66.38	44.26
PNGL	**68.95**	**46.62**

表5.模型对长短文结构识别性能比较

5.3 不同模型结构识别的比较

图4从左到右展示了只考虑局部信息、只考虑全局信息以及考虑全局和局部信息的模型对chtb_0756（文章内容及标准结构树见附录A）的预测结果。MVM应用栈和队列，采用移进规约的方法，考虑栈顶的篇章单元和队首的段落能否合并成一个更大的篇章单元，如果可以合并则采取规约操作，否则采取移进操作。由于MVM只考虑局部信息，在从左到右进行结构识别的时候，未能识别出来相邻两个段落之间是否要合并成一个大的篇章单元，因此采用了一系列的移进操作，当队列中为空之后，又采取一系列的规约操作，最终形成如图所示的结构树。

PNGL（-local）采用栈数据结构，通过自顶向下的方法递归确定文章的切分位置，从而形成结构树。PNGL(-local)首先会对$DU_{(1,1)}$和$DU_{(2,5)}$、$DU_{(1,2)}$和$DU_{(3,5)}$、$DU_{(1,3)}$和$DU_{(4,5)}$、$DU_{(1,4)}$和$DU_{(5,5)}$这四个语义联系的紧密程度进行排序，确定$DU_{(1,4)}$和$DU_{(5,5)}$之间的语义联系最松散，然后递归地对$DU_{(1,4)}$进行以上过程，确定$DU_{(1,2)}$和$DU_{(3,4)}$之间的语义联系最松散，最终形成如图所示的结构树。但由于篇章单元中往往有多个段落，包含的语义信息比较复杂，如果只考虑全局信息，会使得模型很难对相邻篇章单元之间的紧密程度进行正确排序。而本文的模型PNGL通过加入相邻两个段落之间的语义联系（局部信息），考虑到了篇章单元边界的信息，从而提升了模型结构识别的能力。

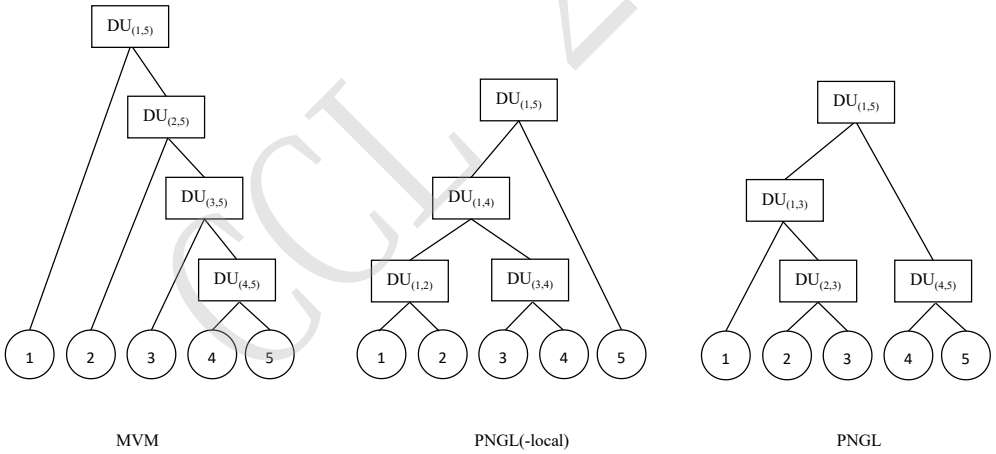


图 4. 不同模型构建的文章chtb_0756的篇章结构树

6 总结

本文针对宏观汉语篇章结构识别任务，提出了一种融合全局和局部信息的指针网络模型PNGL用于自顶向下的识别篇章结构，构建篇章结构树。其中，段落编码层采用多头注意力机制，可以有效地捕获词语之间的长距离依赖；段落交互层通过多头注意力交互机制捕获段落和段落之间的语义联系，即局部信息；指针网络的编码层用来捕获全局信息，解码层会融合全局和局部信息进行解码，自顶向下的识别篇章结构，构建篇章结构树。在MCDTB实验结果表明，本文的模型PNGL比传统机器学习的方法LD-CM性能提高了3.71%，比目前最好的模型MVM性能提高了2.31%，证明了融合全局和局部信息在篇章结构识别任务中的有效性。由于模型识别短文的性能比较好，因此在下一步工作中将融入话题分割的思想，尝试将长文划分成短文本，从而提高长文的结构识别的性能。

(c) 2020 中国中文信息学会计算语言学专业委员会

参考文献

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2007. RST Discourse Treebank.Current and new directions in discourse and dialogue. pages 85–112.

Kyunghyun Cho, Bart Van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1724–1734.

Xiaomin Chu, Xuefeng Xi, Feng Jiang, Sheng Xu, Qiaoming Zhu, and Guodong Zhou. 2020. Macro discourse structure representation schema and corpus construction. *Journal of Software*, 31(2):321–343.

Junyoung Chung, Caglar Gulcehre, Kyung Hyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *Computer Research Repository*.

Arman Cohan and Nazli Goharian. 2015. Scientific article summarization using citation-context and article's discourse structure. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 390–400.

Fengyu Guo, Ruifang He, Di Jin, Jianwu Dang, Longbiao Wang, and Xiangang Li. 2018. Implicit discourse relation recognition using neural tensor network with interactive attention and sparse learning. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 547–558.

Hugo Hernault, Helmut Prendinger, David A. Duverle, and Mitsuru Ishizuka. 2010. HILDA: A discourse parser using support vector machine classification. *Dialogue and Discourse*, 1(3):1–33.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, pages 1735–1780.

Yangfeng Ji and Jacob Eisenstein. 2014. Representation learning for text-level discourse parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13–24.

Feng Jiang, Peifeng Li, Xiaomin Chu, Qiaoming Zhu, and Guodong Zhou. 2018a. Recognizing macro Chinese discourse structure on label degeneracy combination model. In *Proceedings of the 7th CCF International Conference on Natural Language Processing and Chinese Computing*, pages 92–104.

Feng Jiang, Sheng Xu, Xiaomin Chu, Peifeng Li, Qiaoming Zhu, and Guodong Zhou. 2018b. MCDTB: A macro-level Chinese discourse treebank. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3493–3504.

Shafiq Joty, Giuseppe Carenini, Raymond Ng, and Yashar Mehdad. 2013. Combining intra- and multi-sentential rhetorical parsing for document-level discourse analysis. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 486–496.

Maria Liakata, Simon Dobnik, Shyamasree Saha, Colin Batchelor, and Dietrich Rebholz-Schuhmann. 2013. A discourse-driven content model for summarising scientific articles evaluated in a complex question answering task. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 747–757.

Xiang Lin, Shafiq Joty, Prathyusha Jwalapuram, and M Saiful Bari. 2019. A unified linear-time framework for sentence-level discourse parsing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4190–4200.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 27th Annual Conference on Neural Information Processing Systems*, pages 3111–3119.

Mathieu Morey, Philippe Muller, and Nicholas Asher. 2017. How much progress have we made on RST discourse parsing? A replication study of recent results on the RST-DT. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1319–1324.

Caroline Sporleder and Alex Lascarides. 2004. Combining hierarchical clustering and machine learning to predict high-level discourse structure. In *Proceedings of the 20th International Conference on Computational Linguistics*, pages 43–49.

(c) 2020 中国中文信息学会计算语言学专业委员会

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the 28th Annual Conference on Neural Information Processing Systems*, pages 3104–3112.

Dijk T A Van. 1980. Macrostructures : An interdisciplinary study of global structures in discourse, interaction, and cognition. *hillsdale, n.j.*

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31th Annual Conference on Neural Information Processing Systems*, pages 5998–6008.

Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2015. Pointer networks. In *Proceedings of the 29th Annual Conference on Neural Information Processing Systems*, pages 2692–2700.

Sheng Xu, Tishuang Wang, Peifeng Li, and Qiaoming Zhu. 2019. Multi-layer attention network based Chinese implicit discourse relation recognition. *Journal of Chinese Information Processing*, 33(8):12–19.

Yi Zhou, Xiaomin Chu, Peifeng Li, and Qiaoming Zhu. 2019. Constructing Chinese macro discourse tree via multiple views and word pair similarity. In *Proceedings of the 8th CCF International Conference on Natural Language Processing and Chinese Computing*, pages 773–786.

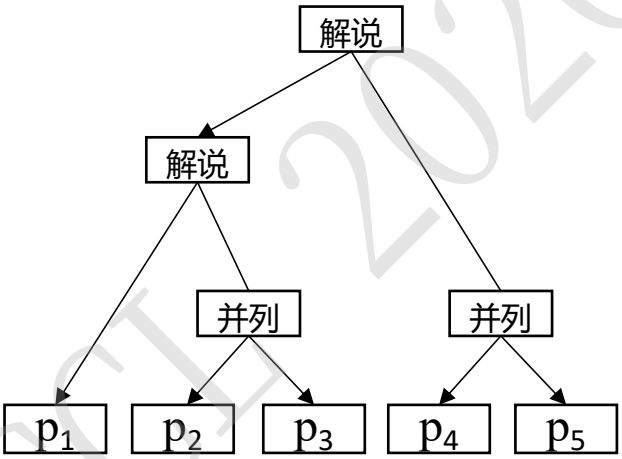
CCL 2020

(c) 2020 中国中文信息学会计算语言学专业委员会

A chtb_0756文章内容及标准结构树

p_1:由于当天公布的一份报告表明美国消费者对经济前景具有信心，纽约股市２９日全面走高。道－琼斯３０种工业股票平均价格指数上升９４．２３点，收于９３２０．９８点，增幅达百分之一。
p_2:道－琼斯指数在过去８个交易日里连续上升。到目前为止，该指数已比今年初上涨了百分之十七点九，比１１月２３日创造的最高记录９３７４．２７点也只有５３点之遥。
p_3与此同时，标准普尔５００种股票指数和以技术股为主的纳斯达克指数２９日均创下了最高纪录。标准普尔指数上升了６．３２点，收于１２４１．８１点。纳斯达克指数则上升了１．４７点，收于２１８１．７７点。此外，纽约证券交易所和美国证券交易所指数以及以小公司为主的罗斯２０００股票指数都告上升。
p_4:在当日的交易中，上涨股票以零售业为主。而前几个交易日紧俏的因特网股则因获利回吐而下跌。
p_5:当天，纽约证交所的上升股与下跌股之比为７比５，成交额从前一交易日的５．２６亿股微升到５．８２亿股。（完）

纽约股市全面上涨（chtb_0756）



宏观篇章结构树（chtb_0756）

CCL 2020

(c) 2020 中国中文信息学会计算语言学专业委员会