

文章编号: 1003-0077(2020)12-0030-09

基于篇章主题的中文宏观篇章主次关系识别方法

孙振华, 周懿, 朱巧明, 蒋峰, 李培峰

(苏州大学 计算机科学与技术学院, 江苏 苏州 215006)

摘要: 篇章分析是自然语言处理领域研究的热点和重点。作为篇章分析的任务之一, 篇章主次关系研究篇章的主要和次要内容, 从而更好地理解 and 把握篇章的核心内容。该文重点研究宏观领域的中文篇章主次关系, 提出了一种基于篇章主题的中文宏观篇章主次关系识别方法。该方法利用篇章单元间、篇章单元与篇章主题间的语义交互来识别主次关系, 并有选择地应用篇章主题信息, 有效提高了主次关系核心的识别。在中文宏观汉语篇章树库(MCDTB)上的实验结果显示, 该方法优于目前性能最好的基准系统。

关键词: 篇章分析; 宏观篇章主次识别; BERT; 篇章主题

中图分类号: TP391

文献标识码: A

Recognizing Chinese Macro Discourse Nuclearity Based on Discourse Topic

SUN Zhenhua, ZHOU Yi, ZHU Qiaoming, JIANG Feng, LI Peifeng

(School of Computer Science and Technology, Soochow University, Suzhou, Jiangsu 215006, China)

Abstract: Discourse analysis is a hot topic in the field of Natural Language Processing. Discourse nuclearity recognition, a subtask of discourse analysis, focuses on recognizing the main and secondary content of a discourse, to better understand and grasp its core content. This paper focuses on the task of macro Chinese discourse nuclearity recognition and proposes a recognition method based on discourse topic. This method introduces the semantic interaction between different discourse units and that between the discourse unit and its topic to identify the nuclearity. Moreover, it applies the selection mechanism of the discourse topic to further improve the performance of nuclearity recognition. Experimental results on MCDTB show that the proposed method outperforms the state-of-the-art baselines.

Keywords: discourse parsing; macro discourse nuclearity recognition; BERT; discourse topic

0 引言

当前, 自然语言处理的研究内容正从传统的词汇、句法分析、语义角色标注等浅层语义分析领域深入到了基于深层语义理解的篇章分析领域。篇章分析是自然语言处理研究的热点和重点, 旨在识别篇章单元(句子、段落、章节或篇章)间的语义关系, 确定篇章结构, 从而挖掘出自然语言文本的结构信息和语义信息^[1]。

篇章分析由篇章结构树构建、主次识别、关系识别三个子任务构成。其中, 篇章主次识别研究旨在分析篇章的主要内容和次要内容, 进而理解篇章主题思想、展开思路 and 核心内容^[2]。篇章主要内容是

指篇章中具有支配地位、起决定性作用的部分, 而次要内容是指篇章中居于辅助地位、不起决定作用的部分。修辞结构理论(rhetorical structure theory, RST)^[3]将篇章主次关系分为单核关系和多核关系。单核关系拥有核心(nucleus)和卫星(satellite), 其中核心表示主要信息, 卫星为其提供附加信息; 多核关系拥有两个及其以上核心。

从文本颗粒度上来分, 篇章主次关系分为微观和宏观两个层面。在微观角度, 篇章主次关系表现为句子之间、句群之间的主要和次要关系; 在宏观角度, 篇章主次关系表现为段落之间、章节之间的主要和次要关系。结合 RST 理论, 篇章主次关系有以下三种类型: ①核心—卫星(NS), 即主要在前, 次要在后; ②卫星—核心(SN), 即主要在后, 次要在前;

收稿日期: 2019-09-11 定稿日期: 2019-10-19

基金项目: 国家自然科学基金(61836007, 61773276, 61772354); 江苏省高校优势学科建设工程资助项目

③核心—核心(NN),即同等重要。其中 NS、SN 属于单核关系,NN 属于多核关系。

本文的研究内容是宏观篇章主次关系的识别。宏观篇章主次关系识别研究不仅能够帮助分析篇章主题,而且也利于理解和挖掘篇章宏观主题与篇章各部分之间的语义联系和脉络结构,可以广泛应用于自然语言处理中的其他任务,包括问答系统^[4]、自动文摘^[5]、情感分析^[6]以及信息抽取^[7]等。本文以宏观汉语篇章树库(macro Chinese discourse tree-bank, MCDTB)^[8]中的一个例子(chtb_0019,见附录)来说明宏观篇章主次关系,该例宏观篇章关系结构如图 1 所示。

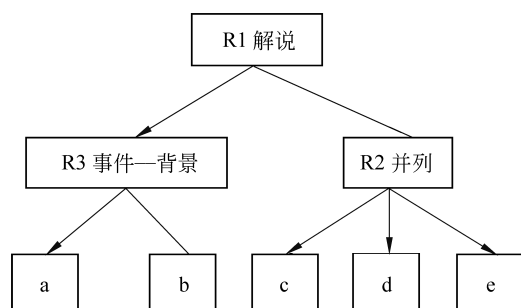


图 1 宏观篇章关系结构(chtb_0019)

其中,叶子节点(a~e)是基本篇章单元(elementary discourse units, EDUs),分支节点是篇章单元(discourse units, DUs),也是关系节点,表示该节点连接的两个子节点之间的语义关系。箭头指向的节点为核心,否则为卫星。在该例中,段落 a 提出“宁波保税区经过三年建设已取得丰硕成果”这一事实,段落 b 讲述了宁波保税区的基本情况,段落 c、d、e 分别从进出口贸易、实行政策以及企业管理机制三个方面阐述了宁波保税区取得显著成就的原因。因此,段落 a 与 b 构成事件—背景关系,段落 a 为主要内容;段落 c、d、e 从不同角度阐述原因,三者同等重要,构成并列关系。

在篇章分析上已经有很多研究,然而只有少数研究专注于篇章主次关系识别。面向汉语的篇章主次识别,尤其是宏观篇章主次关系识别的研究更少^[9]。随着 MCDTB 构建完成,语料资源匮乏问题得到了极大的缓解,但是由于宏观篇章构成复杂,基本篇章单元(自然段落)长度长,段落内部表达形式多样,段间关系复杂,给宏观篇章主次关系识别造成了一定困难。

在汉语微观篇章主次识别任务上,Xu 等^[10]提出的 TMN(text matching network)神经网络模型

获得了较好的性能。目前,还没有神经网络方法应用于汉语宏观篇章主次关系识别任务上。由于微观和宏观主次关系识别从某种程度上来讲存在相似性,本文将 TMN 模型迁移到了宏观主次识别任务上。然而,TMN 模型采用 BiLSTM 和 CNN 来编码语义,将其应用到宏观领域会面临以下三个问题。

(1) 宏观篇章单元比微观篇章单元更长,整体语义建模难度更大。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时,仍然力不从心。

(2) 词序列长度的增加导致篇章单元中出现更为复杂的词间依赖与交互模式。LSTM 主要是随着时间推移来顺序处理文本,当文本中相距较远的词之间存在语义依赖时,LSTM 和 CNN 都无法捕获这种信息。

(3) BiLSTM 无法做到深度双向编码语义。BiLSTM 虽然以从左到右和从右到左两个方向来编码语义,但是在某一时刻,BiLSTM 编码输入时,仅能考虑当前词的上文信息或者下文信息,无法在编码时同时利用上文和下文信息。这个缺点在建模更富有语义的宏观篇章单元时,将丢失大量的语义特征。

为了解决以上问题,本文首次将 BERT(bidirectional encoder representations from transformers)^[11]应用于汉语宏观篇章主次关系识别任务。TMN 模型认为:①语义相似度高的两个篇章单元更有可能是多核关系;②在语义上更靠近段落主题的则更有可能是核心。受该思想的启发,本文在宏观篇章主次关系识别中利用篇章主题与篇章单元之间的语义交互来识别篇章单元间的主次关系。TMN 先学习篇章单元及其所在段落的句子级表示,然后再计算三者之间的语义相似度和语义交互。本文从词间依赖入手,通过捕获句内词间依赖来学习句子级的语义表示;通过捕获跨句词间依赖来学习句子间的语义交互,从而能够充分捕获篇章单元间、篇章单元与篇章主题间的交互信息和依赖信息。同时,在应用篇章主题上,本文提出“选择性应用篇章主题”的方法,只对含有篇章第一自然段的样本应用篇章主题。在 MCDTB 语料库上的实验结果表明,本文的方法解决了篇章单元过长、难以建模的问题。通过选择性引入篇章主题,仅仅依赖语义信息,就在微平均 F_1 值、宏平均 F_1 值以及各个主次关系类型的 F_1 值上均有所提升,尤其是识别出了 SN 这一小类样本。

1 相关工作

在英文方面,涉及到篇章主次关系的语料库资源主要有修辞结构篇章树库(RST discourse tree-bank, RST-DT)^[12]。中文方面主要有汉语篇章树库 CDTB^[13]和宏观汉语篇章树库 MCDTB^[8]等。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。汉语篇章树库(CDTB)依据连接依存树的篇章结构理论,在宾州大学汉语树库(CTB)上标注了 500 篇微观篇章结构。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、主次关系、语义关系等。这三个语料库中,RST-DT 和 CDTB 进行了微观篇章主次关系的标注,MCDTB 进行了宏观篇章主次关系的标注。

目前,篇章主次关系识别研究大多属于微观篇章分析。在 RST-DT 上,Hernault 等^[14]提出了基于支持向量机的篇章分析器 HILDA,该模型以贪婪的方式自底向上自动构建篇章结构树;Joty 等^[15]利用动态随机场模型分别构建了句内和句间两个层级的篇章分析器,并使用动态规划算法对篇章树构建进行了优化;Li 等^[16]基于篇章单元的子树结构,使用递归神经网络来获取篇章单元表示,并构建了两层前馈神经网络来识别两个篇章单元之间的关系;Li 等^[17]立足于篇章的分层结构,提出了基于 Attention 的分层 BiLSTM 网络,从词序列中学习基本篇章单元(EDU)的表示,依据学到的 EDU 表示来建模篇章单元(DU),并使用基于张量的变换函数来捕获篇章单元特征之间的相互关系。在 CDTB 上,Chu 等^[18]使用上下文、词对、词和词性等特征来识别篇章主次关系。Kong 等^[19]使用语义相似度和上下文特征,采用最大熵模型构建了一个端到端的篇章结构分析器;Xu 等^[10]提出 TMN 模型,使用 BiLSTM+CNN 来对篇章单元及其段落进行编码,通过计算语义相似度和语义交互进行主次关系识别。

宏观篇章分析研究极少。在 MCDTB 上,蒋等^[9]提出融合基于 Word2Vec 和 LDA 的主题相似度,并利用最大熵分类器进行宏观篇章主次关系识别。该模型通过最大化词间相似度和主题相似度来获取语义特征,并结合部分组织结构特征来识别主次关系。然而,这种捕获语义特征的方法,不仅丢失

了篇章单元的连贯性,而且也无法充分捕获篇章单元的语义信息。

2 NRDT 模型

本文提出了基于篇章主题的中文宏观篇章主次识别模型 NRDT(nuclearity recognition based on discourse topic),其架构如图 2 所示。该模型包含三个部分:①输入模块、②编码模块、③主次识别模块。

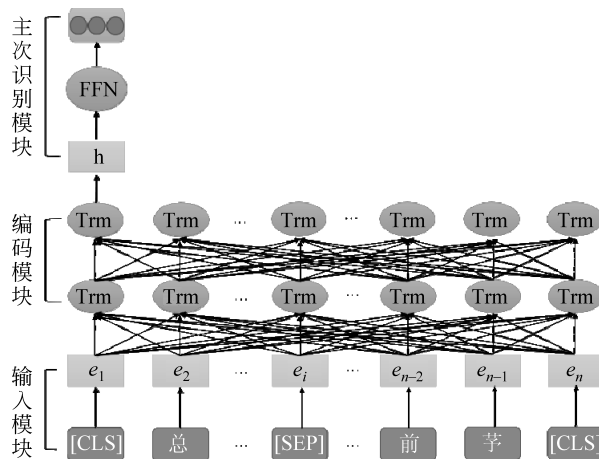


图 2 模型架构

2.1 输入模块

篇章主次关系中,大多数是像图 1 中 R1、R3 这样的二元关系,也存在像 R2 这样的多元主次关系。本文遵循了蒋等^[9]对多元关系的处理方法,将多元主次左二义化。因此问题转换成句子对的三分类任务。即:给定篇章单元 Arg1 与 Arg2,判断两者之间的主次关系(NS、SN 或 NN)。样本数据形式化为三元组(Arg1, Arg2, label),label 表示 Arg1 与 Arg2 之间的主次关系。考虑到篇章单元与篇章主题之间的语义交互可以在更高层次上表现出该单元所蕴含的语义信息,本文将篇章标题作为篇章主题 topic,与篇章单元一起作为输入。本文用 Arg 来表示篇章单元,用 Topic 来表示篇章主题。

由于编码层使用 BERT 模型来编码输入篇章单元,其在处理篇章单元对时,将篇章单元对整合成一个序列,并用特殊标记[SEP]来区分篇章单元 1 和篇章单元 2:

[CLS]...Arg1...[SEP]...Arg2...[SEP]

每个输入序列的开始词是一个特殊分类标记[CLS],该标记对应的最终隐向量融合了输入序列

的语义表示。考虑本文的输入由三部分构成,本文模型的输入最终可形式化为:

[CLS]...Topic...Arg1...[SEP]...Arg2...[SEP]

为了适应编码层对输入的要求,避免输入对预训练好的 BERT 模型产生干扰,本文并没有在 Topic 与 Arg1 之间添加[SEP]标记。本文用“Arg1 前”“Arg1 后”“Arg2 前”以及“Arg2 后”来表示 Topic 添加的四种位置,具体如下:

Arg1 前: [CLS]... Topic ... Arg1 ... [SEP] ... Arg2...[SEP]

Arg1 后: [CLS]... Arg1 ... Topic ... [SEP] ... Arg2...[SEP]

Arg2 前: [CLS]... Arg1 ... [SEP] ... Topic ... Arg2...[SEP]

Arg2 后: [CLS]... Arg1 ... [SEP] ... Arg2 ... Topic...[SEP]

实验发现,将 Topic 放在 Arg1 前性能最好。其余三个位置在微平均 F_1 上要低约 0.5 个百分点。

本文对比了三种将篇章主题 Topic 加入到输入的策略。第一种是对所有的篇章单元输入都对都加入篇章主题。样本形式化为:

(Topic, Arg1, Arg2, label)

第二种是当篇章单元 Arg1 包含篇章第一自然段时,加入篇章主题,样本形式同上,否则不加篇章主题,样本形式化为:

(Arg1, Arg2, label)

第三种是第二种策略的反面。即:当篇章单元 Arg1 不包含篇章第一自然段时,添加篇章主题。与第二种策略形成正反对比。

第二种添加策略基于如下考虑:

(1) 人类在辨别篇章单元间的主次关系时,并不会一直参考篇章主题。有一些篇章单元对,仅仅依赖自身语义就可识别出主次。篇章主题的引入反而可能会对判别造成干扰,使得单核关系中的核心识别错误。

(2) 篇章一般采用“总分总”或“总分”结构。尤其在新闻体裁中,更是如此。篇章的开始段落更有可能总领全文,归属于核心。然而开始段落只有下文信息,缺乏上文信息,篇章主题在一定程度上可以弥补这一点,增强对篇章开始段落的核心识别。

(3) 篇章中的后续篇章单元通常是从不同角度为篇章主题服务,多属于 NN,两者语义上相似度较高,可以仅仅依赖自身语义就可识别出来。

本文模型的输入向量与 Vanilla BERT 一致,由

三部分构成。本文用 w 表示词向量, s 表示句向量, p 表示位置向量,则对于第 i 个词来说,其向量表示,如式(1)所示。

$$e_i = w_i + s_i + p_i \quad (1)$$

对于每一个构造好的输入序列 q ,模型的最终输入是该序列中 n 个词的向量表示,如式(2)所示。

$$q = [e_1, e_2, e_3, \dots, e_n] \in \mathbb{R}^{n \times d} \quad (2)$$

其中, d 表示每一个输入向量的维度。

2.2 编码模块

编码层的主体是 BERT,由多层 Transformer Encoder^[20]构成。该模型在大规模语料库上进行遮蔽语言模型(masked language model, MLM)和下一句预测(next sentence prediction, NSP)联合训练。MLM 任务同时学习被遮蔽掉词的左右两边信息来预测出该词。通过这个类似于完形填空的任务, BERT 能够充分学习词语的语义表征,捕获词间语义交互。NSP 即输入两个句子,让模型来判断第 2 句是否是第 1 句的下一句。通过这个简单任务,模型能够学会综合考虑词语间语义交互,从而计算句子的语义表征。BERT 的输入由三部分构成,分别是词向量、句子向量以及位置向量。输入的第一个词是特殊分类标记[CLS],最后一个词是特殊分类标记[SEP]。两个句子中间用一个[SEP]分开。[SEP]标记和句子向量作用类似,都是帮助模型区分输入中的两个句子。给定词序列输入 $X = \{x_1, x_2, \dots, x_n\}$, BERT 的输出为 $H = \{h_1, h_2, \dots, h_n\}$ 。

本文采用预训练好的 BERT-base 模型做编码层。该模型由 12 个 Transformer encoder 块堆叠而成,每一块包含 12 个自注意力头,768 个隐层神经元。模型的核心由多头自注意力模块和两层前馈神经网络模块构成。多头自注意力模块计算方法如式(3)~式(5)所示。

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (3)$$

$$\text{MultiHead}(Q, K, V) = [\text{head}_1, \dots, \text{head}_h]W^O \quad (4)$$

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \quad (5)$$

其中, $[\dots]$ 代表级联, $Q = K = V \in \mathbb{R}^{n \times d}$, $W_i^Q, W_i^K, W_i^V \in \mathbb{R}^{d \times d_k}$, $W^O \in \mathbb{R}^{hd_k \times d}$, $d_k = d/h$, $h = 12$ 。

前馈神经网络模块进行计算如式(6)所示。

$$\text{FFN}(x) = \text{relu}(xW_1 + b_1)W_2 + b_2 \quad (6)$$

其中, $W_1 \in \mathbb{R}^{d \times d_{ff}}$, $W_2 \in \mathbb{R}^{d_{ff} \times d}$, $b_1 \in \mathbb{R}^{d_{ff}}$, $b_2 \in \mathbb{R}^d$ 为要训练的参数。 $x \in \mathbb{R}^d$, $d_{ff} = 4d$ 。

自注意力模块使用 12 个独立的注意力头,在不

同的特征子空间中计算任意词对之间的点积。这使得每一个词可以从不同的角度,与其他词发生交互,进行语义解析,从而对复杂的词间依赖进行建模。自注意力将词间距离视为 1,直接计算任意词对间的依赖关系,能够学习句子的内部结构,从而捕获长序列的局部信息。

级联所有注意力头的输出,输入到前馈神经网络模块。前馈神经网络模块具有复杂非线性能力,能综合考虑注意力模块捕获到的不同角度的词间依赖信息和语义交互信息,并对这些信息进行语义融合,从而捕获序列的全局信息。

2.3 主次识别模块

根据 Arg_1 、 Arg_2 以及 Topic,通过编码模块,主次识别模块不仅获得了篇章单元间的语义交互信息,也获得了篇章单元与篇章主题间的语义交互信息。同时也得到了篇章单元与篇章主题各自的语义信息。本文以第一个词[CLS]对应的最终隐向量 $\mathbf{h} \in \mathbb{R}^d$ 作为模型计算的表示,然后通过两层前馈神经网络,实现篇章主次关系识别,如式(7)~式(10)所示,其中 $\mathbf{W}_3 \in \mathbb{R}^{d \times d}$ 、 $\mathbf{W}_4 \in \mathbb{R}^{d \times 3}$ 、 $\mathbf{b}_3 \in \mathbb{R}^d$ 、 $\mathbf{b}_4 \in \mathbb{R}^3$ 为参数矩阵:

$$\mathbf{u} = \tanh(\mathbf{W}_3^T \mathbf{h} + \mathbf{b}_3) \quad (7)$$

$$\mathbf{z} = \text{dropout}(\mathbf{u}) \quad (8)$$

$$\mathbf{y} = \text{softmax}(\mathbf{W}_4^T \mathbf{z} + \mathbf{b}_4) \quad (9)$$

$$p = \arg\max \mathbf{y} \quad (10)$$

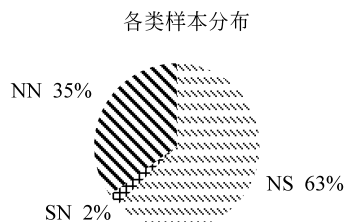
向量 \mathbf{y} 表示模型预测的概率分布。本文取概率最大的类别 p 作为样本的主次关系预测。

3 实验

3.1 实验设置

本文使用了 MCDTB 语料库,每条数据包含两个篇章单元、篇章主题、篇章单元对应的组织结构特征以及主次关系标签。MCDTB 语料库标注了 720 篇新闻语料,由于宏观篇章主次识别以段落为基本篇章单元,因此数据样本仍然不足。考虑到小样本训练集的不稳定性,实验采用了五折交叉验证。图 3 显示了一折数据 segdata0 训练集的各类样本分布。

本文实验采用深度学习框架 Pytorch。模型的预测层使用了 Dropout,比例为 0.1。对学习率,使用了线性热身,比例为 0.1。实验参数如表 1 所示。



NS 主要在前, SN 主要在后, NN 同等重要

图 3 segdata0 训练集各类样本分布

表 1 模型参数设置

参数名	参数值
Bert-base 参数	默认值
隐藏层维度	768
学习率	2e-5
学习率热身比例	0.1
Dropout 比例	0.1
训练轮数	10
批处理大小	16

实验采用多类交叉熵损失函数,如式(11)所示。

$$L = -\frac{1}{N} \sum_{i=1}^N y_i \log(\hat{y}_i) \quad (11)$$

其中, N 代表样本总数, y_i 为指示变量(0 或 1),若该类别与样本 i 的类别相同就是 1,否则为 0。 \hat{y}_i 为样本 i 属于当前类别的预测概率。在训练过程中,使用 Adam 优化器来优化模型参数,通过最小化交叉熵损失,对模型进行端到端的 fine-tuning (Bert 参数随着训练进行 fine-tuning)。

本文采用精确率(precision)、召回率(recall)、微平均 micro_F_1 和宏平均 macro_F_1 来衡量模型的分类质量。精确率 P 、召回率 R 、 F_1 值的定义如式(12)~式(14)所示。

$$P = \frac{TP}{TP + FP} \quad (12)$$

$$R = \frac{TP}{TP + FN} \quad (13)$$

$$F_1 = \frac{2PR}{P + R} \quad (14)$$

其中, TP 表示正例被正确预测的个数, FP 表示负例被错误预测的个数, FN 表示正例被错误预测的个数。在实验评估中,采用 sklearn 工具^①进行

① <https://scikit-learn.org/>

上述指标的计算。

3.2 实验结果

3.2.1 模型

本文将文中提出的模型 NRDT 与基准系统进行对比。基准系统分为两种：①只用语义信息；②语义信息和结构信息相结合。

(1) **ME**: 性能最好的传统模型^[9],该模型以基于 Word2Vec 的主题相似度、基于 LDA 的主题相似度以及组织结构为特征,使用最大熵分类器。为了公平对比,本文复现了蒋等^[9]提出的 ME 模型,并在五折数据集上进行了相关实验。

(2) **BiLSTM+Att**: 在宏观篇章主次关系研究中,目前还没有相应的神经网络模型可作为基准系统。因此本文构造了 BiLSTM+Attention 模型作为本文的第一个神经网络基准系统。使用 BiLSTM 分别对 Arg₁、Arg₂ 和 Topic 建模,经过 Attention 层之后,级联得到的三个语义表示,进行分类。

(3) **TMN**: 本文复现了微观篇章主次关系识别中性能优异的 TMN 模型^[10]。

(4) **BERT_{base}**: 使用 BERT 对两个篇章单元进行联合语义交互编码。

(5) **NRDT**: 即本文提出的模型,使用篇章单元信息以及篇章主题信息,不使用结构信息。使用 BERT 来编码篇章单元间、篇章单元与篇章主题间的语义交互。

(6) **NRDT+stru**: 可选模型,在本文提出模型

的基础上,添加 MLP 编码后的结构特征。此处使用了五种组织结构特征(篇章单元的开始段落、结束段落、篇章单元到篇章开始段落和结束段落的距离以及篇章单元包含的段落数)。本文使用 v_1, v_2, \dots, v_5 分别表示这五种结构特征,将其级联,经过 MLP 编码后,得到 v 。在该模型中,本文对式(8)进行了修改,从而融入了结构信息, $[\dots]$ 表示级联,如式(15)所示。

$$z = [\text{dropout}(u), v] \tag{15}$$

3.2.2 实验结果分析

实验结果如表 2 和表 3 所示。无论是“只使用语义信息”还是“语义和结构相结合”,本文提出的模型 NRDT 都取得了最好性能,相比基准系统在宏平均 F_1 和微平均 F_1 均得到了较大提升。当“只使用语义信息”时, NRDT 模型在宏平均上达到了 59.59%,微平均达到了 84.06%。当“语义和结构联合”时, NRDT 模型在宏平均 F_1 上达到了 58.34%,微平均 F_1 达到了 84.15%。

(1) 只使用语义信息

如表 3 所示,在“只使用语义信息”的情况下, NRDT 比 TMN 和 BiLSTM+Att 在宏平均和微平均 F_1 值上均获得了较大提升。这是因为宏观篇章单元包含的词相对微观领域更多,而且序列长度更长,词间依赖更加错综复杂。无论是使用 BiLSTM 还是 BiLSTM+CNN 做语义编码,它们都无法充分捕获篇章单元的全局信息和复杂的词间依赖等局部信息。TMN 模型则在语义匹配时没有考虑跨句的

表 2 不同模型的性能比较(语义+结构)

模型	主在前(NS)			主在后(SN)			同等重要(NN)			Macro- F_1	Micro- F_1
	P	R	F_1	P	R	F_1	P	R	F_1		
ME	92.41	81.79	86.78	0	0	0	71.90	91.28	80.43	55.73	83.15
TMN	82.10	88.21	85.04	0	0	0	74.87	70.79	72.77	52.60	79.84
BiLSTM+Att	90.96	83.37	87.00	0	0	0	72.97	88.69	80.06	55.69	83.21
NRDT+stru	89.81	86.63	88.18	15.71	4.06	6.37	76.23	85.28	80.46	58.34	84.15

表 3 不同模型的性能比较(语义)

模型	主在前(NS)			主在后(SN)			同等重要(NN)			Macro- F_1	Micro- F_1
	P	R	F_1	P	R	F_1	P	R	F_1		
BiLSTM+Att	73.42	78.41	75.79	0	0	0	56.91	53.87	55.23	43.67	67.83
TMN	75.57	84.87	79.95	0	0	0	61.63	52.48	56.68	45.54	71.64
NRDT	89.36	87.31	88.31	36.97	6.58	10.43	76.76	83.68	80.03	59.59	84.06

词间依赖。NRDT 直接计算任意词之间的语义交互,能够充分捕获句内和跨句词间依赖,挖掘语义信息。NRDT 仅仅通过语义特征,获得了比其他基准系统更好的性能。这在一定程度上说明,句子间语义交互与跨句词间依赖关系有很大联系,同时从词间依赖入手有助于对更长的序列进行语义建模。

(2) 语义和结构相结合

在宏观主次识别上,目前性能最好的模型是蒋等^[9]的 ME 模型。该模型采用“语义和结构相结合”的特征来做主次识别。为了公平比较,本文对 TMN、BiLSTM+Att 以及 NRDT 在原有的语义基础上,添加了结构信息。如表 2 所示,在添加结构信息之后,NRDT 模型在宏平均 F_1 和微平均 F_1 上仍取得了最好性能。

BiLSTM+Att、TMN 以及 NRDT 在添加结构信息之后,微平均 F_1 上分别获得了 15.38%、8.20%、0.09% 的提升。前文提到,LSTM 模型无法对更长的文本序列进行良好的语义建模,因此前两

个模型受限于各自的长距离依赖编码能力,无法充分挖掘宏观篇章单元的语义信息。结构信息作为一种强特征,弥补了这一点,因而 BiLSTM+Att 和 TMN 这两个模型获得的提升较大。NRDT 模型能够充分提取输入文本的语义信息,仅仅依靠语义就获得了很好的性能,结构信息的加入带来的提升略小,而且对 SN 的识别造成了一定的影响。这可能是因为 SN 样本太少,其结构信息不具有代表性,该特征的引入带来的噪声较大。

(3) 篇章主题对主次识别的作用以及篇章主题添加策略分析

如表 4 所示,对比 BERT_{base} 模型和 NRDT 模型的性能,可以看出篇章主题对篇章主次关系识别有很大帮助,微平均 F_1 提高了约 4.49 个百分点,宏平均 F_1 提高了约 3.43 个百分点,在 NS、SN 以及 NN 上都有提升。这很大程度上证明了本文所依据假设的合理性,篇章主题与篇章单元间的语义交互有助于主次关系中核心的识别。

表 4 篇章主题对主次识别的作用分析

模型	主在前(NS)			主在后(SN)			同等重要(NN)			Macro- F_1	Micro- F_1
	P	R	F_1	P	R	F_1	P	R	F_1		
BERT _{base}	83.49	86.55	84.98	23.98	7.81	10.34	74.14	72.30	73.15	56.16	79.57
NRDT _{策略 1}	86.44	86.62	86.47	27.28	12.87	14.34	74.99	76.93	75.80	58.87	81.38
NRDT _{策略 2}	89.36	87.31	88.31	36.97	6.58	10.43	76.76	83.68	80.03	59.59	84.06
NRDT _{策略 3}	83.51	86.99	85.20	38.38	5.83	10.04	73.74	72.59	73.11	56.11	79.88

本文也对在模型中添加篇章主题 Topic 的三种策略做了对比实验,在表 4 中策略 1、策略 2、策略 3 分别代表 2.1 节中提到的三种策略。比较这三种策略,可以看到策略 2 的性能最好,即“篇章单元包含第一自然段时添加篇章主题,否则不添加”这一策略性能最好。策略 2 与策略 1 的对比说明,选择性添加篇章主题要比对所有篇章单元都添加篇章主题的性能要好。策略 2 与策略 3 的对比印证了本文之前的考虑,篇章主题的引入会对篇章非开始段落的篇章单元对造成干扰,而这些篇章单元对仅仅通过自身语义就能识别出来。策略 2 的性能最好,可能是因为篇章主题一定程度上充当了 Arg1(包含第一段)的上文信息,从而提高了含有第一段的 Arg1 为核心的识别率。

(4) NRDT 性能分析

本文提出的 NRDT 相比其他基准系统,在三种篇章主次类型上均获得了最好的性能,尤其是识别出了 SN 这一小类样本。基准系统受限于自身语义

特征提取能力弱,又面临着 SN 样本数量较少的问题,导致模型无法学习到 SN 样本应有的特征。本文模型能够捕获跨句词间的依赖,充分计算篇章单元与篇章主题间的语义交互,语义编码能力强,因而能够从少量的 SN 样本中学习到的较为有力的特征,准确识别出其中的核心。

但是该模型在三种篇章类型上的表现并不相同。如表 3 所示,各类别上的性能表现并不均衡。对于 NS 和 NN 类型,由于这两类样本数量较多,模型识别效果较好,通过表 5 的混淆矩阵可以看出,这两类各自有个别样本被误分类为 SN,绝大多数分类错误发生在 NS 和 NN 之间的混淆,这可能是篇章主题的引入导致。如例 2 所示,相比于 Arg₂,Topic 与 Arg₁ 有更多的词相似,模型容易将该 NN 样本识别成 NS。对于 SN 类型,该类样本数量太少,虽然模型特征提取能力强,但也仅识别出少量样本。SN 样本多半为因果或评价关系,宏观篇章单元间没有显式连接词,篇章单元与篇章主题间的语

义交互又难以区分此类语义关系。

表 5 NRDT 实验结果的混淆矩阵(%)

预测值	真实值		
	NS	SN	NN
NS	87.31	0.29	12.39
SN	39.37	6.87	53.75
NN	15.54	0.78	83.68

例 2 识别错误样本

Arg1: 国家建设部将在全国选择四十家大型建筑施工企业给予重点支持。
Arg2: 为此,建设部将采取七项扶植政策。
Topic: 中国将重点扶持四十家大型建筑施工企业。

表 6 展示了 NRDT 模型错误识别的主次关系在篇章位置中的分布。本文将篇章按照段落均分为三部分,并统计了各个位置上三种主次关系识别错误的分布。可以看出错误大多发生在篇章开始部分为 NN、中间部分以及结尾部分为 NS 这两种。这可能与本文采用的应用篇章主题的策略、篇章复杂的行文结构以及人工标注时的主观判断有关。

表 6 NRDT 错误识别样本的位置分布(%)

位置	主次		
	NS	SN	NN
开始	40.58	17.39	42.03
中间	56.70	12.47	30.82
结尾	58.64	9.02	32.33

4 总结与展望

本文提出的基于篇章主题的中文宏观篇章主次关系识别模型 NRDT 是 MCDTB 上的第一个神经网络模型。该模型通过对篇章单元间、篇章单元与篇章主题间的语义交互和信息依赖进行建模,从而识别主次关系,实验结果在宏平均 F_1 、微平均 F_1 上均有提高。这说明仅仅依赖语义信息,就可以很好地完成主次识别任务。本文提出的模型,通过对词间依赖进行建模,进而捕获句间语义交互,这在一定程度上说明,句子间语义交互与跨句词间依赖有很大联系,同时从词间依赖入手有助于对更长的序列进行语义建模。有选择地加入篇章主题信息后,模型性能得到了进一步的提高,这说明篇章主题对篇章单元间的主次识别很有帮助。本文将篇章标题作为篇章主题,但有些篇章标题并不能充分地代表篇

章主题。因此,在下一步工作中,将尝试构建主题模型来学习篇章全局和局部主题,以进一步提高性能。篇章语义关系和主次关系有很大的相关性。构成并列关系的篇章单元对之间往往是多核关系;构成因果关系的篇章单元对之间往往是单核关系。接下来会尝试基于以自注意力机制为核心的模型对这两个任务进行联合学习,例如,Transformer 及其相关变体。

参考文献

- [1] 徐凡,朱巧明,周国栋. 篇章分析技术综述[J]. 中文信息学报, 2013, 27(3): 20-33.
- [2] 褚晓敏,朱巧明,周国栋. 自然语言处理中的篇章主次关系研究[J]. 计算机学报, 2017, 40(4): 842-860.
- [3] Mann W C, Thompson S A. Rhetorical structure theory: Toward a functional theory of text organization[J]. Text-Interdisciplinary Journal for the Study of Discourse, 1988, 8(3): 243-281.
- [4] Liakata M, Dobnik S, Saha S, et al. A discourse-driven content model for summarising scientific articles evaluated in a complex question answering task[C]// Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, 2013: 747-757.
- [5] Cohan A, Goharian N. Scientific article summarization using citation-context and article's discourse structure [C]//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal: Association for Computational Linguistics, 2015: 390-400.
- [6] Zhou L, Li B, Gao W, et al. Unsupervised discovery of discourse relations for eliminating intra-sentence polarity ambiguities[C]//Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2011: 162-171.
- [7] Zou B, Zhou G, Zhu Q. Negation focus identification with contextual discourse information[C]//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, 2014: 522-530.
- [8] Jiang F, Xu S, Chu X, et al. MCDTB: A macro-level Chinese Discourse TreeBank[C]//Proceedings of the 27th International Conference on Computational Linguistics, 2018: 3493-3504.
- [9] 蒋峰,褚晓敏,徐昇. 基于主题相似度的宏观篇章主次关系识别方法[J]. 中文信息学报, 2018, 32(1): 43-50.
- [10] Xu S, Li P, Zhou G, et al. Employing text matching network to recognise nuclearity in Chinese discourse [C]//Proceedings of the 27th International Conference on Computational Linguistics, 2018: 525-535.

- [11] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 29th Conference of the North American Chapter of the Association of Computational Linguistics, 2018: 4171-4186.
- [12] Carlson L, Okurowski M E, Marcu D. RST discourse treebank[M]. Linguistic Data Consortium, University of Pennsylvania, 2002.
- [13] Li Y, Kong F, Zhou G. Building Chinese discourse corpus with connective-driven dependency tree structure[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, 2014: 2105-2114.
- [14] Hernault H, Prendinger H, Ishizuka M. HILDA: A discourse parser using support vector machine classification[J]. Dialogue & Discourse, 2010, 1(3): 1-33.
- [15] Joty S, Carenini G, Ng R, et al. Combining intra-and multi-sentential rhetorical parsing for document-level discourse analysis[C]//Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, 2013, 1: 486-496.
- [16] Li J, Li R, Hovy E. Recursive deep models for discourse parsing[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, 2014: 2061-2069.
- [17] Li Q, Li T, Chang B. Discourse parsing with attention-based hierarchical neural networks[C]//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. 2016: 362-371.
- [18] Chu X, Wang Z, Zhu Q, et al. Recognizing nuclearity between Chinese discourse units[C]//Proceedings of the 2015 International Conference on Asian Language Processing. IEEE, 2015: 197-200.
- [19] Kong F, Wang H, Zhou G. A CDT-styled end-to-end Chinese discourse parser[M]. Natural Language Understanding and Intelligent Applications. Springer, Cham, 2016: 387-398.
- [20] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Proceedings of Advances in Neural Information Processing Systems, 2017: 5998-6008.

附录

例 1 chth_0019 文章内容

宁波保税区建设成就显著

- (a) 总面积二点三平方公里的宁波保税区,经过三年建设,已取得丰硕成果。
- (b) 宁波保税区是中国十三个保税区之一,于一九九二年经国务院批准设立。目前,保税区的各项功能已初具规模,开发水平在中国各保税区中名列前茅。
- (c) 据统计,至去年年底,宁波保税区累计完成进出口贸易额八点一二亿美元,仅去年一年通过保税区海关的进出口贸易额就达三点六五亿美元。目前,区内已有十个保税仓库,仓储面积达八万多平方米;仅去年一年,区内储有货物就达二十六点二七亿元人民币。
- (d) 随着从今年四月开始中国对保税区外有关特殊政策的调整,保税区免证、免税,保税政策的稳定性优势显得更为明显,国内外一大批实业加工项目相继在区内落户。到去年十二月底,区内已累计设立企业一千六百一十四家,总投资达十二亿美元,其中外商投资企业二百六十家,实际利用外资一点一三亿美元。另外,众多国内企业也通过保税区与国际市场接轨。
- (e) 为了在运行机制上与保税区相配套,宁波保税区率先在中国实施了企业依法注册直接登记制的试行一站式管理,一次性办理。同时,保税区大力抓好区内信息高速公路的网络体系建设,为实现现代化管理创造良好的配套条件。(完)



孙振华(1995—),硕士研究生,主要研究领域为自然语言处理。
E-mail: 20185227016@stu.suda.edu.cn



周懿(1995—),硕士研究生,主要研究领域为自然语言处理。
E-mail: yzhou0928@stu.suda.edu.cn



朱巧明(1963—),通信作者,博士,教授,主要研究领域为自然语言处理。
E-mail: qmzhu@suda.edu.cn