doi:10.3969/j.issn.1001-4616.2026.02.008

融合双语信息的汉语篇章主次识别方法

李艳翠,郭鹏程,苗国义

(1.河南师范大学计算机与信息工程学院,河南 新乡 453007)
(2.教育人工智能与个性化学习河南省重点实验室,河南 新乡 453007)
(3.河南省教学资源与教育质量评估大数据工程研究中心,河南 新乡 453007)

[**摘要**] 在主次识别中,汉语句子间的显式衔接手段较少,因此其主次识别具有极大的挑战性. 英语大多用明确的主从结构或连接词来表示句子的主次关系,而现有方法在训练模型时没有利用英语信息. 与现有方法在训练模型时单独使用中文数据不同,提出的方法在训练模型时使用平行双语数据. 对双语文本编码时使用多语言预训练模型,在得到的编码上应用多头注意力机制,捕获显式或隐含于句中的主从信息. 在汉语篇章树库(Chinese Discourse Treebank,CDTB)上的实验显示,提出的模型和方法比之前最好的 GMN-Nu 模型在宏平均 $F1$ 值和微平均 $F1$ 值上提高了 8.7%和 6.1%;相较于仅使用预训练模型和单语数据训练的方法,融合双语信息的主次识别方法对于 mBERT、mT5、XLM-R 3 种模型在微平均 $F1$ 值上分别提高了 1.6%、3.5%、1.3%. 在汉英篇章结构平行语料库(Chinese-English Discourse Treebank,CEDT)上的实验显示,融合双语信息的主次识别方法比单语言的主次识别方法在微平均 $F1$ 值和宏平均 $F1$ 值上分别提升了 10.2%和 5.8%.

[**关键词**] 篇章分析,主次识别,预训练模型,双语信息
[**中图分类号**]TP391　[**文献标志码**]A　[**文章编号**]1001-4616(2026)02-0074-11

Integration of Bilingual Information for Nuclearity Recognition in Chinese Discourse

Li Yancui, Guo Pengcheng, Miao Guoyi

(1.School of Computer and Information Engineering, Henan Normal University, Xinxiang 453007, China)
(2.Henan Key Laboratory of Educational Artificial Intelligence and Personalized Learning, Xinxiang 453007, China)
(3.Henan Engineering Research Center of Teaching Resources & Education Quality Evaluation Big Data, Xinxiang 453007, China)

Abstract: Chinese nuclearity recognition encounters inherent difficulties owing to limited explicit inter-sentential connectives. In contrast, English systematically marks nuclearity through subordinate constructions and discourse markers. Current approaches trained models exclusively on Chinese corpora without leveraging English signals. Our methodology addresses this gap by incorporating parallel bilingual training data. A multilingual pre-trained model processed the bilingual texts, and a multi-head attention mechanism captured explicit and implicit nuclearity features. Experiments on the Chinese Discourse Treebank(CDTB) showed that our model achieved 8.7% and 6.1% improvements in Macro-$F1$ and Micro-$F1$ scores over the previous state-of-the-art GMN-Nu model. Compared to monolingual training with mBERT, mT5, and XLM-R, the bilingual fusion strategy increased Micro-$F1$ by 1.6%, 3.5% and 1.3%, respectively. Additional tests on the Chinese-English Discourse Treebank(CEDT) demonstrated 10.2% and 5.8% gains in Micro-$F1$ and Macro-$F1$ over monolingual methods.

Key words: discourse analysis, nuclearity recognition, pretrained models, bilingual information

篇章分析是使用机器自动分析自然语言语篇的过程、技术和方法,它研究文本中的篇章结构关系,是自动摘要、机器翻译、问答系统等下游应用的基础. 篇章分析中的修辞结构理论(Rhetorical Structure Theory,RST)[1]认为:一个篇章通常由若干连续的基本篇章单元(Elementary Discourse Unit,EDU)组成,而

收稿日期:2025-03-24.
基金项目:教育部人文社会科学研究项目(22YJCZH091)、河南省科技攻关项目(252102210102、262102210084)、河南省自然科学基金项目(262300421797).
通讯作者:苗国义,博士,研究方向:自然语言处理. E-mail:miaoguoyi@htu.edu.cn

相邻的基本篇章单元之间经由特定的修辞关系和核性连接形成上层篇章单元,重复上述操作,可以为该篇章构造出一棵带修辞关系的树状结构,称为篇章结构树.

在构建篇章结构树的过程中包括结构识别、核性识别和修辞关系识别. 核性识别即为主次识别,是篇章分析中的一个重要子任务,核性或主次指的是 EDU 分为核心(Nucleus,N)和卫星(Satellite,S)两种角色. 核心就是主句,指的是起主要作用、具有相对完整语义的 EDU;卫星就是从句,指的是起解释性作用、用于补充说明的 EDU. RST 中存在单核和多核关系. 单核关系表现为 N-S、S-N 两种类型,多核关系表现为 N-N 类型. 其中 N-S 代表篇章结构树的左子树是核心、右子树是卫星,S-N 代表右子树是核心、左子树是卫星,N-N 代表左右子树都是核心.

篇章的主次关系按照文本颗粒度分为微观和宏观两个层面. 在微观方面,主次关系指的是句子与句子、句群与句群之间核心与卫星的关系;在宏观方面,主次关系指的是段落与段落、章节与章节之间核心与卫星的关系.

主次识别的目的是识别出篇章结构树的左右子树中哪个是核心、哪个是卫星. 对主次识别的研究能帮助把握篇章的主旨、认识和理解篇章的中心主题和展开思路,更有效地挖掘篇章的宏观主题和各部分之间的语义关联.

受 RST 的启发,Li 等[2]提出一种基于连接依存树(Connective-driven Dependency Tree,CDT)的汉语篇章结构表示体系,该体系叶子节点是 EDU,中间节点是连接词,根据汉语本身特点并参照依存理论确定篇章关系的主次. 在 CDT 体系的指导下,他们手动标注了 500 个文档的汉语篇章结构语料库(Chinese Discourse Treebank,CDTB).

本文以 CDTB 中的一个例子来说明篇章中的主次关系. 如例 1 所示,a~g 分别为段落中的 EDU,这些 EDUs 是篇章结构树的叶子节点. 由这些 EDUs 构成的篇章结构树如图 1 所示. EDUs 通过特定的修辞关系连接在一起,包括本例中的总分、解说和并列关系. 主次关系则由箭头指出,箭头所指的为篇章单元中的核心,否则为卫星. 具体来说,图 1 篇章结构树根节点的左右子树中,右子树的内容是对左子树的解释说明,因此两棵子树之间的修辞关系为总分,而主次关系为 N-S. 由于 d、e 与 f、g 在篇章中的作用相同,都为具体描述,因此它们之间的修辞关系为并列,主次关系为 N-N.

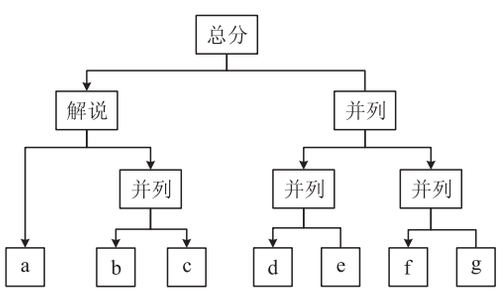

图 1　例 1 的篇章结构树

Fig. 1　Discourse structure parsing of example 1

例 1　a:据海关总署提供的统计数据,今年 1—2 月中国对外贸易进出口继续保持增长势头,b:进出口总值达 361 亿美元,c:比去年同期增长 13.9%,d:其中出口为 178.3 亿美元,e:比去年同期下降 1.3%;f:进口 182.7 亿美元,g:增长 34.1%.

a:According to the statistics provided by the General Administration of Customs, China's foreign trade imports and exports continued to grow from January to February this year, b:reaching a total value of 361 billion US dollars, c:which is a 13.9% increase compared to the same period last year, d:of this, exports accounted for 178.3 billion US dollars, e:marking a 1.3% decrease from the same period last year, f:while imports amounted to 182.7 billion US dollars, g:showing a 34.1% increase.

从例 1 可以发现,汉语侧重于"意合",其句子结构灵活,并依赖于语义和上下文来传达信息. 相反,英语侧重于"形合",依靠严格的语法规则、标点符号、连词、短语和从句来表达逻辑关系. 根据 Zhou 等[3]的统计,中文树库(Chinese Treebank,CTB)的篇章关系中没有连接词的隐式关系占 82%,而英文的宾州篇章树库(Penn Discourse Treebank,PDTB)中隐式关系只占 54.5%,这说明中文隐式关系的比例远高于英文.

考虑到中英文的特点,本文提出一种融合汉英双语信息的汉语篇章主次识别方法. 通过在训练过程中引入汉语及其对应的英文翻译信息,模型不仅能学习中文的语义和句法特征,还能从英文中学习更丰富的语义表示,在语义空间中将中文和英文形成某种程度上的对齐. 这有助于模型理解句子间的关系,提升主次识别效果. 本文的主要贡献如下:

(1)提出了一种基于预训练模型和注意力机制的中文主次识别模型. 使用预训练模型作为编码器提取文本通用特征,在提取的特征上应用注意力机制进一步捕获对主次识别有帮助的信息. 在中文主次识别上的多个性能指标均高于现有模型.

(2)提出了一种融合双语信息的中文篇章主次识别方法,与其他单独使用中文数据训练的方法不同,本文在模型训练过程中融入相应的英文翻译,能有效提升模型在中文主次识别上的精确率、召回率、宏平均 $F1$ 值和微平均 $F1$ 值.

1 相关工作

自动构建篇章结构树需要进行结构识别、核性识别和修辞关系识别 3 个子任务. 在早期的研究中,对核性的识别通常被视为篇章修辞结构分析中的一项辅助工作,没有得到足够的重视[4]. 这一任务可以从微观和宏观角度来进行,微观角度关注段落内句子之间的核性,而宏观角度则关注不同段落之间的核性. 这些研究大多数集中在英文修辞结构篇章树库(RST Discourse Treebank,RST-DT)[5]、CDTB 和宏观汉语篇章树库(Macro Chinese Discourse Treebank,MCDTB)[6]上,MCDTB 根据 RST 标注宏观语篇关系,包含 720 篇文章,Kappa 值超过 0.6. 在 RST-DT 上,Joty 等[7]使用条件随机场计算篇章单元的修辞结构关系,并使用动态规划算法进一步优化篇章结构树的构建. Li 等[8]提出使用依存结构直接表示 EDU 之间的关系,他们用 Eisner 算法和最大生成树算法构建基于依存结构的篇章分析器. Ji 等[9]采用表示学习方法将表面特征转化到潜在空间,所得到的 Shift-Reduce 篇章分析器在 RST-DT 上的核性预测率达到 71.13%. 在 CDTB 上,Kong 等[10]的 ME 模型利用语义相似度、词汇特征,通过最大熵模型进行核性识别. Zhang 等[11]提出了一个自上而下的文本级篇章分析器,将篇章分析视为递归分裂点的排序任务,在核性识别上取得了 57.3%的微平均 $F1$ 值.

近年来,随着深度学习技术的发展,核性识别逐渐被视为独立的任务. 目前,对于微观核性识别任务,Xu 等[12]使用双向长短时记忆网络(Bidirectional Long Short-Term Memory,Bi-LSTM)和卷积神经网络(Convolutional Neural Network,CNN)分别获得文本全局编码信息和局部编码信息,进而使用文本匹配方法建模语义交互,最后输入到分类器进行识别. 王体爽等[13]使用与 Xu 等[12]相同的 Bi-LSTM 和 CNN 编码器获取编码信息,然后融合两部分信息送入门控记忆网络,使用该门控网络捕获语义特征,进而对主次关系识别,他们的方法在 CDTB 上取得了 71.1%的微平均 $F1$ 值. 对于宏观核性识别任务,孙振华等[14]在 MCDTB 上使用 BERT 作为编码器进行主次识别,其工作在多个方面的指标都优于之前的传统神经网络方法.

先前的工作大多数使用 Word2Vec[15]等预训练词向量进行编码,文本经过编码后得到的是静态词嵌入,无法融入上下文的全局信息. 随着预训练模型如 BERT[16]、GPT[17]的提出,预训练加微调的方法所达到的效果逐渐超过了传统神经网络的方法,这是由于预训练模型在大量数据集上训练而具备强大的泛化能力,模型本身包含了数据的通用特征和信息.

在使用双语信息解决中文分析问题方面,杨紫怡等[18]引入双语信息对中文进行零指代识别和消解,通过引入一系列双语词对齐特征,包括候选先行词的英文翻译及其词性、中文零指代项在英文翻译中对应的词及其词性等,系统的 $F1$ 值提升了 5%. Wang 等[19-20]和 Wang 等[21]都引入双语信息来提升其模型在翻译省略代词(Dropped Pronouns,DP)时的性能. 具体来说,前者通过在双语平行语料库中标注中文 DP,并在其机器翻译系统中利用这些预测的 DP,使 BLUE 值提高了 1.58%. 后者通过在模型编码端和解码端隐藏层中嵌入被省略的 DP 信息来提升整体翻译性能.

本文将多语言预训练模型作为编码器应用于中文微观主次识别,并提出一种融合双语信息的主次识别方法. 相较于之前最好的工作,本文提出的模型和方法在宏平均 $F1$ 值和微平均 $F1$ 值上比 Wang 等[13]使用 Bi-LSTM 和 CNN 编码的方法提高 8.7%和 6.1%;相较于仅使用预训练模型和单语数据训练的方法,本文提出的融合双语信息的主次识别方法对于 mBERT、mT5、XLM-R 3 种模型在微平均 $F1$ 值上分别提高了 1.6%、3.5%、1.3%. 在汉英篇章结构平行语料库(Chinese-English Discourse Treebank,CEDT)[22]上的实验进一步证明本文所提出的融合双语信息的主次识别方法的有效性.

2 实现方法

图 2 展示了本文所提出方法的实现流程,主要包含数据处理和模型训练两个部分. 首先,原始中文数据会被翻译成对应的英文,并加入训练集,如“2.1”节所述. 接着,训练集的数据被打乱顺序送入模型,经过多语言预训练模型的编码得到通用特征,并通过注意力机制提取局部特征,最后进行分类. 模型部分见“2.2”节和“2.3”节.

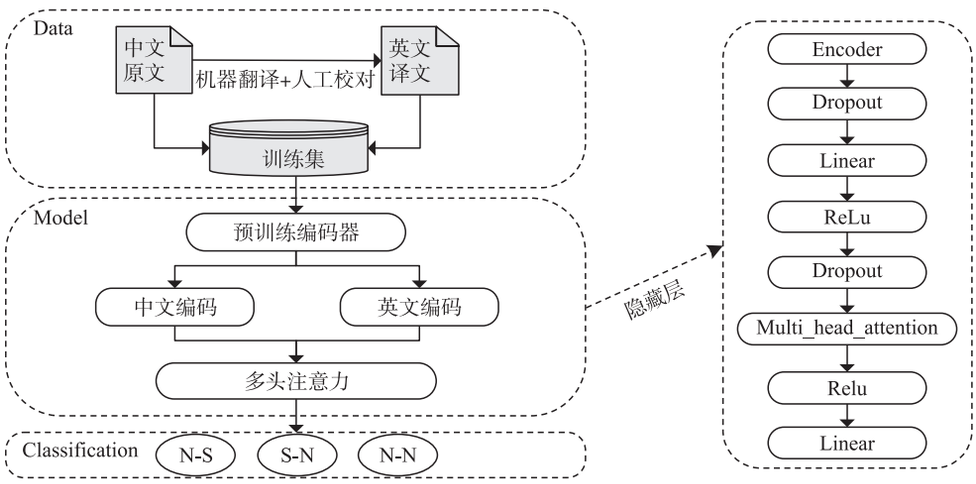

图 2　融合双语信息的主次识别方法实现流程图

Fig. 2　Flowchart of the bilingual information fusion method for nuclearity recognition

2.1　数据预处理

在某些存在多核关系的篇章结构树中,多核关系节点的孩子节点数量超过 2 个. 对于这些节点,将最左侧篇章单元作为该节点的左孩子,而其余节点作为该节点的右孩子,且所有节点均标注为核心,如此反复,直到最终所有树节点均只包含 2 个孩子节点为止. 该过程称为篇章树的二分化,这种做法的目的是降低数据复杂度,便于深度学习模型计算. 调研结果显示,以往的绝大多数主次识别工作在预处理时均进行二分化操作[23].

本文分别在 CDTB 和 CEDT 2 个语料库上进行了实验,CDTB 语料库是汉语单语语料库,为了获取对应的英语语料,本文使用机器翻译将原始数据翻译为英语后加入训练集,并打乱顺序. 为了检验本文方法对于不同质量翻译文本的鲁棒性,本文使用不同的翻译引擎获取了不同的英语语料,翻译引擎包括百度翻译、谷歌翻译、Deepl 翻译.

在进行实验时,每条输入模型的样本为 2 个相邻的篇章单元,每个 batch 内既包含中文样本,又包含英文样本,但每条样本只包含一种语言的句子.

2.2　编码器

本文对比了使用不同跨语言预训练模型作为编码器的效果,如 mBERT[24]、XLM-R[25]等. 两种模型都在上百种语言上进行了预训练,跨语言泛化能力强. 由于预训练模型具有强大的信息融合能力,分词后的 token 经过编码后已经基本融合了句子上下文的信息. 因此,仅仅使用预训练模型进行编码和分类的效果就已经超过了之前工作使用 Word2Vec 编码并用 LSTM 和 CNN 进行分类的效果.

BERT 在训练过程中采用 MLM 和 NSP 两种任务,模型能够有效捕获双向上下文信息. mBERT 是 BERT 的多语言版本,在训练时使用了 104 种语言,共有 12 个隐藏层,隐藏层维度为 768 维,每层有 12 个头,拥有大约 110 M 参数.

与 BERT 类似,XLM-R 在预训练时同样使用了 MLM 任务,不同的是没有使用 NSP 任务,这可以帮助模型更专注于理解语言内部的复杂性. 此外,XLM-R 使用了比 mBERT 更多的训练数据、更长的训练时间和更大的 batch size. 其使用了 2.5 T 的 CommonCrawl 数据,包括了 100 种语言文本,共拥有 270 M 参数.

双语信息的编码如式(1)所示. 式中($x_{zh} \parallel x_{en}$)表示将中文和英文样本打乱顺序并列输入 mini batch 中,编码器 E 同时对这两类输入进行独立编码,两类样本在序列维度上并列处理. 该操作不涉及跨样本的

序列拼接,通过预训练模型内在的多语言能力实现跨语言编码.

$$\text{Encoding}_{\text{bilingual}} = E(x_{zh} \parallel x_{en}). \tag{1}$$

2.3 注意力层

在主次识别任务中,模型需要确定文本中的主要和次要元素,由于编码器的输出为序列中每个单词的上下文表示,无法突出关键信息. 而注意力可以加强对关键信息的强调、对文本进行更细粒度的分析,区分主题和细节. 因此本文在编码器的输出上应用了多头注意力机制[26],这可使模型对于输入序列的不同部分具有不同的关注程度,从而能更好地捕捉主句和从句的信息.

应用多个不同的头可以专注于输入数据的不同方面,某些头可以专注于捕捉语法结构关系,而其他头则可能关注语义相关性. 所有头的输出随后被拼接并通过一个最终的线性层来整合,形成一个统一的输出表示.

此外,BERT、XLM-R 在预训练时使用的语料是 BooksCorpus、英语维基百科[16],这些大多是在非特定任务上的通用数据. 本文所使用的 CDTB 和 CEDT 语料库主要来自新闻、广播等,这些数据具有篇章单元长度长、语义跨度大、涵盖信息多、语法正式等特点,使用注意力机制在这些数据上进行微调,能更好地捕捉数据相关细节和特性. 例如,在 CEDT 中的句子“国有企业继续居于主导地位,外商投资企业仍然发挥重要的作用”中,注意力机制能给关键词“国有企业”、“外商投资企业”、“主导地位”、“重要的作用”更多的注意力分数,帮助模型捕捉句子中的层次结构,区分主次信息,进而提升主次识别的正确率.

注意力机制首先涉及 3 组线性变换,这些变换将输入映射到一个新空间,便于后续注意力分数的计算,线性映射如式(2)所示. 其中 $\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$ 分别为查询(Query)、键(Key)和值(Value)矩阵,H 是输入的隐藏状态,$\boldsymbol{W}^{\boldsymbol{Q}}$、$\boldsymbol{W}^{\boldsymbol{K}}$、$\boldsymbol{W}^{\boldsymbol{V}}$ 是对应的权重矩阵.

$$\boldsymbol{Q} = \boldsymbol{W}^{\boldsymbol{Q}}H, \boldsymbol{K} = \boldsymbol{W}^{\boldsymbol{K}}H, \boldsymbol{V} = \boldsymbol{W}^{\boldsymbol{V}}H, \tag{2}$$

$$\text{Attention} = \text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{dk}}\right)\boldsymbol{V}. \tag{3}$$

进一步,模型通过计算 Query 和 Key 之间的点积来得到原始的注意力分数. 为了防止计算过程中分数过大导致梯度消失问题,分数会被 Key 矩阵维度的平方根所缩放. 随后,应用 Softmax 函数将这些分数转换成概率形式,这些概率即为注意力权重. 最后,这些权重和矩阵 $\boldsymbol{V}$ 进行加权求和,生成每个头的输出. 注意力分数的计算见式(3). 多头的整合如式(4)所示,h 是头数,$\boldsymbol{W}^{\boldsymbol{O}}$ 是输出线性变换的权重矩阵.

$$\text{MultiHead}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \text{Concat}(head_1, \cdots, head_h)\boldsymbol{W}^{\boldsymbol{O}}. \tag{4}$$

2.4 分类层

由于注意力机制主要涉及加权与线性操作,而在语料中,词与词之间的关系往往非常复杂,并非简单的线性组合. 因此本文在注意力层后应用了如式(5)所示的 ReLU 非线性激活函数. 使用非线性函数可以增强模型的非线性处理能力,这对于理解和区分句子的主次结构尤为关键.

$$\text{ReLU}(x) = \max(0, x). \tag{5}$$

进一步,对于 ReLU 的输出,本文使用线性变换将隐藏层维度转换为 3,这是由于共有 3 种主次关系,所以目标类别数为 3.

最后用如式(6)所示的 Softmax 激活函数将线性输出转换为概率分布. 其中 $\boldsymbol{W}$ 是权重矩阵,$\boldsymbol{b}$ 是偏置向量,x 是来自线性层的输出.

$$p = \text{Softmax}(\boldsymbol{W}x + \boldsymbol{b}). \tag{6}$$

3 实验

3.1 数据集

3.1.1 汉语篇章结构语料库(Chinese Discourse Treebank,CDTB)

CDTB 共有 500 个文档(chtb001-chtb0657),文本全部来自 CTB6.0,句子标号从 1 到 6 648. 每个段落标注为一棵连接依存树,共有效标注 2 342 个篇章(段落). CDTB 共包含 10 643 个子句,每棵篇章结构树平均包含 4.5 个子句. 平均每个有效标注的句子包含 2 个子句,每个子句平均长度为 22 个汉字. 共有主次关系 7 310 个,其中,N-S 类型有 2 108 个,S-N 类型有 1 447 个,N-N 类型有 3 755 个.

3.1.2 汉英篇章结构平行语料库(Chinese-English Discourse Treebank,CEDT)

CEDT 采用“结构对齐,关系对齐”[27]的标注策略,对 CDTB 中的前 175 个文档(chtb001-chtb175)进行了双语平行对齐标注,即为汉英翻译文本标注了对齐篇章结构信息,包括将双语篇章中的每个 EDU 对齐、EDU 的主次关系对齐等策略. 冯文贺等[22]对 CEDT 的切分对齐、结构对齐、关系对齐、连接词对齐、关系角色与中心对齐进行了评估,结果表明 CEDT 标注质量高,具有一定的可计算性和实用性. CEDT 的中文主次关系共有 1 467 个,其中 N-S 类型有 453 个,S-N 类型有 93 个,N-N 类型有 921 个.

3.2 实验设置

在 CDTB 上的实验使用了与 Xu 等[12]和王体爽等[13]相同的数据集与数据集划分方法,数据集划分及主次关系数量统计如表 1 所示. 为了防止过拟合,在编码层和第一个 ReLU 层后分别添加了 dropout 层,dropout 率为 0.3,注意力层的头数设置为 8. 在训练过程中使用 AdamW 优化器,权重衰减系数为 0.01,batch size 大小为 48. 使用学习率预热策略,学习率设置为 3e-5,预热阶段占整个训练过程的 1%,在验证集 loss 最小时保存模型. 为保证实验效果,本文进行了 5 次实验,报告的是 5 次结果的平均值.

表 1 CDTB 数据集划分

Table 1 Partitioning of the CDTB dataset

	训练集	验证集	测试集
文档数量	400	50	50
文档列表	0001-0080,0101-0180,0201-0280,0301-0325, 0400-0454,0520-0554,0590-0596,0600-0637	0081-0090,0181-0190,0281-0290, 0500-0509,0638-0647	0091-0100,0191-0200,0291-0300, 0510-0519,0648-0657
主次关系	N-S:1 728 S-N:1 193 N-N:2 964	N-S:173 S-N:150 N-N:407	N-S:207 S-N:104 N-N:384

对 CEDT 的主次关系数量统计如表 2. 在 CEDT 数据集上的实验使用了同样的超参数,但由于实验数据有限,没有提前划分数据集,而是使用 5 折交叉验证法,采取早停机制,即如果在测试集上的准确率经过 3 个连续的 epoch 仍未提升 1%,则停止训练.

表 2 CEDT 数据集统计

Table 2 CEDT dataset statistics

	N-S	S-N	N-N
中文	453	93	921
英文	468	96	928

3.3 实验结果

首先,本文对比了之前的工作——ME 模型[10]、TMN 网络[12]、GMN-Nu 网络[13]与使用 mT5[28]、mBERT、XLM-R 预训练模型进行主次识别的实验效果. 报告了模型在 3 种主次关系上的精确率(P)、召回率(R)、$F1$ 值及整体的宏平均 $F1$ 值(Macro-$F1$)和微平均 $F1$ 值(Micro-$F1$). 表 3 结果显示,在微平均 $F1$ 值上,XML-R 取得了最高值 75.9;在宏平均 $F1$ 值上,XML-R 取得了最高值 69.6. 此外,XML-R 在其他大多数性能指标中都为最佳.

表 3 使用预训练模型与传统模型的对比实验

Table 3 Comparative experiment between pre-trained models and traditional models

Model	N-S $P/R/F1$	S-N $P/R/F1$	N-N $P/R/F1$	Macro-$F1$	Micro-$F1$
ME	32.2/15.1/20.5	40.0/15.0/21.8	65.6/87.8/75.0	42.3	60.5
TMN	69.1/45.4/54.8	39.2/49.0/43.6	76.2/83.3/79.6	60.4	69.0
GMN-Nu	60.8/59.9/60.3	47.3/41.4/44.1	79.6/82.3/80.9	61.9	71.1
mT5	63.5/63.4/63.4	54.6/45.2/48.2	80.4/82.7/81.5	64.4	72.8
mBERT	69.6/57.6/62.6	53.8/47.1/48.2	80.6/87.0/83.6	64.8	74.2
XLM-R	69.1/66.4/67.6	55.1/61.0/57.6	84.0/83.2/83.6	69.6	75.9

从实验结果可以看出,仅仅使用预训练模型进行分类的效果在绝大多数性能指标上就已经优于之前的模型. 主要原因为编码器在以下两方面有所不同:

在上下文编码方面,之前的模型在获取词嵌入时使用的是静态的 Word2Vec 词向量,而预训练模型获取的是能感知上下文的动态词向量,其依赖于整个句子,能捕获更复杂的句子关系. 其次,在训练数据方面,Word2Vec 是在包含 10 亿词的 Google 内部新闻数据集上训练的,而 mBERT 是在包含 104 种语言的维

基百科上训练的,仅英文单语言的维基百科就包含 25 亿词,二者的数据量具有较大差距. XLM-R 和 mT5 也都在超过上百种语言上进行了预训练,因此具有强大的泛化能力.

接着,本文检验了注意力层的作用,分别在 3 种预训练模型的输出上应用多头注意力机制,注意力的头数为 8,dropout 率为 0.01. 表 4 结果显示,相较于直接使用预训练模型分类,在预训练权重的输出上应用多头注意力可以进一步提升模型在主次识别上的效果,具体表现为应用了注意力的预训练模型比原始预训练模型的多个性能指标均有提升. 对于微平均 $F1$ 值,注意力使 XLM-R、mBERT、mT5 分别提升了 1.3%、2.6%、3.3%;对于宏平均 $F1$ 值,注意力使 XLM-R、mBERT、mT5 分别提升了 0.8%、4.4%、4.5%.

表 4 注意力机制的消融实验

Table 4 Ablation study of attention mechanisms

Model	N-S *P*/*R*/*F*1	S-N *P*/*R*/*F*1	N-N *P*/*R*/*F*1	Macro-*F*1	Micro-*F*1
XLM-R	69.1/66.4/67.6	55.1/61.0/57.6	84.0/83.2/83.6	69.6	75.9
XLM-R+Att	70.9/66.7/68.7	59.7/57.3/57.8	83.5/85.9/84.6	70.4	77.2
mBERT	69.6/57.6/62.6	53.8/47.1/48.2	80.6/87.0/83.6	64.8	74.2
mBERT+Att	66.8/61.4/64.0	62.0/54.8/58.2	83.1/88.0/85.5	69.2	76.8
mT5	63.5/63.4/63.4	54.6/45.2/48.2	80.4/82.7/81.5	64.4	72.8
mT5+Att	68.5/68.0/68.1	58.6/52.9/54.7	83.5/84.5/83.9	68.9	76.1

注:Att 代表注意力机制.

之前的工作使用 RNN 来捕获长距离依赖关系,此类序列模型的问题在于信息衰减,距离较远的位置存在梯度消失或爆炸. 而本文使用的自注意力通过全连接权重矩阵直接建立任意两个位置的关联. 此外,8 个注意力头可以并行计算不同子空间的注意力,部分头关注语法结构,另一部分头捕捉语义重点,对于句子的不同成分给予不同的关注程度.

然后,本文在预处理后的 CDTB 语料库上进行了融合双语信息的实验,具体做法为使用机器翻译将原始中文训练集翻译为英文后打乱顺序一同训练. 最后单独使用中文的测试集进行测试. 表 5 展示了使用双语数据训练和添加注意力层后使用双语数据训练的结果.

表 5 融合双语数据的实验

Table 5 Monolingual *vs*. bilingual training comparisons

Model	N-S *P*/*R*/*F*1	S-N *P*/*R*/*F*1	N-N *P*/*R*/*F*1	Macro-*F*1	Micro-*F*1
mBERT	69.6/57.6/62.6	53.8/47.1/48.2	80.6/87.0/83.6	64.8	74.2
mBERT*	71.6/57.8/63.8	55.2/53.8/54.0	81.4/88.0/84.4	67.4	75.7
mBERT+Att	71.5/57.5/63.1	56.6/47.5/50.3	80.2/88.2/83.9	65.7	74.9
mBERT+Att*	69.7/63.1/66.1	58.6/52.1/54.9	81.4/86.4/83.8	68.3	75.8
mT5	63.5/63.4/63.4	54.6/45.2/48.2	80.4/82.7/81.5	64.4	72.8
mT5*	71.8/61.1/65.9	61.6/50.0/54.0	80.5/88.0/84.0	68.0	76.0
mT5+Att	68.5/68.0/68.1	58.6/52.9/54.7	83.5/84.5/83.9	68.9	76.1
mT5+Att*	71.9/63.6/67.3	61.8/51.4/54.8	81.2/87.1/83.9	68.7	76.3
XLM-R	69.1/66.4/67.6	55.1/61.0/57.6	84.0/83.2/83.6	69.6	75.9
XLM-R*	73.7/63.7/67.9	59.4/56.9/58.1	82.2/87.3/84.6	70.2	77.2
XLM-R+Att	70.9/66.7/68.7	59.7/57.3/57.8	83.5/85.9/84.6	70.4	77.2
XLM-R+Att*	73.9/60.9/66.7	58.9/61.9/60.3	82.4/87.5/84.8	70.6	77.2

注:* 代表使用双语训练.

结果显示,对于 3 种预训练模型,相较于使用单语言训练和原始模型分类,应用注意力和双语训练的方法普遍效果更好. 具体来说,相较于单独使用 mBERT 模型在单语言上训练,使用注意力和双语训练分别在宏平均 $F1$ 值和微平均 $F1$ 值上提高了 3.5%和 1.6%;同样,对于 XLM-R 模型,宏平均 $F1$ 值和微平均 $F1$ 值分别提高了 1.0%和 1.3%;对于 mT5 模型,宏平均 $F1$ 值和微平均 $F1$ 值分别提高了 4.3%和 3.5%.

在所有的训练方法中,使用 XLM-R+Att 在双语上训练的效果最好,比之前最好的 GMN-Nu 模型在宏平均 $F1$ 值上提高了 8.7%,在微平均 $F1$ 值上提高了 6.1%.

为了验证本文方法对于译文质量的鲁棒性,本文分别使用百度翻译、谷歌翻译、Deepl 翻译 3 种翻译引擎对原始中文数据翻译,所得译文均为机器生成的原始输出,未进行任何人为筛选或加工. 在此基础上,

分别基于3种译文构建双语输入数据,并在此数据上使用XLM-R、mBERT与mT5 3种多语言预训练模型开展融合双语信息的主次识别实验.

实验结果如表6所示. 在所有翻译引擎生成的译文数据上,3种模型在引入双语信息后均表现出较明显的性能提升. 以微平均$F1$值为例,相较于单语训练,mBERT模型在百度、谷歌、Deepl翻译数据上分别提升了1.5%、1.8%、1.1%;mT5模型分别提升了3.2%、3.1%、2.3%;XLM-R模型分别提升了1.3%、0.8%、1.0%.

上述结果表明,尽管3种翻译引擎在翻译质量上存在一定差异,但融合双语信息所带来的性能增益在不同译文下仍较为一致,表明所提方法对译文质量具有较强的鲁棒性与泛化能力. 此外,模型在未经人工优化的机器译文上依然能有效学习英文结构中的主从信息,说明该方法不依赖高质量人工翻译即可在实际应用中发挥作用,具有较强的现实可行性.

表6 使用不同翻译引擎数据的实验

Table 6 Experiment using data from different translation engines

Model	翻译引擎	N-S $P/R/F1$	S-N $P/R/F1$	N-N $P/R/F1$	Macro-$F1$	Micro-$F1$
mBERT	单语	69.6/57.6/62.6	53.8/47.1/48.2	80.6/87.0/83.6	64.8	74.2
	百度翻译	71.6/57.8/63.8	55.2/53.8/54.0	81.4/88.0/84.4	67.4	75.7
	谷歌翻译	72.1/60.2/64.9	59.5/48.7/52.7	80.9/88.7/84.4	67.4	76.0
	Deepl 翻译	66.1/67.8/66.7	52.5/57.5/54.3	85.7/82.3/83.9	68.3	75.3
mT5	单语	63.5/63.4/63.4	54.6/45.2/48.2	80.4/82.7/81.5	64.4	72.8
	百度翻译	71.8/61.1/65.9	61.6/50.0/54.0	80.5/88.0/84.0	68.0	76.0
	谷歌翻译	70.6/64.9/67.6	56.7/49.0/52.1	81.5/86.4/83.8	67.8	75.9
	Deepl 翻译	74.3/58.2/65.0	53.3/51.3/51.8	79.8/87.3/83.3	66.7	75.1
XLM-R	单语	69.1/66.4/67.6	55.1/61.0/57.6	84.0/83.2/83.6	69.6	75.9
	百度翻译	73.9/60.9/66.7	58.9/61.9/60.3	82.4/87.5/84.8	70.6	77.2
	谷歌翻译	70.2/67.4/68.4	58.2/59.0/58.2	83.9/84.5/84.2	70.3	76.7
	Deepl 翻译	72.9/64.4/68.2	55.3/58.1/56.4	83.4/86.2/84.7	69.8	76.9

此外,对于更高质量的译文,本文使用mBERT+Att模型在CEDT上训练,CEDT为中英双语平行语料,其对CDTB中的前175个文档进行双语平行对齐标注,英语翻译经过人工校对. 表7展示了在CEDT上的实验结果,本文报告了模型在3种不同语言数据上训练时的表现. 在单独使用中文和英文训练时,分别使用中文和英文测试;在使用中文和英文双语混合训练时,使用中文测试. 中英混合后,相较于单独使用中文,3种主次类型的精确率、召回率、$F1$值均有提高. 具体来说,对于N-S类型,双语训练的精确率比单语提高了8.2%,召回率提高了7.7%,$F1$值提高了8.0%;对于S-N类型,精确率提高了10.1%;对于N-N类型,精确率提高了5.5%,召回率提高了2.6%,$F1$值提高了4.2%. 总体而言,融合双语信息的训练方法在中文微观篇章主次识别上的宏平均$F1$值与微平均$F1$值分别提高了10.2%和5.8%.

表7 在CEDT上融合双语数据训练的实验

Table 7 Monolingual *vs.* bilingual training on CEDT

语言	N-S $P/R/F1$	S-N $P/R/F1$	N-N $P/R/F1$	Macro-$F1$	Micro-$F1$
mBERT+Att(zh)	90.0/90.7/90.3	84.8/74.7/78.3	94.0/95.8/94.8	87.8	92.7
mBERT+Att(en)	92.0/91.2/91.6	75.1/72.0/72.6	92.8/97.0/94.8	86.3	92.5
mBERT+Att*	98.2/98.4/98.3	94.9/99.3/96.8	99.5/98.4/99.0	98.0	98.5

注:zh代表使用中文训练,en代表使用英文训练,* 代表使用双语训练.

受到英语信息能帮助汉语主次识别的启发,本文验证了与英语同属印欧语系的德语和法语在信息迁移方面的能力,使用谷歌翻译获取了德语和法语的译文,并用同样的双语训练方法进行了实验,结果如表8所示. 结果显示,对于3种多语言预训练模型,分别使用其他语言对(中法、中德)的数据训练同样能提升在中文主次识别上的效果. 具体来讲,对于mBERT,使用中法和中德训练分别在Micro-$F1$上提升了2.6%和2.2%;对于mT5,使用中英、中法和中德训练在Micro-$F1$上提升相同,都为3.1%;对于XLM-R,使用中法和中德训练分别在Micro-$F1$上提升了1.2%和1.1%. 可以看出,使用英、法、德3种语言数据分别和中文混合训练均能有效提升多语言模型的主次识别效果,本文认为主要有以下几点原因:

一方面,3种语言同属印欧语系,在句法结构上均体现出较强的显性连接特征. 例如,德语中存在严格

的语序规则,常通过将从句动词置于句末等方式明确区分主从句结构. 而法语的连接词系统更加复杂,且在主从句之间具有严格的时态搭配规则,有助于凸显从属关系. 实验结果表明本文的双语框架具有扩展到其他语言对的潜力,尤其适用于与中文在句法结构上互补的语言对.

另一方面,英语、德语和法语都是模型的预训练语料中占比较高的语言. 例如,XLM-R 的训练语料中,英语、德语和法语分别占比 12.92%、2.86%和 2.44%;mT5 的训练语料中,英语、德语和法语分别占比 5.67%、3.05%和 2.89%;mBERT 的训练语料同样包含这 3 种语言. 这表明对于这些语言,模型在预训练阶段已经积累了丰富的语言知识,当这些语言参与到下游任务时,能更有效地激活模型的多语言共享语义空间,帮助模型跨语言结构对齐.

表 8 使用不同语言译文的实验

Table 8 Experiment using data from different target languages

Model	译文语言	N-S *P*/*R*/*F*1	S-N *P*/*R*/*F*1	N-N *P*/*R*/*F*1	Macro-*F*1	Micro-*F*1
mBERT	单语	69.6/57.6/62.6	53.8/47.1/48.2	80.6/87.0/83.6	64.8	74.2
	英语	72.1/60.2/64.9	59.5/48.7/52.7	80.9/88.7/84.4	67.4	76.0
	法语	72.7/60.4/66.0	55.9/50.0/52.8	81.7/89.5/85.4	68.1	76.8
	德语	67.0/69.6/68.0	61.5/47.6/53.6	83.2/85.5/84.3	68.6	76.4
mT5	单语	63.5/63.4/63.4	54.6/45.2/48.2	80.4/82.7/81.5	64.4	72.8
	英语	70.6/64.9/67.6	56.7/49.0/52.1	81.5/86.4/83.8	67.8	75.9
	法语	74.2/55.6/63.5	63.3/48.1/54.6	78.1/90.5/83.9	67.4	75.9
	德语	76.5/56.5/65.0	56.0/45.2/50.0	78.7/90.7/84.3	66.4	75.9
XLM-R	单语	69.1/66.4/67.6	55.1/61.0/57.6	84.0/83.2/83.6	69.6	75.9
	英语	70.2/67.4/68.4	58.2/59.0/58.2	83.9/84.5/84.2	70.3	76.7
	法语	73.9/65.7/69.6	54.1/63.5/58.4	84.1/85.0/84.5	70.8	77.1
	德语	71.8/73.9/72.9	49.7/69.2/57.8	88.6/80.0/84.1	71.6	77.0

3.4 实验结果分析

关于应用注意力及融合双语信息为何能提高对中文微观篇章主次识别的效果,本文认为主要有以下几点原因:

3.4.1 应用注意力能帮助模型捕捉隐式关系

应用注意力有助于模型学会忽略非关键信息,专注于获取决定篇章单元结构和意义的核心语言特征,这些核心语言特征是帮助识别主次的关键,有益于模型在面对中文数据时做出更准确的判断.

如例 2 所示,未应用注意力之前,模型将其预测为 N-N 结构,应用注意力之后,模型将其正确预测为 N-S 结构. 究其原因,这句话由一个介宾、三个动宾结构组成,没有显式连接词引导,模型难以区分不同句子结构的区别. 而注意力机制能够捕捉到隐含的逻辑关系,给“以……为重点”更多的权重分数,这些词在语义上起到了引导主次关系的作用. 进而能够识别出“以西部塔里木盆地等新油气区为重点”是句子的主要目标,而“打开勘探新局面”和“拿到更多的油气资源”是实现这一目标的具体措施,从而正确判断为 N-S 结构.

例 2 在科技攻关方面,要以西部塔里木盆地等新油气区为重点,打开勘探新局面,|拿到更多的油气资源.

In terms of scientific and technological research,we should focus on new oil and gas areas such as the western Tarim Basin,open up new prospects for exploration,| and get more oil and gas resources.

3.4.2 融合双语数据能促进知识迁移

模型接收的是来自不同语言的语法、用词和表达方式,因此融合双语数据训练的一个显著优势是增加了数据的多样性和丰富性,这能促进句法和语义层面的知识迁移. 尽管中英两种语言在句法结构上有所不同,但它们所蕴含的修辞关系和语义信息是相似的,模型学习这些蕴含于句中的修辞关系和语义信息有助于对主次关系的判断.

例 3 贷款将向能源、交通、电力等基础设施产业倾斜,|尤其以国外大公司在华设立的大中型企业为重点.

The loans will be in favor of infrastructure industries such as energy sources,transportation,and electrical power,etc.,| especially focusing on large- and medium-scale enterprises set up in China by large foreign companies.

例 3 是逗号分隔的流水句,未融入英文信息之前,模型将其预测为 N-N 结构,这很大程度上受到“尤其”、“重点”等单词的影响,这些部分虽然是具体的描述,但在中文里并没有明显的从属结构标记.

而在英文中,前半句为主谓宾结构,是包含完整信息的主句. 中文“贷款将向……倾斜”也是完整的句法结构,模型通过对齐双语的句法结构以强化对主句的理解和识别. 英文后半句句首副词“especially”明确标示后面是用于补充说明或强调的非核心内容,这与中文里“尤其”引导的非主干成分建立映射,帮助模型识别到连接词后的结构是次要信息.

后续的“focusing on…”是一个明确的现在分词短语作状语从句,用于补充说明主句,是从属关系的标志. 同时中文“尤其以……为重点”在语法上不独立,与英文的从句具有相似的语义功能. 中英两句在语义上形成映射,英文中的从属关系标志帮助模型捕捉中文里的主次结构边界. 通过对齐中英句子,模型可以抽象出“结构、语义完整”是主句的一个强特征. 因此融入英文信息之后,模型将其正确预测为 N-S 结构.

例 4 种种主张碰壁之余,中国人才找到了马克思列宁主义,|才认定只有社会主义才能救中国.

Only after the repeated failures of various propositions did the Chinese people finally discover Marxism-Leninism,| thereby coming to the conviction that socialism alone can save China.

融入双语之前,模型将例 4 预测为 N-S 结构,而融入双语信息之后,模型成功将其预测为 N-N 结构. 分析该错误可发现,中文第二个子句由副词“才”引出,具有较强的承接与因果色彩,容易被模型判断为对前句的附属推论,从而归类为从句. 并且中文中缺乏显性连接词或句法标记,进一步加大了主次识别的困难.

引入英文翻译后,模型得以利用英语中更为显式的句法结构对该句进行更准确的解析. 首先,句首“Only after”引导的时间状语从句是用于修饰主干的从句,清晰地表明了主句发生的条件,揭示第一个子句为主干成分. 相比之下,中文中“之余”虽然也有从属意味,但其语法不如英文结构明晰,在形式上更为松散,尤其在缺乏明确连词的情况下更难界定其是否为主句. 倒装结构在句法上揭示整句话的主干成分为“did the Chinese people finally discover…”,与“only after”引导的内容不谋而合.

其次,“thereby”引导非限定性结果状语结构,该结构虽然与前一子句在逻辑上是递进关系,但在语法上并非依附于主句的从句,而是与前一子句并列推进的延伸信息. 而中文中的“才认定”具有承接意味,容易在形式上被模型认定为从属关系. 模型在学习中可以借助英语中清晰的语义延伸的形式,纠正其在中文中因承接词而将子句误判,正确地将两个子句都识别为主句.

4 结论

本文提出了一种融合双语信息的中文篇章主次识别方法,与传统的使用单语数据训练的方法不同,该方法在模型训练阶段使用中文和英文两种语言的数据. 使用多语言预训练模型对原始文本编码,在得到的编码上应用注意力机制,通过有效利用蕴含在不同语言文本中的句子结构信息,增强了模型的泛化能力和抗噪能力,有效提升了在中文上的主次识别效果. 在 CDTB 上的实验表明,本文提出的模型和方法比当前最优的 GMN-Nu 模型在微平均 $F1$ 值上提高了 6.1%,在宏平均 $F1$ 值上提高了 8.7%. 在 CEDT 上的实验进一步证实本文所提出的方法的有效性.

在未来的工作中,将尝试把此种方法应用于篇章分析中的宏观篇章主次识别、篇章修辞关系识别等任务中,以进一步检验本文方法在不同任务中的作用效果.

[参考文献]

[1] Mann W C, Thompson S A. Rethorical Structure Theory: toward a functional theory of text organization[J]. Text-Interdisciplinary Journal for the Study of Discourse, 1988, 8(3): 243-281.

[2] Li Y C, Feng W H, Sun J, et al. Building Chinese discourse corpus with connective-driven dependency tree structure[C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Doha: Association for Computational Linguistics, 2014: 2105-2114.

[3] Zhou Y P, Xue N W. PDTB-style discourse annotation of Chinese text[C]//Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics. Jeju Island: Association for Computational Linguistics, 2012: 69-77.

[4] 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.
[5] Carlson L,Marcu D,Okurowski M E. Building a discourse-tagged corpus in the framework of rhetorical structure theory[M]// Current and New Directions in Discourse and Dialogue. Dordechat,Netherlands:Klumer Academic Publishers,2003:85-112.
[6] Jiang F,Xu S,Chu X M,et al. MCDTB:A macro-level Chinese discourse TreeBank[C]//Proceedings of the 27th International Conference on Computational Linguistics. Santa Fe:Association for Computational Linguistics,2018:3493-3504.
[7] Joty S R,Carenini G,Ng R T,et al. Combining intra- and multi-sentential rhetorical parsing for document-level discourse analysis[C]//Proceedings of the 51st Annuel Meeting of the Association for Computational Linguistics. Bulgauria:Association for Computational Lingustics,2013:486-496.
[8] Li S J,Wang L,Cao Z Q,et al. Text-level discourse dependency parsing[C]//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Baltimore:Association for Computational Linguistics,2014:25-35.
[9] Ji Y,Eisenstein J. Representation learning for text-level discourse parsing[C]//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Baltimore:Association for Computational Linguistics,2014:13-24.
[10] Kong F,Wang H L,Zhou G D. A CDT-Styled end-to-end Chinese discourse parser[J]. ACM Transactions on Asian and Low-Resource Language Information Processing,2017,16(4):387-398.
[11] Zhang L Y,Xing Y Q,Kong F,et al. A Top-Down neural architecture towards text-level parsing of discourse rhetorical structure[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online:Association for Computational Linguistics,2020:6386-6395.
[12] Xu S,Li P F,Zhou G D,et al. Employing text matching network to recognise nuclearity in Chinese discourse[C]// Proceedings of the 27th International Conference on Computational Linguistics. Online: Association for Computational Linguistics,2018:525-535.
[13] 王体爽,李培峰,朱巧明. 基于门控记忆网络的汉语篇章主次关系识别方法[J]. 中文信息学报,2019,33(5):39-46.
[14] 孙振华,周懿,朱巧明,等. 基于篇章主题的中文宏观篇章主次关系识别方法[J]. 中文信息学报,2020,34(12):30-38.
[15] Mikolov T,Sutskever I,Chen K,et al. Distributed representations of words and phrases and their compositionality[C]// Conference on Neural Information Processing Systems. Lake Tahoe:Neural Information Processing Systems Foundation,2013: 3111-3119.
[16] Devlin J,Chang M W,Lee K,et al. Bert:Pre-training of deep bidirectional transformers for language understanding[C]// Proceedings of NAACL-HLT 2019. Minneapolis:Association for Computational Linguistics,2019:4171-4186.
[17] Brown T,Mann B,Ryder N,et al. Language models are few-shot learners[C]//Conference on Neural Information Processing Systems. Vancouver:Neural Information Processing Systems Foundation,2020:1877-1901.
[18] 杨紫怡,贡正仙,孔芳,等. 基于中英文可比较语料的中文零指代消解[J]. 北京大学学报(自然科学版),2017,53(2): 279-286.
[19] Wang L Y,Tu Z P,Zhang X J,et al. A novel approach to dropped pronoun translation[C]//North American Chapter of the Association for Computational Linguistics. San Diego:Association for Computational Linguistics,2016:983-993.
[20] Wang L Y,Tu Z P,Zhang X J,et al. A novel and robust approach for pro-drop language translation[J]. Machine Translation, 2017,31(1):65-87.
[21] Wang L Y,Tu Z P,Shi S M,et al. Translating pro-drop languages with reconstruction models[C]//Proceedings of the AAAI Conference on Artificial Intelligence. New Orleans:AAAI Press,2018:4937-4945.
[22] 冯文贺,李艳翠,任函,等. 汉英篇章结构平行语料库的对齐标注评估[J]. 中文信息学报,2017,31(3):86-93.
[23] 张龙印. 融合多层次知识的篇章修辞结构解析研究[D]. 苏州:苏州大学,2022.
[24] Pires T,Schlinger E,Garrette D. How multilingual is multilingual BERT? [C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence:Association for Computational Linguistics,2019:4996-5001.
[25] Conneau A,Khandelwal K,Goyal N,et al. Unsupervised cross-lingual representation learning at scale[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics, 2020:8440-8451.
[26] Vaswani A,Shazeer N,Parmar N,et al. Attention is all you need[C]//31st Conference on Neural Information Processing Systems. Long Beach:Neural Information Processing Systems Foundation,2017:5998-6008.
[27] 冯文贺. 汉英篇章结构平行语料库的对齐标注研究[J]. 中文信息学报,2013,27(6):158-164.
[28] Xue L,Constant N,Roberts A,et al. Mt5:A massively multilingual pre-trained text-to-text transformer[C]//North American Chapter of the Association for Computational Linguistics. Online:Association for Computational Linguistics,2021:483-498.

[责任编辑:黄　敏]