DOI:10.19651/j.cnki.emt.2417748

考虑多特征的英语篇章关系分析识别方法*

张 静[1] 宗 欣[2] 郑 渊[3]
(1.山西医科大学汾阳学院 吕梁 032200;2.北京理工大学国际交流合作处 北京 100086;3.太原理工大学计算机科学与技术学院(大数据学院) 晋中 030600)

摘 要:为解决篇章级关系分析识别领域中存在多实体多匹配类别、上下文关系信息复杂等缺陷,本文提出一种融合实体特征信息和上下文特征信息的考虑多特征的英语篇章关系分析识别方法。首先,提出考虑多特征的关系分析识别框架。然后,对其中实体识别单元和上下文关系识别单元的机制进行详细介绍。最后,通过公开数据集和自选数据集进行对比实验和消融实验,验证并分析本文模型的优越性,并对模型识别效率以及相关参数的影响进行实验和分析。在 2 个数据集上,本文所提方法在 F_1 和 $Ign\ \ F_1$ 两个指标上均能保持最优性能,较其他次优识别模型在 F_1 指标上分别提升 1.65%和 1.13%,在 $Ign\ \ F_1$ 指标上分别提升 2.78%和 1.58%。实验表明:本文所提模型能够提取出表征篇章关系的关键特征信息,帮助理解篇章脉络以及各部分之间的关系,把握文章整体结构。

关键词:英语;篇章关系;分析识别;多特征;上下文关系
中图分类号:TP183;TN99 **文献标识码**:A **国家标准学科分类代码**:510.40

English discourse structure analysis recognition method considering multiple features

Zhang Jing[1] Zong Xin[2] Zheng Yuan[3]
(1. Fenyang College, Shanxi Medical University, Lyuliang 032200, China; 2. International Student Center, Beijing Institute of Technology, Beijing 100086, China; 3. College of Computer Science and Technology (College of Data Science), Taiyuan University of Technology, Jinzhong 030600, China)

Abstract: In order to solve the problems of multi-entity and multi-matching categories and complex context information in the field of discourse level relation analysis and recognition, this paper proposes an English discourse structure analysis and recognition method considering multi-features by fusing entity feature information and context feature information. Firstly, a structural analysis and recognition framework considering multiple features is proposed. Then, the mechanisms of entity recognition unit and context relation recognition unit are introduced in detail. Finally, comparative experiments and ablation experiments are carried out through public datasets and self-selected datasets to verify and analyze the superiority of the proposed model, and the recognition efficiency of the model and the influence of related parameters are experimented and analyzed. On the two datasets, the proposed method can maintain the optimal performance in both indicators F_1 and $Ign\ \ F_1$. Compared with other suboptimal recognition models, the proposed method improves 1.65% and 1.13% respectively on F_1, 2.78% and 1.58% respectively on $Ign\ \ F_1$. Experiments show that the proposed model can extract the key feature information representing the discourse relation, help to understand the context and the relationship between each part of the text, and grasp the overall structure of the text.

Keywords: English; discourse structure; analysis recognition; multi-feature; context relation

0 引 言

篇章关系分析识别任务是通过分析并提取篇章中实体对的关系,以此能够方便梳理出篇章脉络以及各部分之间的关系,有助于把握文章整体结构以及对语义进行更深层次的理解,在文本理解、文本生成、知识图谱构建等自然语言处理领域应用广泛[1-3]。相对于传统基于句子层次的关系信息抽取,篇章层次的关系信息提取更能够贴近实际应

收稿日期:2024-12-28
* 基金项目:2024 年山西省高等学校一般性教学改革创新立项项目(J20241693)资助

用,提取效果也相对较好[4-6]。

国内外针对英语篇章关系分析识别领域已有一定进展,主要是利用神经网络或者深度学习技术,构建篇章结构与实体、上下文关系、知识图谱、句法等信息的非线性映射关系。从当前研究内容看,篇章语料库主要包括基于层次化篇章关系的语料库[7]、基于扁平化篇章关系的语料库[8]等内容。王俊等[9]基于双向长短期记忆网络模型,考虑上下文信息,对篇章中时序关系进行分析与识别。Kurfali等[10]利用外部信息监督机制,对篇章关系隐藏信息进行分析识别。Guo 等[11]基于论元编码机制,结合双向长短期记忆网络和卷积神经网络模型,考虑动态分块的最大池化模型,对篇章关系进行识别。

上述研究在解决英语篇章关系分析识别领域有一定适用性,但存在两个关键问题没有被考虑:1)篇章中多实体多匹配类别问题。在英语篇章关系分析中,大概率存在多个实体,且相同的实体与其他不同的实体进行匹配,存在匹配类别多样的问题。2)上下文关系信息复杂。与句子级别的关系提取相比,篇章级别的关系更为复杂,一方面,实体对之间的关系存在于上下文中,单纯对实体对信息进行匹配,难以获得准确结果。另一方面,篇章中上下文信息多样,若对多样化的上下文信息不加筛选进行使用,会在模型中引入噪声干扰。

为解决上述缺陷,本文引入人工智能技术[12-13],融合实体特征和上下文信息特征对篇章级别的关系进行分析识别,为了解决篇章中多实体多匹配类别问题,设计实体识别模型,筛选出能够表征篇章特征的实体特征,去除冗余实体信息。为了解决上下文关系信息复杂问题,将实体对提及进行相互组合,并对篇章进行区分,基于 max-pooling 机制对多样化的上下文信息进行提取,最后利用注意力机制提取关键上下文信息。

本文主要贡献如下:1)设计实体识别模型,筛选出能够表征篇章特征的实体特征,去除冗余实体信息,解决篇章中多实体多匹配类别问题;2)根据实体对提及的位置,提取出关键的上下文信息,解决上下文关系信息复杂问题;3)利用公开数据集和自选数据集,进行大量实验并进行分析,验证本文模型的优越性。

1　考虑多特征的关系分析识别框架

英语篇章关系分析主要目的是理解英文篇章之间的整体结构以及各段落之间的语义关联,本文提出的考虑多特征的关系分析识别框架主要包含实体提取单元(entity extraction unit, EEU)和上下文关系提取单元(context relation extraction unit, CREU)两个模块,如图 1 所示。

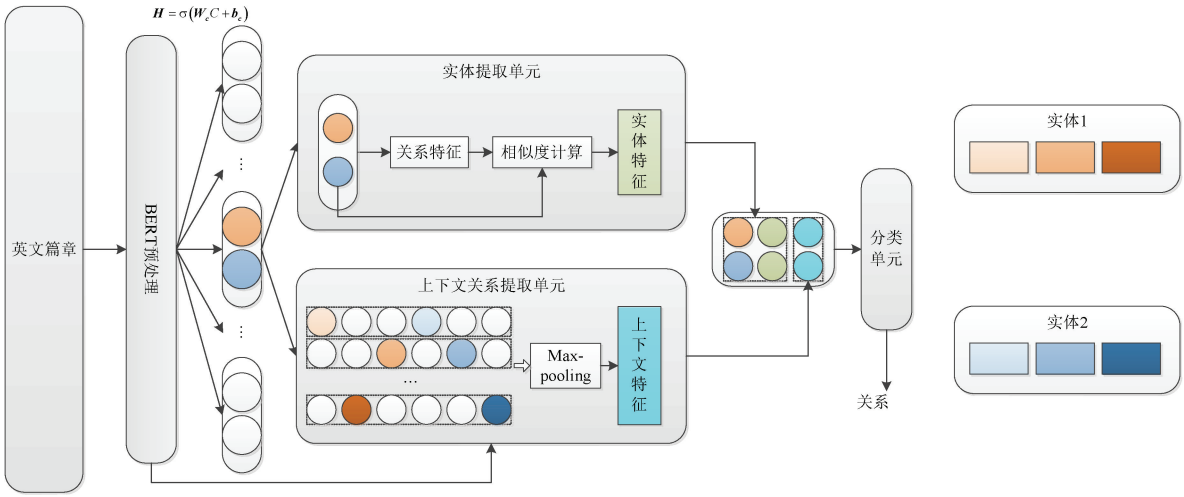

图 1　考虑多特征的关系分析识别框架

Fig. 1　Relationship analysis and recognition framework considering multiple features

如图 1 所示,实体提取单元主要是对英语篇章的实体特征进行提取,对篇章中存在的大量的冗余实体进行去除,获得关键实体。框架中专门设置实体提取单元,主要考虑如下:一方面由于固定的实体和其他不同实体进行匹配,经常会获得不同的结构关系,就表明该实体包含的特征信息较少,不足以识别出稳定的结构关系。另一方面,结合结构关系类别,对实体进行筛选,能够避免实体匹配过程中出现的多类别问题。

上下文关系提取单元主要是对英语篇章的上下文特征进行提取,框架中专门设置上下文关系提取单元,主要考虑如下:单独依赖实体确定英语篇章的结构关系具有一定局限性,上下文信息能够为识别模型提供重要的时序和因果信息,对结构关系的确定也具有重要的辅助作用。

1.1　BERT 预训练模型

BERT(bidirectional encoder representations from transformers)是一种预训练模型,该方法通过将双向变压器编码网络结构,能够获取文本数据中的语义信息[14]。将文本数据转化为向量序列,便于后续实体特征和上下文特

征提取。BERT 层中，输入语句中前置[CLS]，后置[SEP]，然后将字符的字、词和位置向量进行分别叠加处理，获得最后的数据向量，BERT 层内部机理如图 2 所示，该预训练方法能够充分挖掘出文本数据的信息，提升模型知识提取效率。

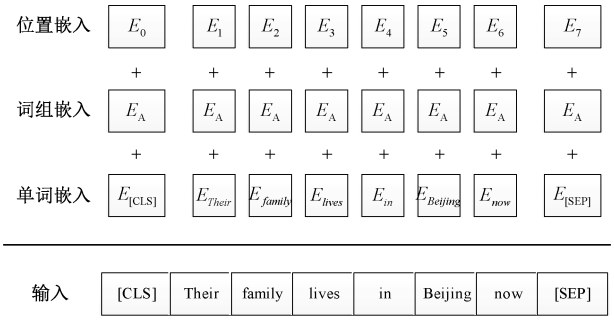

图 2　BERT 层内部机理

Fig. 2　Internal mechanism of BERT layer

针对篇章中包含 n 个标识的单词 $\boldsymbol{Doc}=[c_1,c_2,\cdots,c_n]$，利用 BERT 对全部文章编码，过程如下：

$$[\boldsymbol{CLS},\boldsymbol{C}]=BERT([c_1,c_2,\cdots,c_n]) \tag{1}$$

$$\boldsymbol{H}=\sigma(\boldsymbol{W}_e\boldsymbol{C}+\boldsymbol{b}_e) \tag{2}$$

式中：c_i 为单词表示，$\boldsymbol{CLS}$ 为篇章特征，$\boldsymbol{C}$ 为标识编码处理后的隐藏层向量，$\boldsymbol{W}_e$ 和 $\boldsymbol{b}_e$ 分别为权重和阈值，$\sigma(\cdot)$ 为激活函数，$\boldsymbol{H}$ 为单词标识经过变换后的隐藏层向量，用 $[\boldsymbol{h}_1,\boldsymbol{h}_2,\cdots,\boldsymbol{h}_n]$ 表示。

然后利用实体识别单元和上下文关系识别单元对 $\boldsymbol{H}$ 进行处理，实现 BERT 与 EEU 和 CREU 模块进行集成。同时，BERT 预训练模型通过将单词向量化，能够全面描述篇章中实体以及实体间的关系特征，并且能不错地表示篇章中上下文的语义信息，提升整体识别的性能。

1.2　实体识别单元

实体识别单元主要是对实体对信息进行选择，确定关键实体信息。在篇章关系分析识别过程中，同一个实体与其他不同的实体会存在匹配类别多样，或者同一个实体与多个实体匹配类别一致，表明实体对中的实体信息量不一致，存在差异较大，因此需要结合实体特征，有针对性设计出实体识别单元，提取关键特征信息。实体识别单元的具体结构如图 3 所示，其训练过程主要分两个步骤：

步骤 1)实体关系特征提取。主要是对英语篇章中的实体关系特征进行粗略分析和提取。

步骤 2)相似度计算。根据相似度高的实体能够包含更多关系信息的原则，本文将相似度低的实体进行删除，提取关键实体特征。

1)实体关系特征提取

由于英语篇章相关文本数据具有非结构化、数据规模大等特点，传统的长短期记忆网络、隐马尔科夫模型等知识识别方法已经难以适应，导致知识提取效率偏低，因此本文考虑到数据文本的特点，兼顾模型泛化和提取效率，

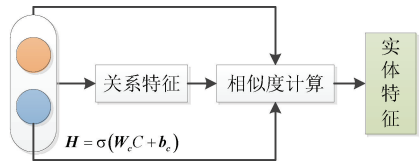

图 3　实体识别单元结构

Fig. 3　Structure of entity recognition unit

将引入注意力机制的双向门控循环神经网络(bidirectional gated recurrent unit-attention，BiGRU-attention)和条件随机场(conditional random field，CRF)结合起来，构成 BiGRU-Attention-CRF 模型，对英文篇章中的实体特征进行识别。该模型在文本数据描述表示、语义信息学习以及最佳标注序列获取等方面均有一定优势，主要有 BiGRU-Attention 层以及 CRF 层，具体框架如图 4 所示。

如图 4 所示，BiGRU-Attention 层以及 CRF 层具体描述如下：

(1)BiGRU-Attention 层。BiGRU 网络由正向和反向的 GRU 网络构成，BERT 层获得的输入数据在两层 GRU 网络中分别计算，然后进行整合获得最终的输出数据。GRU 网络结构能够兼顾模型的性能和计算的复杂度，参数较为简单，能有效避免过拟合缺陷，如图 5 所示。Attention 是将注意力机制引入到 BiGRU 网络中，根据网络中数据的重要程度分别赋予差异化的权重值，引导网络模型更加注重关键环节与数据，以此提高网络模型识别准确率，其中激活函数使用 Softmax 函数。权重值 $Attention(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})$ 表示为：

$$Attention(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=Softmax\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{3}$$

式中：$\boldsymbol{Q}$、$\boldsymbol{K}$ 和 $\boldsymbol{V}$ 分别表示查询、键和代表值，通过计算 $\boldsymbol{Q}$ 和 $\boldsymbol{K}$ 的相似度得到的权重与 $\boldsymbol{V}$ 进行加权求和处理，获得最终的权重值。

(2)CRF 层[15]。CRF 层能够通过对数据序列解码标注，从而获得最佳标注序列。其中目标函数通过考虑输入数据和标签的状态特征，可以更加准确提取出数据的特征以及标签的信息流转原理。

2)相似度计算

相似度是通过对多源数据信息进行整合处理，消除数据冗余性，确保识别出关键实体特征信息。通过实体识别处理后，仍会出现多词以及同义等现象，需要利用知识融合对知识进行统一处理，提升实体特征信息识别质量。本文基于 Jaccard 相关系数对实体之间相似度进行计算[16]，计算公式为：

$$Sim(s_1,s_2)=\frac{|A(s_1)\cap A(s_2)|}{|A(s_1)\cup A(s_2)|} \tag{4}$$

式中：s_1,s_2 分别表示英语篇章中的实体，$A(s)$ 表示实体 s 的属性字符串，实体之间的相似度与 Sim 数值成正相关，通过设置阈值判断两个实体是否能够融合为同一个。

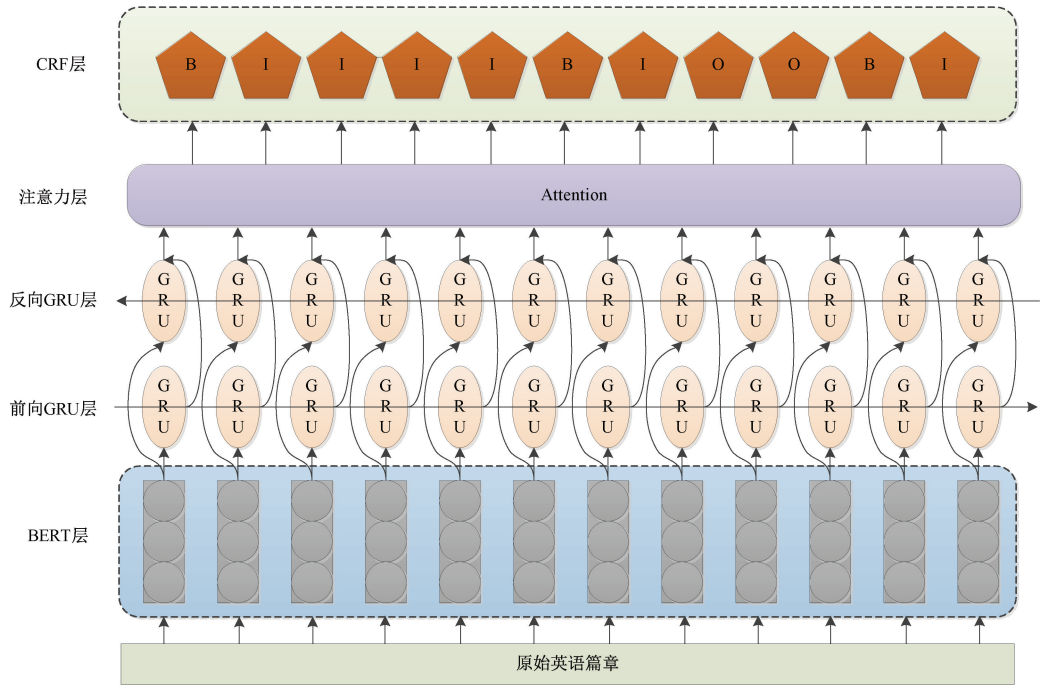

图 4　基于 BERT-BiGRU-Attention-CRF 模型的实体识别框架

Fig. 4　Entity recognition framework based on ERT-BiGRU-Attention-CRF model

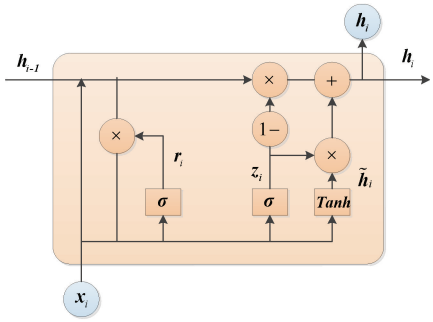

图 5　GRU 网络结构

Fig. 5　GRU network structure

1.3　上下文关系识别单元

上下文关系识别单元主要是对英语篇章中的实体关系特征进行分析与提取,避免冗余的噪声实体对信息干扰的识别结果。其基本原理为:首先将实体对的提及两两组合,然后根据提及的位置将英语篇章切分,利用 *max-pooling* 操作提取篇章中不同的上下文信息,最后利用 Attention 机制提取关键的上下文信息。

上下文关系识别单元的步骤如下:

1)实体组合。对实体对 (e_i,e_j) 在英语篇章中的所有提及进行相互组合,若实体 $e_i=[m_i^1,m_i^2,\cdots,m_i^{k_1}]$ 和 $e_j=[m_j^1,m_j^2,\cdots,m_j^{k_2}]$ 分别表示在文中提及次数为 k_1 和 k_2,两两组合次数为 $k_1\times k_2$。

2)关系信息计算。根据提及对在英语篇章中的位置,将英语篇章区分成 3 部分,基于 *max-pooling* 操作对实体对关系进行计算,即取局部接受域中值最大的点,主要是减少空间维度,降低计算量和参数数量,同时保留重要的特征信息,公式如下:

$$\boldsymbol{c}'_i=[\max(h_{0:p}),\max(h_{p:q}),\max(h_{q:n})] \tag{5}$$

$$\boldsymbol{c}_i=\tanh(\boldsymbol{W}_{c1}\boldsymbol{c}'_i+b_{c1}) \tag{6}$$

式中:p,q 分别为实体提及对在英语篇章中的位置,$\boldsymbol{c}_i$ 为次序为 i 的提及对表示的上下文关系信息,$\boldsymbol{W}_{c1}\in\mathbb{R}^{3d\times d}$,$b_{c1}\in\mathbb{R}^d$ 分别表示模型中的参数和阈值。由于实体提及对在英语篇章中的位置存在差异,因此可以对英语篇章中存在差异的上下文特征信息进行提取,基于 Attention 机制对英语篇章中的提及对上下文关系信息进行提取,即无需复杂的特征学习机制,可以自动从数据中学习特征,降低训练量,公式如下:

$$\begin{cases}\boldsymbol{c}=\sum_i\beta_i\boldsymbol{c}_i\\ \beta_i=softmax(\boldsymbol{W}_{c2}\boldsymbol{c}_i)\end{cases} \tag{7}$$

式中:β_i 为提及对上下文关系信息权重,$\boldsymbol{W}_{c2}$ 为模型权重,$\boldsymbol{c}$ 为提及对上下文关系信息。

3)信息融合。将识别出的实体特征信息和上下文关系特征信息进行融合处理,计算出综合的实体对关系信息,公式如下:

$$\begin{cases}\boldsymbol{z}_i=\mathrm{FFNN}([\boldsymbol{e}_i,\boldsymbol{e}_r,\boldsymbol{c}])\\ \boldsymbol{z}_j=\mathrm{FFNN}([\boldsymbol{e}_i,\boldsymbol{e}_r,\boldsymbol{c}])\end{cases} \tag{8}$$

式中:$\boldsymbol{z}_i,\boldsymbol{z}_j\in\mathbb{R}^d$ 为实体对关系特征信息向量,$\boldsymbol{e}_r$ 为关键实体关系向量,FFNN 为前馈神经网络模型。

1.4　分类单元

分类单元主要是对给定的实体对特征信息进行匹配,

获得类别信息，若实体对特征信息为 $(\boldsymbol{z}_i,\boldsymbol{z}_j)$，则可基于双线性变换操作，计算实体对特征信息的关系概率，公式如下：

$$P(r\mid \boldsymbol{e}_i,\boldsymbol{e}_j)=\sigma(\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{W}_r\boldsymbol{z}_j+\boldsymbol{b}_r) \tag{9}$$

式中：$\boldsymbol{W}_r\in\mathbb{R}^{d\times d}$ 和 $b_r\in\mathbb{R}^d$ 为模型参数，σ 表示 Sigmoid 激活函数。

为降低双线性变换操作计算的复杂度，基于双线性组将隐藏层划分为不同的组，数量为 l，公式如下：

$$\begin{cases}\boldsymbol{z}_i=[z_i^1,z_i^2,\cdots,z_i^l]\\ \boldsymbol{z}_j=[z_j^1,z_j^2,\cdots,z_j^l]\\ P(r\mid \boldsymbol{e}_i,\boldsymbol{e}_j)=\sigma\left(\sum_n^l(\boldsymbol{z}_i^n)^{\mathrm{T}}\boldsymbol{W}_r^n\boldsymbol{z}_j^n+\boldsymbol{b}_r^n\right)\end{cases} \tag{10}$$

式中：$\boldsymbol{W}_r^n\in\mathbb{R}^{\langle d/l\rangle\times\langle d/l\rangle}$，$b_r^n\in\mathbb{R}^{\langle d/l\rangle}$ 为模型参数。此外，将文献[17]中的自适应阈值机制引入到本文的训练模型中，对损失函数进行定义，并加入 Dropout 函数，避免模型在训练过程中过拟合，提升模型泛化性能。

2　实验

为验证本文所提的考虑多特征的英语篇章关系分析识别方法的有效性与优越性，分别以公开的数据集和收集的英语篇章关系信息为研究对象，进行对比实验和消融实验，并对实验结果进行分析。

2.1　数据集

选择公开的 DocRED 数据集[18]和从 Nature 期刊上筛选出的英语片段作为研究对象，2 个数据集信息如表 1 所示，数据集具有多关系标签的特征。

表 1　数据集信息

Table 1　Information of the dataset

数据集	来源	规模	关系种类	性质
数据集 1	DocRED 数据集	5 053	97	公开
数据集 2	Nature 期刊	485	8	收集

2.2　参数设置

本文在 2 个数据集上的模型参数如表 2 所示。

本文选择 F_1 作为模型的识别性能指标，该指标是精确率指标 *Precision* 和召回率指标 *Recall* 的调和平均数，公式如下：

$$F_1=2\times\frac{Precision\times Recall}{Precision+Recall} \tag{11}$$

此外，为了解决数据集中存在实体对重复的现象，更真实展示模型的识别性能，引入 $Ign\ \ F_1$ 指标，表示去掉训练集和测试集中的共享实体度对的 F_1。

2.3　基线模型

为验证本文所提算法的优越性，将模型与一些较新的基线识别模型进行比较，如下：

表 2　模型参数

Table 2　Parameters of the model

参数	说明或数值	
	数据集 1	数据集 2
开发环境	Tensorflow 1.4.1	Tensorflow 1.4.1
GPU	GeForce GTX 2080Ti	GeForce GTX 2080Ti
内存	16 G	16 G
混合精度训练	APEX 混合精度模块	APEX 混合精度模块
优化器	AdamW	AdamW
批次大小	4	4
l	64	64
迭代次数	30	20
学习率	5×10^{-5}	5×10^{-5}
Dropout	0.1	0.1
分组规模	64	32
梯度裁剪	1.0	1.0

1）HIN-GloVe/HIN-BERT[19]：基于层次推理网络模型，选择 GloVe 或 BERT 作为预训练模型，利用句子、篇章的多元化信息进行篇章关系识别。

2）BERT-FFNN[20]：选择 BERT 作为预训练模型，并利用前馈神经网络对篇章关系进行识别。

3）BiLSTM-RACNN[21]：结合双向长短时记忆网络和循环注意力卷积神经网络，捕捉篇章中的全局和局部特征，对篇章关系进行识别。

4）BMGF-RoBERTa[22]：结合上下文信息、多匹配机制以及全局信息等特征，实现深层次篇章关系挖掘识别。

5）BiLSTM-CNN[11]：结合双向长短时记忆网络和卷积神经网络，引入动态块 max pooling，实现篇章关系识别。

此外，一些经典的深度学习方法也可用来对篇章关系进行识别，比如 CNN、LSTM、Bi-LSTM 等，一并用来作为基线模型。

2.4　对比实验结果与分析

以 2 个数据集为研究对象，将本文所提的考虑多特征的英语篇章关系分析识别方法与基线模型分别进行对比实验，结果如表 3 所示。

可以看出：在 2 个数据集上，本文所提的考虑多特征的英语篇章关系分析识别方法在 F_1 和 $Ign\ \ F_1$ 两个指标上均能保持最优性能，较其他次优识别模型在 F_1 指标上分别提升 1.65%和 1.13%，在 $Ign\ \ F_1$ 指标上分别提升 2.78%和 1.58%。表明实体对特征和上下文特征能够代表篇章中关键特征信息，本文所提算法能够发挥考虑多特征优势，充分挖掘数据的隐藏信息，识别性能优越。

表 3　对比实验结果

Table 3　compares the experimental results

模型	数据集 1		数据集 2	
	F_1 /%	$Ign\ F_1$ /%	F_1 /%	$Ign\ F_1$ /%
CNN	49.85	47.34	78.58	76.41
LSTM	53.47	51.24	83.54	81.74
Bi-LSTM	56.54	54.87	86.47	84.87
HIN-GloVe	58.35	56.46	87.74	85.35
HIN-BERT	59.57	57.14	88.21	86.94
BERT-FFNN	61.24	59.81	90.52	88.47
BiLSTM-RACNN	60.45	58.74	89.77	87.15
BiLSTM-CNN	58.47	56.24	87.83	85.24
本文所提算法	**62.25**	**61.47**	**91.54**	**89.87**

2.5　消融实验结果与分析

为验证本文所提的考虑多特征的英语篇章关系分析识别框架中各模块单元的正向引导作用,以此说明各模块的有效性,在 2 个数据集上分别进行消融实验,同时为说明各模块对模型计算效率的影响,增加评价指标 T,表示模型的运行时间。结果如表 4 所示。

根据消融实验结果,可以看出:

1)识别框架中没有上下文关系提取单元时,在数据集 1 上,F_1 和 $Ign\ \ F_1$ 两个指标分别下降了 1.45%和 1.51%,在数据集 1 上,F_1 和 $Ign\ \ F_1$ 两个指标分别下降了 1.30%和 1.56%。表明上下文关系提取单元对识别性能有正向引导作用。究其原因,在英语篇章中,篇章关系特征不仅仅存在于实体对中,其存在位置往往不固定,忽略对上下文关系的考虑会导致特征信息提取不完整。

表 4　消融实验结果

Table 4　Results of ablation experiments

基线	实体提取单元	上下文关系提取单元	数据集 1			数据集 2		
			F_1 /%	$Ign\ \ F_1$ /%	T/s	F_1 /%	$Ign\ \ F_1$ /%	T/s
√	√		61.35	60.54	493.4	90.35	88.47	48.5
√		√	61.04	59.87	2965.2	89.57	88.15	274.6
√	√	√	**62.25**	**61.47**	**3147.4**	**91.54**	**89.87**	**298.7**

2)识别框架中没有实体提取单元时,在数据集 1 上,F_1 和 $Ign\ \ F_1$ 两个指标分别下降了 1.94%和 2.60%,在数据集 1 上,F_1 和 $Ign\ \ F_1$ 两个指标分别下降了 2.15%和 1.91%。表明实体提取单元对识别性能同样有正向引导作用,且在识别框架中实体提取单元发挥的作用较上下文关系提取单元更大。究其原因,在英语篇章中,实体对是匹配篇章类别的关键信息,且实体对存在多个匹配类别,存在冗余信息,因此准确地识别实体对能够提高模型识别准确率。

3)单独添加上下文关系提取单元的实验运行时间与全模块添加的实验运行时间处于一个数量级,基本相差不大,但去除上下文关系提取单元后实验运行时间大大降低,其原因主要是由于上下文关系提取单元中存在循环抽取实体对提及的模块,会增加运行时间。

2.6　识别效率分析

为了对本文所提的考虑多特征的英语篇章关系分析识别方法的计算效率进行分析,选择模型的运行时间 T、模型参数规模 P 以及模型解码时间 T_d 作为指标,实验结果如表 5 所示。

可以看出:本文所提的识别模型在运行时间 T、模型参数规模 P 以及模型解码时间 T_d 3 个指标上的数值规模均不算大,相应对运行设备性能要求不算太高。同时本文所提的识别模型在 3 个指标上数值均比其他算法大,其原因是由于上下文关系提取单元中存在循环计算模块,增加运行时间。

表 5　识别效率实验结果

Table 5　Experimental results of identification efficiency

识别模型	数据集 1			数据集 2		
	T/s	P	T_d /s	T/s	P	T_d /s
HIN-GloVe	2657.7	116.8	789.6	254.7	12.5	78.1
HIN-BERT	2835.8	132.6	813.5	274.3	14.3	83.9
BERT-FFNN	2962.4	133.4	829.0	287.6	15.7	84.7
本文所提模型	**3147.4**	**145.5**	**887.4**	**298.7**	**16.5**	**85.4**

2.7　参数分析

1)噪声参数对识别结果影响

噪声参数是模型在识别过程中不可避免的干扰因素,模型的抗外界干扰能力直接影响到模型的应用价值,为测试噪声参数对识别结果的影响,验证模型的抗噪声能力,本文区分实体对特征噪声(实验组 1)和上下文关系特征噪声(实验组 2)两种类型,设计 6 组实验,具体信息如表 6 所示。需要说明的是,本文认为噪声是篇章中的固有属性,无需自行添加,并认为本文采取的筛选后的实体对信息、提取上下文特征确定关系信息为去噪后的准确信息,实验 1～3 中天然存实体对的特征噪声,实验 4～6 中天然存在上下文关系特征噪声。

以 F_1 和 $Ign\ \ F_1$ 作为评价指标,以 2 个数据集为研究对象,6 组实验结果如表 7 所示。

表 6　噪声干扰对比实验

Table 6　Comparison experiment of noise interference

实验编号	实验组 1:实体对特征噪声类型	实验编号	实验组 2:上下文关系特征噪声类型
1	单独使用头实体信息	4	BERT 编码后的输出的第一个 CLS 作为上下文信息
2	单独使用为尾实体信息	5	实体对所在篇章句子作为上下文信息
3	使用全部实体对信息	6	不包含实体对的篇章句子作为上下文信息

表 7　噪声参数对识别结果影响

Table 7　Influence of noise parameters on recognition results

实验编号	数据集 1			数据集 2		
	F_1/%	Ign	F_1/%	F_1/%	Ign	F_1/%
1	61.58		60.47	90.15		88.41
2	60.24		59.24	89.14		87.35
3	61.85		60.93	90.46		88.75
4	60.34		59.74	89.34		87.64
5	61.85		60.29	90.74		88.34
6	59.41		58.97	88.14		86.47
本文模型	**62.25**		**61.47**	**91.54**		**89.87**

根据实验 1～3 结果,可以看出:

(1)实验 1 和实验 2 中的识别模型性能在实验组 1 中最差,表明单独使用头实体或尾实体信息时,识别性能有明显下降,不加筛选的单一实体表征的特征信息不足,对识别模型有错误引导。

(2)实验 3 中的识别模型性能位于实验 1、2 和本文模型之间,表明使用筛选后的实体对信息比使用全部实体对信息识别性能更优,表明实体对中存在噪声干扰,需要有针对性地对噪声进行过滤,筛选出最能表征特征信息的实体对信息。

根据实验 4～6 结果,可以看出:

(1)实验 4 和 6 中的识别模型性能在实验组 2 中最差,表明英语篇章中上下文关系中存在噪声干扰,需要对噪声进行处理,提高识别性能。

(2)实验 5 中的识别模型性能较实验 4 和 6 好,表明实体对所在篇章句子作为上下文信息,能够对上下文关系噪声进行一定程度的过滤,提供了更多表征特征的信息。

(3)实验 5 中的识别模型性能较本文模型差,其原因如下:一方面,由于实体对的核心信息位置不固定,有可能存在实体对所在篇章句子中,也有可能存在篇章其他位置。另一方面,本文模型对上下文关系信息进行了赋权,避免了冗余实体对提及信息对识别模型的影响。

2)上下文信息提取方法对识别结果影响

准确的上下文信息信息提取方法能够帮助模型表征更多的特征信息,提高模型识别性能,本文为验证不同上下文信息提取方法对识别结果的影响,凸显本文所提的基于注意力机制的上下文特征提取方法的优势,设计不同的实验对照,具体信息如表 8 所示。需要说明的是,本文模型认为实体提及对具有不同的重要性,通过赋权获得关键关系信息。

表 8　上下文信息提取方法对识别结果影响实验

Table 8　The context information extraction method affects the recognition results

实验编号	上下文信息提取方法
实验 7	随机选择实体提及对作为上下文信息
实验 8	对实体提及对进行平均加权处理,作为上下文信息
实验 9	BERT 编码后的输出的第一个 CLS 作为上下文信息

以 F_1 和 Ign　F_1 作为评价指标,以 2 个数据集为研究对象,3 组实验结果如表 9 所示。

表 9　不同上下文信息提取方法实验结果

Table 9　Experimental results of different context information extraction methods

实验编号	数据集 1			数据集 2		
	F_1/%	Ign	F_1/%	F_1/%	Ign	F_1/%
实验 7	60.47		59.47	89.34		87.12
实验 8	61.57		60.84	90.47		88.74
实验 9	60.04		58.96	88.87		86.65
本文模型	**62.25**		**61.47**	**91.54**		**89.87**

根据实验 7～9 结果,可以看出:

(1)实验 9 中识别模型性能最差,其原因是在英语篇章中,没有区分实体对上下文信息的差异,所有实体对拥有一致的上下文信息,使得识别效果差。

(2)实验 7 中识别模型性能较实验 8 和本文模型差,但比实验 9 性能要好,其原因是随机选择实体提及对作为上下文信息有可能包含核心的上下文信息,但大多数情况可能不包含,导致信息有效性不强,使得性能变差。

(3)实验 8 中识别模型性能仅次于本文模型,其原因是对实体提及对进行平均加权处理作为上下文信息能够包含关键特征信息,但由于没有加以筛选,导致噪声比较多,使得识别效果一般。

3 结　论

为解决英语篇章关系分析识别问题,本文提出了融合了实体特征和上下文信息特征的多特征识别框架,该模型能够提取篇章级别中实体特征以及上下文信息特征,解决篇章中多实体多匹配类别以及上下文关系信息复杂问题。选择公开的 DocRED 数据集和从 Nature 期刊上筛选出的英语片段作为研究对象,通过大量实验验证了本文识别模型的优越性,实验表明本文所提模型能够提取出表征篇章关系的关键特征信息,识别准确率优于基线模型以及消融实验对比组。该模型为梳理出篇章脉络以及各部分之间的关系,把握文章整体结构以及对语义进行更深层次的理解提供参考。

参考文献

[1] ZHANG M SH, LI ZH H, FU G H, et al. Dependency-based syntax-aware word representations[J]. Artificial Intelligence, 2021, 292: 103427.

[2] LI J Q, LIU M, QIN B, et al. A survey of discourse parsing[J]. Frontiers of Computer Science, 2022, 16(5): 165329.

[3] 蒋峰,范亚鑫,褚晓敏,等. 英汉篇章结构分析研究综述[J]. 软件学报,2023,34(9):4167-4194.
JIANG F, FAN Y X, CHU X M, et al. Survey on English and Chinese discourse structure analysis[J]. Journal of Software,2023,34(9):4167-4194.

[4] 周佳伦,李琳宇,马洪彬,等. MRC-PBM:一种中文电子病历嵌套命名实体识别方法[J]. 国外电子测量技术,2024,43(1):159-165.
ZHOU J L, LI L Y, MA H B, et al. MRC-PBM: A Chinese electronic medical record nested name dentity recognition method [J]. Foreign Electronic Measurement Technology,2024,43(1):159-165.

[5] 岳琳,杨风暴,王肖霞. 基于 HRAGS 模型的混合式摘要生成方法[J]. 电子测量技术,2022,45(15):75-83.
YUE L, YANG F B, WANG X X. Hybrid summary generation method based on HRAGS model [J]. Electronic Measurement Technology,2022,45(15):75-83.

[6] 朱芳鹏,王晓峰. 面向船舶工业新闻的文本分类[J]. 电子测量与仪器学报,2020,34(1):149-155.
ZHU F P, WANG X F. Text classification for ship industry news[J]. Journal of Electronic Measurement and Instrumentation,2020,34(1):149-155.

[7] CARLSON L, MARCU D, OKUROWSKI M E. Building a discourse-tagged corpus in the framework of rhetorical structure theory [C]. Current and New Directions in Discourse and Dialogue. Dordrecht: Springer, 2003:85-112.

[8] PRASAD R, DINESH N, LEE A, et al. The penn discourse TreeBank 2.0[C]. Language Resources and Evaluation(LREC). Marrakech: European Language Resources Association, 2008:2961-2968.

[9] 王俊,史存会,张瑾,等. 融合上下文信息的篇章级事件时序关系抽取方法[J]. 计算机研究与发展,2021,58(11):2475-2484.
WANG J, SHI C H, ZHANG J, et al. Document-level event temporal relation extraction with context information[J]. Journal of Computer Research and Developmen,2021,58(11):2475-2484.

[10] KURFALI M, ÖSTLING R. Let's be explicit about that: Distant supervision for implicit discourse relation classification via connective prediction[C]. Workshop on Understanding Implicit and Underspecified Language. Stroudsburg: ACL, 2021: 1-10.

[11] GUO F Y, HE R F, DANG J W. Implicit discourse relation recognition via a BiLSTM-CNN architecture with dynamic chunk-based max pooling [J]. IEEE Access, 2019, 7:169281-169292.

[12] 张宏宏,甘旭升,李双峰,等. 复杂低空环境下考虑区域风险评估的无人机航路规划[J]. 仪器仪表学报,2021,42(1):257-266.
ZHANG H H, GAN X SH, LI SH F, et al. UAV route planning considering regional risk assessment under complex low altitude environment[J]. Chinese Journal of Scientific Instrument,2021,42(1):257-266.

[13] 张宏宏,李文华,郑家毅,等. 有人/无人机协同作战:概念、技术与挑战[J]. 航空学报,2024,45(15):168-194.
ZHANG H H, LI W H, ZHENG J Y, et al. Manned/unmanned aerial vehicle cooperative combat system: Concepts, technologies, and challenges[J]. Acta Aeronautica et Astronautica Sinica, 2024, 45(15):168-194.

[14] 张凤荔,黄鑫,王瑞锦,等. 基于 BERT 多知识图融合嵌入的中文 NER 模型[J]. 电子科技大学学报,2023,52(3):390-397.
ZHANG F L, HUANG X, WANG R J, et al. A Chinese NER model based on BERT with multi knowledge graph fusion and embedding[J]. Journal of University of Electronic Science and Technology of China,2023,52(3):390-397.

[15] 袁里驰. 基于 BiLSTM-CRF 的中文分词和词性标注联合方法[J]. 中南大学学报(自然科学版),2023,54(8):3145-3153.
YUAN L CH. A joint method for Chinese word segmentation and part-of speech tagging based on

BiLSTM-CRF[J]. Journal of Central South University (Science and Technology),2023,54(8):3145-3153.

[16] 武聪,马文明,王冰,等.融合用户标签相似度的矩阵分解算法[J].南京大学学报(自然科学),2022,58(1):143-152.
WU C, MA W M, WANG B, et al. Matrix factorization algorithm combined with user tag similarity[J]. Journal of Nanjing University(Natural Science),2022,58(1):143-152.

[17] ZHOU W X, HUANG K, MA T Y, et al. Document-level relation extraction with adaptive thresholding and localized context pooling[J]. ArXiv preprint arXiv: 2010.11304, 2020.

[18] YUAN Y, YE D M, LI P, et al. DocRED: A large-scale document-level relation extraction dataset[J]. ArXiv preprint arXiv: 1906.06127, 2019.

[19] TANG H ZH, CAO Y N, ZHANG Z Y, et al. HIN: Hierarchical inference network for document-level relation extraction[J]. ArXiv preprint arXiv: 2023.12754, 2020.

[20] 黄河燕,袁长森,冯冲.融合实体和上下文信息的篇章关系抽取研究[J].自动化学报,2024,50(10):1953-1962.
HUANG H Y, YUAN CH S, FENG CH. Document-level relation extraction with entity and context information[J]. Acta Automatica Sinica, 2024, 50(10):1953-1962.

[21] 王秀利,金方焱.融合特征编码和短语交互感知的隐式篇章关系识别[J].电子学报,2024,52(4):1377-1388.
WANG X L, JIN F Y. Implicit discourse relation recognition integrating feature coding and phrase interaction Perception[J]. Acta Electronica Sinica, 2024,52(4):1377-1388.

[22] LIU X, OU J F, SONG Y Q, et al. On the importance of word and sentence representation learning in implicit discourse relation classification[C]. Conference on Artificial Intelligence. Amsterdam: Elsevier, 2020: 3830-3836.

作者简介

张静,本科,实验师,主要研究方向为英语听力教学与管理。

E-mail:578381672@qq.com

宗欣(通信作者),硕士,助理研究员,主要研究方向为外国语言学及应用语言学。

E-mail:zongxin@bit.edu.cn

郑渊,硕士研究生,主要研究方向为大数据技术与工程。

E-mail:574529447@qq.com