计算机科学
COMPUTER SCIENCE

话题性话语标记的自动识别与分类

杨进才，余漠洋，胡满，肖明

引用本文

杨进才，余漠洋，胡满，肖明. 话题性话语标记的自动识别与分类[J]. 计算机科学，2025，52(4)：255-261.

YANG Jincai, YU Moyang, HU Man, XIAO Ming. Automatic Identification and Classification of Topical Discourse Markers [J]. Computer Science, 2025, 52(4): 255-261.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.240100155

话题性话语标记的自动识别与分类

杨进才[1]　余漠洋[1]　胡　满[1]　肖　明[2]
1 华中师范大学计算机学院　武汉 430079
2 华中师范大学语言与语言教育研究中心　武汉 430079

摘　要　话语标记(Discourse Markers)是一种语言标记,具有组织语篇、引导指意、显示情感的作用,因而受到语言学界的广泛关注。对话语标记及其类别的准确识别,对于篇章理解、说话人意图和情感的把握有重要作用。近十年来,国内外学者对话语标记的功能、特征、来源和系统分类展开研究并取得了丰富的成果。然而,因话语标记形式多变、来源多样、特征抽象、变体繁多,机器自动识别的难度较大。对此,以话题性话语标记为研究对象,提出一种融合外部语言学特征的 NFLAT 指针网络模型,实现对语篇中话语标记的自动识别和分类。经实验检验,训练后模型对话题性话语标记的识别及分类精确率(P 值)达 94.55%。

关键词:话语标记;语义增强;特征融合;自动识别与分类
中图分类号　TP391

Automatic Identification and Classification of Topical Discourse Markers

YANG Jincai[1], YU Moyang[1], HU Man[1] and XIAO Ming[2]
1 School of Computer Science, Central China Normal University, Wuhan 430079, China
2 Research Center for Language and Language Education, Central China Normal University, Wuhan 430079, China

Abstract　Discourse markers, a kind of linguistic markers at the pragmatic level which have functions of organizing discourse, guiding signifier, and expressing emotions, have attracted extensive attention in linguistics. The accurate identification of discourse markers and categories plays an important role in the comprehension of text and the grasp of the speaker's intention and emotion. In the past decade, scholars at home and abroad have conducted research on function, characteristics, sources and systematic classification of discourse markers and achieved rich results. However, due to the changeable forms, diverse sources, abstract features, and variants, it is difficult for machines to automatically identify discourse markers. In this paper, an NFLAT pointer network model integrating external linguistic features is proposed, which takes topical discourse markers as the research object, and realizes the automatic recognition and classification of discourse markers in discourse. Experimental results show that the precision of the trained model for the recognition and classification of topical discourse markers reaches 94.55%.

Keywords　Discourse marker, Semantic enhancement, Feature fusion, Automatic identification and classification

1　引言

话语标记译自术语"discourse marker",是语用层面的一种语言标记,初见于 20 世纪 50 年代西方英语语言学界。随着 70 年代语篇分析、语用学、会话分析等理论的建立和完善,话语标记的研究渐入佳境。经数十年的发展,进入 21 世纪后,对话语标记的研究保持上升趋势,国内发展势头迅猛,近 5 年来的研究文献量均维持在较高水平[1]。

话语标记作为一种语用标记,其作用主要表现在语篇的连接、言外之意的引导以及特殊感情的凸显 3 个方面,具有话语组织(篇章)功能、人际(人际交互)功能和元话语功能这三大功能[2]。话语标记可与词语、短语、小句同形,它与其他语言形式的关系在于其区别性的特征。因此,语言学家通过描述话语标记在语音、句法、语用和语义等方面的特征来对其进行定义。在语音层面,Liu[3]认为话语标记在语音上有弱化现象,可通过停顿、重音等将其从其他句法单位中识别出来。在句法层面上,其特征则主要体现在句法的可取消性、可分离性和句法位置的不确定性上[2]。在语义层面上,Xu[4]认为话语标记较少含有词汇意义或概念意义,Li[5]认为话语标记在发挥作为话语标记的作用时,并不表达概念语义。以上这些特征是我们判断准标记词(即候判标记词)是否发挥话语标记功能的重要依据。

话语标记词的形式丰富,是语言词汇化、语法化、语用化的结果,是一个半封闭的集合。部分话语标记词有数个变体,

到稿日期:2024-01-22　返修日期:2024-05-16
基金项目:国家社会科学基金(19BYY092);教育部人文社科规划基金(20YJA740047)
This work was supported by the National Social Science Fundation of China(19BYY092) and Humanity and Social Science Foundation of Ministry of Education of China(20YJA740047).
通信作者:杨进才(jcyang@mail.ccnu.edu.cn)

例如引导性话题标记“对X来说”的变体有“就X来说”和“对于X而言”等。同一语言片段有时充当话语标记，有时则不充当话语标记，这是识别工作中的难点问题。

例1 “这是五百万，天尧要我交给你，你们之间，到此为止。”她惺惺作态，佯装同情可儿。（夙云《人妖新娘》）

例2 “算了，不要为一个女人花这么多心思，不值得。到此为止。”宋思明暗暗想。（六六《蜗居》）

在例句1中，“到此为止”充当谓语，有实在的意思，不属于话语标记。例2中，“到此为止”与上文叙述的内容没有结构或内容上的关系，属于结束类话语标记。

此外，一个话语标记也存在兼属多种话语标记种类的情况。

例3 “别再孩子气了！就这样，你马上过来这里，不然我就去接你过来。”（于萍《迷糊娇娇女》）

例4 “你只要告诉我，我能不能在两点之前收到传真，就这样。”她迎上他的目光，那双深邃黝黑的眸子，平静的令人憎恨！（沈亦《碎心情劫》）

话语标记词“就这样”在例句3和例句4中分别充当转换性话题标记和结束性话题标记。这种话语标记跨类兼用的情况，也是分类任务中面临的挑战。

目前，对话语标记的发掘和功能分类工作主要依靠语言学家对其进行逐一分析和归纳分类，工作繁杂，并且人工统计所得数量有限，大量的话语标记还未被统计。话语标记的特征抽象，传统特征工程等方法不容易提取，鲜见系统的话语标记的自动识别或分类工作。本文以话题性话语标记为研究对象，构建话题性话语标记语料库，综合运用深度学习模型，实现对现代汉语话语标记的自动识别与分类。

2 相关工作

话语标记的分类体系是识别工作的基础，目前国内学者对现代汉语话语标记的功能分类有二分法、三分法、四分法和多分法[6]。

二分法中，Li[7]运用元话语理论，从语篇功能和人际功能的角度对元话语标记进行分类。在三分法研究中，Liu[3]从话语标记的形式分类、对语境的依存关系、功能分类这3个角度展开研究。Li[8]采用四分法，将话语标记的功能分为话语组织功能、人际互动功能、情态表达功能和衔接连贯功能。多分法体系中，Xi[9]对比了英汉语用标记语，从表达语用信息和元话语功能的角度将语用标记语分为7类。Zhou[6]从语用目的和语用功能的角度出发，将话语标记分为语篇类标记和语义类标记两大类。其中，语篇类标记分为话题性话语标记和衔接性话语标记，语义类标记分为理据性话语标记、解说性话语标记和推论性话语标记，下属总共22小类[6]。

其中，话题性话语标记用于对话题的控制，主要包括引导性话题标记、顺序性话题标记、转换性话题标记和结束性话题标记，其分别具有引导和提示话题、对话语内容排序、转换话题内容和结束话题的功能。话题性话语标记在口语表达和书面写作中出现频率高，变体丰富，数量多，作为话语标记及其类别的自动识别任务的研究对象具有代表性作用。本文以周明强所提分类体系中的话题性话语标记这一大类及其下属二级分类的47个话语标记为研究对象，进行语料数据集的标注和整理。

话语标记自动识别相关的现有研究工作中，Zhao[10]提出了一种基于多维度频谱图的话语标记语音特征识别系统，Xiao等[11]基于依存关系图对8个固定的话语标记进行识别。上述研究均未涉及对话语标记功能类别、话语标记变体以及新话语标记的识别。本文基于这些问题展开了进一步的研究，通过融合词性特征的NFLAT指针网络模型，对话语标记进行自动识别与功能分类。

3 融合词性特征的NFLAT指针网络

本文模型的整体设计围绕解决对准标记词是否在语句中充当话语标记的识别、对话语标记词的变体进行识别和对部分多功能性话语标记的分类展开。模型的整体架构如图1所示。

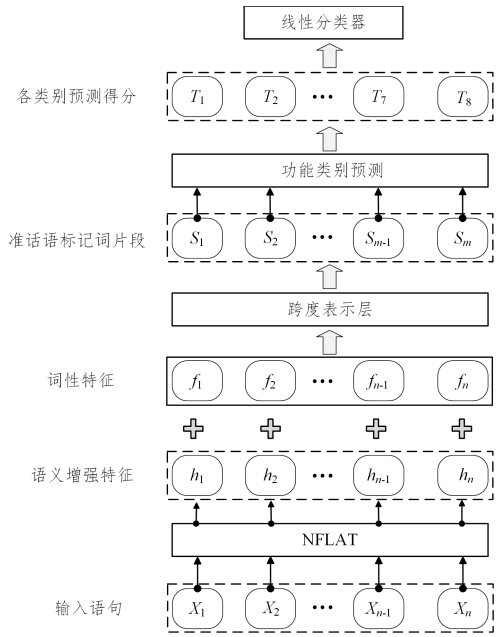

图1 模型架构图

Fig. 1 Model architecture diagram

输入语句经过NFLAT网络模型训练和词性标注特征表示，再由全局指针网络对语句中的话语标记词进行预测，并对预测出的话语标记词进行具体功能分类的识别。

3.1 NFLAT语义增强方法

中文分词任务存在词汇边界模糊、复杂组合、实体嵌套和词长不定等问题[12]。因此，近年来中文NER任务主要针对外部数据进行处理，例如词典信息[13]、字形信息[14-15]和语义信息[16]，以提升模型性能，或基于字符、融合词汇信息提出增强模型[17]。其中，Wu等[18]针对Li等[19]提出的Flat-Lattice-Transformer模型中因计算量和内存消耗巨大而难以处理长跨度词汇的问题，提出一种更有效的语义词汇增强方法，即非扁平晶格网络NFLAT（Non-Flat-Lattice）。本文融合NFLAT模型的前两个阶段，即InterFormer层和Transformer编码层，对输入语句分别进行词汇边界信息及语义信息融合处理，对包含词汇信息编码。具体网络层次如图2所示。首先，InterFormer对输入语句进行处理，得到融合词汇边界

语义信息的字符表征；然后，通过 Transformer 编码器对上下文进行编码，得到指针网络的输入序列。

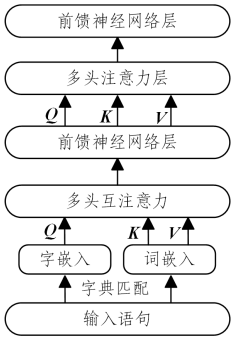

图 2 NFLAT 网络结构

Fig. 2 NFLAT network structure

InterFormer 方法中包含一个多头注意力机制（Multi-Head Inter Attention）和一个前馈神经网络层（Feed-forward Neutal Network），其目的是构建一个非平面格（Non-Flat-Lattice）网络结构。输入序列通过词汇匹配处理得到字向量 $\boldsymbol{C}=\{c_1,c_2,\cdots,c_n\}$ 和词向量 $\boldsymbol{W}=\{w_1,w_2,\cdots,w_m\}$，通过嵌入查找表处理，可以得到字嵌入向量 $\boldsymbol{X}^c=\{x_{c_1},x_{c_2},\cdots,x_{c_n}\}$ 以及词嵌入向量 $\boldsymbol{X}^w=\{x_{w_1},x_{w_2},\cdots,x_{w_m}\}$。其中，字嵌入向量经过线性变换得到输入注意力层的查询向量 $\boldsymbol{Q}$，词嵌入向量经过线性变换得到输入注意力层的键向量 $\boldsymbol{K}$ 和值向量 $\boldsymbol{V}$。具体计算如式（1）所示：

$$[\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}]=[\boldsymbol{X}^c\boldsymbol{W}_q,\boldsymbol{X}^w\boldsymbol{W}_k,\boldsymbol{X}^w\boldsymbol{W}_v] \tag{1}$$

其中，$\boldsymbol{W}_q$，$\boldsymbol{W}_k$ 以及 $\boldsymbol{W}_v$ 均为可学习的超参数。互注意力机制用来融合词汇边界信息，如式（2）所示：

$$InterAtt(\boldsymbol{A},\boldsymbol{V})=\mathrm{softmax}(mask(\boldsymbol{A}))\boldsymbol{V} \tag{2}$$

$$\boldsymbol{A}=(\boldsymbol{Q}_i+u)^{\mathrm{T}}\boldsymbol{K}_i+(\boldsymbol{Q}_i+v)^{\mathrm{T}}R_{ij}^* \tag{3}$$

其中，$mask()$ 是字和词的得分掩码，其对序列中的空位赋值为 1×10^{-5}，使该部分的权重在经 softmax 全连接层正则化后结果接近 0。式（3）中注意力 $\boldsymbol{A}$ 的计算方式参考 Dai 等在 TransformerXL[20] 中提出的引入相对位置编码机制的自注意力计算公式。其中，$1\leqslant i\leqslant n$，$1\leqslant j\leqslant m$，u 和 v 均为可学习的超参数。相对距离 R_{ij}^* 的计算如式（4）所示：

$$R_{ij}=RELU(\boldsymbol{W}_r(p_{h_i^c-h_j^w}\oplus p_{t_i^c-t_j^w})) \tag{4}$$

其中，$\boldsymbol{W}_r$ 为可学习的超参数，h 和 t 表示输入文本中词语的第一个字符和最后一个字符的位置编号，c 和 w 分别代表字和词。$h_i^c-h_j^w$ 为第 i 个汉字和第 j 个词语开头的位置偏移量，$t_i^c-t_j^w$ 为第 i 个字和第 j 个词语词尾的位置偏移量，即相对距离。位置编码 p 采用 Vaswani 等[21] 提出的正弦函数和余弦函数生成方式，如式（5）和式（6）所示：

$$p_{\mathrm{span}}^{(2k)}=\sin\left(\frac{span}{10\,000^{2k/d_{\mathrm{model}}}}\right) \tag{5}$$

$$p_{\mathrm{span}}^{(2k+1)}=\cos\left(\frac{span}{10\,000^{2k/d_{\mathrm{model}}}}\right) \tag{6}$$

其中，$span$ 即为两个相对距离，k 表示 k 维，d_{model} 表示隐藏层深度。

由于多头注意力信息之间的互补作用，我们在实验中发现多头的互注意力机制可以更有效地融合词边界等信息，从而得到更好的语义增强效果。多头间注意力计算如式（7）和式（8）所示：

$$H^{(s)}=InterAtt(X^{C,(s)},X^{W,(s)}) \tag{7}$$

$$MultiHead(X^C,X^W)=[H^{(1)},H^{(2)},\cdots,H^{(l)}] \tag{8}$$

其中，l 为注意力头数，$H^{(s)}$ 表示第 s 个互注意力头在字和词汇向量子空间的输出结果，$X^{C,(s)}$ 和 $X^{W,(s)}$ 为字和词汇在其子空间的向量表示。注意力编码器的前馈神经网络层采用两个全连接层实现，如式（9）所示：

$$FFN(x)=\max(0,x\,W_1+b_1)W_2+b_2 \tag{9}$$

InterFormer 通过上述过程融合了字符表征与词典信息后，由 Transformer 编码器进行上下文编码。在编码器中使用未缩放的自注意力机制[22]，取得的结果会更好。

3.2 词性特征融合

为提高模型对句中话语标记及其变体的识别准确率，本实验着重研究了话语标记总体特征（语音、句法、语义和语用 4 个方面）的可利用性。在语义特征层面上，语言学家认为话语标记本身可能有概念语义，但发挥话语标记作用时通常不表达概念语义，应侧重于语用意义的理解。例如标记“不瞒你说”，其概念语义为“我下面说的话是没有瞒着你的”，作为话语标记时，则有“我说的话是真实的”和“我对你是信任的，没有隐瞒”等语用意义。

基于这一特征，我们在对语料数据进行词性分析时发现，准话语标记词意义不同时，其对应的词性标注结果亦有所不同。

例 5 闲话 休提 ， 且 说 “ 升官

v v wp c v wp v

怎么 会 和 红楼 联上 的 ？

r v p n v u wp

例 6 且 说 文人 哭 穷

c v n v a

例 7 只能 凭 个人 感觉 ， 姑且 说

v p n n wp d v

之 。

r wp

如上所示的例句中，可以观察到词性标注结果的不同之处。例句 5 和例句 6 中，准标记词“且说”为引导性话题标记，旨在引出后面关于“升官”和“文人哭穷”话题的讨论，其标注结果均为“c+v”（连词+动词）。而例句 7 中的“且说”则与前后内容构成更大的句法单位，在此处不充当话语标记词，标注结果为“d+v”（副词+动词），与上两例并不相同。因此，我们在实验中融入词性标注特征。

词性特征的提取借助哈工大语言技术平台 LTP[23]（Language Technology Platform）实现. 本实验中，特征提取部分的实现模型结构如图 3 所示。

在共享编码层，对于给定的输入序列 $\boldsymbol{X}=[x_1,x_2,\cdots,x_n]$（$x_i$ 为输入字符），经过预训练模型 ELECTRA 后，输出对应的隐藏层编码序列 $\boldsymbol{H}=[h_{[\mathrm{CLS}]},h_1,h_2,\cdots,h_n,h_{[\mathrm{SEP}]}]$。在词性标注（Part-of-Speech，POS）层，通过线性解码器对每个字符进行分类：

$$y_i=\mathrm{Softmax}(W_{\mathrm{POS}}h_i+b_{\mathrm{POS}}) \tag{10}$$

其中，y_i 表示第 i 个字符的词性标注标签概率分布，W_{POS} 和 b_{POS} 为可训练的超参数。

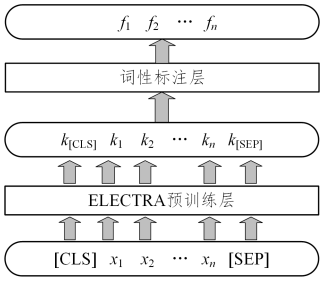

图 3 词性特征提取

Fig. 3 Part-of-speech feature extraction

3.3 Global Pointer NER 全局指针网络模型

本文摒弃了传统序列到序列及文本分类的实验方法，采用全局指针网络(Global Pointer，GP)[24]对话语标记及其类别进行识别。全局指针网络由指针网络[25]发展而来，是一种基于子序列分类的解码方法。对于输入长度为 n 的序列，可以枚举出 $n\times(n+1)/2$ 个候选实体，再从中判断是否存在话语标记词。将识别问题转化为一个从 $n\times(n+1)/2$ 中选取 k 个标签的多分类问题。网络模型将实体的起始位置作为一个整体进行识别，可以无差别地识别重叠实体和长实体，具有全局性特征，识别效率高[26]。以指针的形式预测句中准关键词的开始位置和结束位置，抽取出所有候选话语标记词的子序列。对于输入的句子序列，全局指针网络通过一个上三角矩阵遍历所有的子序列，如图 4 中所示：以转换性话题(即话题性话语标记，简称话题标记)标记不同类型的话语标记分类器对子序列进行分类，得到该类型下的识别结果。

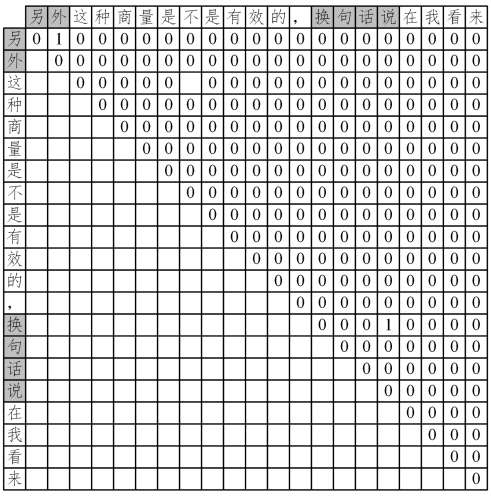

图 4 转换性话题标记识别矩阵

Fig. 4 Inflectivet opical discourse marker recognition matrix

为充分利用边界位置，模型中引入了相对位置信息。在实体表示中引入 RoPE[27]相对位置编码，并参与话语标记类型的概率分布计算。对于给定的输入序列 $\boldsymbol{X}=[x_1,x_2,\cdots,x_n]$($x_i$为输入字符)，通过 *NFLAT* 语义增强方法和词性标注特征模型得到跨度表示层的输入序列的嵌入向量表示 $\boldsymbol{K}=[k_{[CLS]},k_1,k_2,\cdots,k_n,k_{[SEP]}]$。在跨度表示层中，实现基于全局指针网络的实体标签跨度解码：模型通过两个基于实体起、止索引的前馈神经网络层，得到准话语标记的实体表示，如式(11)和式(12)所示。

$$q_{i,\alpha}=W_{q,\alpha}h_i+b_{q,\alpha} \tag{11}$$

$$p_{i,\alpha}=W_{p,\alpha}h_i+b_{p,\alpha} \tag{12}$$

其中，$q_{i,\alpha}\in R^d$，$p_{i,\alpha}\in R^d$，为识别类型为 α 时所预测实体的向量表示。然后，通过跨度 $s[i:j]$计算每个候选子序列被预测为该类型的话语标记词的概率，如式(13)所示。

$$S_\alpha(i,j)=q_{i,\alpha}^{\mathrm{T}}p_{j,\alpha} \tag{13}$$

引入满足$R_i^{\mathrm{T}}R_j=R_{j-i}$的 RoPE 相对位置信息编码，可提高全局指针网络对实体长度信息的利用效率，减少不同实体首位相互组合的情况。

$$R_i^{\mathrm{T}}R_j=R_{j-i} \tag{14}$$

$$S_\alpha(i,j)=(R_iq_{i,\alpha})^{\mathrm{T}}(R_jp_{j,\alpha}) \tag{15}$$

其中，$\boldsymbol{R}$ 为旋转位置编码变换矩阵，可标注输入序列中字符的相对位置。实验中，通过损失函数缓解类不平衡问题。在单类别分类实验中，交叉熵损失函数如式(16)所示。

$$L=\log\left(1+\sum_{i=1,i\neq t}^{n}\mathrm{e}^{S_i-S_t}\right) \tag{16}$$

其中，S_i是非目标类别得分，S_t是目标得分。那么在多标签分类的情况下，网络模型的训练目标是使目标类别的得分不低于非目标类别的得分，因此损失函数如式(17)所示。

$$\log\left(1+\sum_{i\in\Omega_{\mathrm{neg}}}\mathrm{e}^{S_i}\sum_{j\in\Omega_{\mathrm{pos}}}\mathrm{e}^{-S_j}\right) \tag{17}$$

其中，Ω_{neg}和Ω_{pos}分别为正样本和负样本集合，则对于话语标记的某一类型 α，损失函数为：

$$\log\left(1+\sum_{(q,p)\in P_\alpha}\mathrm{e}^{-S_\alpha(q,p)}\right)+\log\left(1+\sum_{(q,p)\in Q_\alpha}S_\alpha(q,p)\right) \tag{18}$$

其中，q 和 p 为一个跨度的开始和结尾索引，P_α表示包含 α 类型话语标记的跨度集合，Q_α表示不含话语标记的跨度或含有其他类型话语标记的跨度的集合，$S_\alpha(q,p)$为跨度 $s[p:q]$为 α 类型话语标记的分数。

4 实验

4.1 数据集

本文中语料选自北京大学中国语言学研究中心语料库 CCL(Center for Chinese Linguistics PKU)、北京语言大学语料库中心 BCC(BLCU Corpus Center)和媒体语言语料库 MLC(Media Language Corpus)。将它们分成 2 个语料集：话题性话语标记语料集，以及新话语标记语料集。

话题性话语标记语料集包含 7 900 条语句，涉及 47 个话题性话语标记。由于部分话语标记词具有多功能性(见表 1)，因此识别类型共涉及 8 个。

表 1 多功能话语标记

Table 1 Multifunctional discourse markers

多功能标记	功能	分类
对/就 X 来说	引导性话题标记	引话
	对象性推论标记	其他
这/那就是说	转换性话题标记	转话
	判断性解说标记	其他
(你)比方/如说	转换性话题标记	转话
	示例性衔接标记	示衔
	方式性解说标记	其他
再说(了)	转换性话题标记	转话
	过渡性衔接标记	过衔
到此为止	转换性话题标记	转话
	结束性话题标记	结话
就这样(吧)	转化性话题标记	转话

话题性话语标记语料集的分布情况如表2所列。每条语句中均包含话题性话语标记的准标记词，准标记词在句中是否发挥话语标记功能的正、负用例均有涉及。

新话语标记语料集共包含849条语句，该语料集中的准标记词均未在话题性话语标记语料集中出现，为待识别的新标记。新话语标记语料集的具体分布如表3所列。语料集中的语句亦均包含准标记词，语料涉及每个话语标记词的正、负样本。

表2 话题性话语标记语料集

Table 2 Topical discourse marker corpus

话题性话语标记	话语标记	语料数量
引导性话题标记（引话）	话说；且说；却说；说起来；是这样的；是这样；常言道；俗话说；就（拿）X来说；对（于）它来说/讲；对（于）X而言；说起X	2763
顺序性话题标记（顺话）	首先、其次、再次、最后；第一、第二…最后；“是”字组；“其”字组；“方面”组；“则”字组	913
转折性话题标记（转话）	换言之；换而言之；换句话说；比如说；比方说；你比如说；就是说；这/那就是说；也就是说；或者说；简而言之；话（又/再）说回来；话还得说回来；到此为止；顺便说/问一下；话虽/是这么说；再说（了）；扯远了；另外；说到X；说到这里；要论X；要说X；（就）这样吧；就这样；要不这样；那这样；充其量	3556
结束性话题标记（结话）	就这样（吧）；要不这样；那这样；充其量	668

表3 新话语标记语料集

Table 3 Untrained discourse marker corpus

话语标记	语料分布
至于说	148
我说	102
要我说	156
归根到底	189
我想说的是	88
进一步说	175

本文在话题性话语标记语料集和新话语标记语料集上分别进行实验，以检测模型对训练过的话语标记自动识别与分类的功能和未经训练的新话语标记自动识别与分类的功能。

4.2 实验参数与评价指标

本实验的开发环境为Google T4 GPU，训练参数如表4所列。

表4 实验相关参数

Table 4 Experimental parameters

参数	参数值
batch_size	32
epoch	20
hidden_size	224
learning rate	1×10^{-4}
max_seq_length	302
decay_rate	0.9
decay_steps	50
dropout	0.1

实验中对话语标记的识别与分类结果的评价指标为精确率（Precision，P）、召回率（Recall，R）和F_1值（F_1-score）。上述评价指标的计算如式（19）—式（21）所示。

$$Precision=\frac{TP}{TP+FP} \tag{19}$$

$$Recall=\frac{TP}{TP+FN} \tag{20}$$

$$F_1=2\times\frac{Precision\times Recall}{Precision+Recall} \tag{21}$$

4.3 实验结果分析

本文在话题性话语标记语料集和新话语标记语料集上分别进行了实验1和实验2，以检测各阶段模型的性能。

实验1中训练集和测试集由话题性话语标记语料集中各个标记对应的例句按比例划分，测试各阶段模型对已学习话语标记的识别与分类能力。由表5可知，融合词性特征的NFLAT指针网络模型在7个类别上取得最优F_1值，对“过渡性衔接”这一类别的话语标记识别完全准确，对多功能话语标记的“示例性衔接”类别语料识别的准确率有待提高。总体来说，融合词性特征的NFLAT指针网络模型对于训练过的话语标记识别及分类能力最佳。

实验2的训练数据集为话题性话语标记数据集中的所有语料，测试集为新话语标记数据集中的所有语料，以测试各阶段模型对全新话语标记的识别和分类能力。由表6可知，在新话语标记的识别与分类任务中，在“至于说”“我说”“归根到底”“我想说的是”和“进一步说”这5个话语标记上，融合词性特征的NFLAT指针网络模型取得最佳F_1值；在“要我说”的语料识别结果中，由于一些准标记词出现在话轮开端，但与上下文构成句法结构，不属于话语标记范畴的用例，因此被错误预测为话语标记，导致精确度降低。综合来说，融合词性特征的NFLAT指针网络在各阶段模型中仍为最优。

表5 各类别识别结果

Table 5 Rrecognition result

（%）

模型	指标	结话	示衔	过衔	引话	转话	延衔	顺话	其他
GP（Global Pointer）	*Precision*	75.00	76.74	72.80	76.29	73.12	77.69	75.29	76.19
	Recall	64.45	65.36	65.24	65.46	64.73	72.31	65.44	65.46
	F_1	69.32	70.39	68.73	70.25	71.01	74.64	70.24	70.25
POS+GP	*Precision*	73.77	75.53	74.91	75.19	75.19	78.62	75.19	75.76
	Recall	67.23	67.74	66.56	67.70	67.71	71.97	67.70	67.77
	F_1	70.24	**71.33**	70.43	71.16	71.16	75.12	71.16	71.46
NFLAT+POS+GP	*Precision*	86.66	53.38	100.00	98.49	94.29	90.00	94.30	91.08
	Recall	92.85	73.33	100.00	97.03	89.95	81.81	95.86	90.12
	F_1	**89.65**	61.11	**100.00**	**97.76**	**92.07**	**85.71**	**95.08**	**90.60**

表 6 新话语标记的识别和分类结果

Table 6 Untrained discourse markers recognition and classification result

(%)

模型方法	指标	至于说	我说	要我说	归根到底	我想说的是	进一步说
GP(Global Pointer)	*Precision*	91.36	45.45	71.92	37.50	64.47	59.61
	Recall	100.00	100.00	97.61	100.00	94.23	83.78
	F_1	95.48	62.49	**82.81**	54.54	76.55	69.66
POS+GP	*Precision*	93.38	46.15	70.59	37.91	65.82	48.65
	Recall	100.00	100.00	100.00	100.00	100.00	97.30
	F_1	96.57	63.15	82.75	54.97	79.39	64.86
NFLAT+POS+GP	*Precision*	93.33	77.78	69.42	37.91	66.67	62.26
	Recall	99.21	93.33	100.00	100.00	96.15	89.19
	F_1	**96.18**	**84.84**	81.95	**54.97**	**78.74**	**73.33**

4.4 实验方法对比

在对话语标记自动识别任务的探索中，为验证融合词性特征的 NFLAT 指针网络模型的性能，在基于 XLNet[28] 的下游任务微调模型和 GPT-3[29] 预训练语言模型上和 GPT-3.5 预训练语言模型上对话题性话语标记数据集中的语料进行了话语标记的自动识别。结果如图 5 所示，可以看出融合词性特征的 NFLAT 指针网络模型的 3 个评价指标均最优。

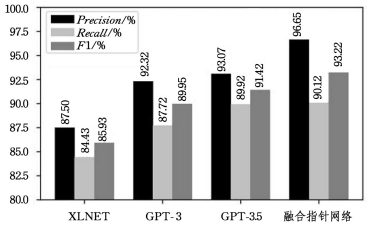

图 5 对比实验结果

Fig. 5 Comparison experiment results

4.5 消融实验

本文在实验 1 和实验 2 上进行了是否融合词性特征、是否使用语义增强方法的消融实验，各项指标如表 7 所列。可以看出，词性特征提高了话语标记识别任务的精确率和 F_1 值，NFLAT 语义增强方法使模型在各个指标上均有提高，融合词性特征的 NFLAT 指针网络模型的识别性能综合最佳。此外，由图 6 可知，使用 NFLAT 语义增强方法后，模型在处理语料时的工作效率(inference time)也得到了提高。

表 7 消融实验

Table 7 Ablation experiment

(%)

实验	模型	*Precision*	*Recall*	F_1
实验 1	GP	85.68	89.13	87.37
实验 1	POS+GP	87.75	93.61	90.58
实验 1	NFLAT+POS+GP	**93.07**	**90.12**	**91.57**
实验 2	GP	70.19	73.06	71.59
实验 2	POS+GP	75.47	71.24	73.29
实验 2	NFLAT+POS+GP	**94.55**	**88.88**	**91.63**

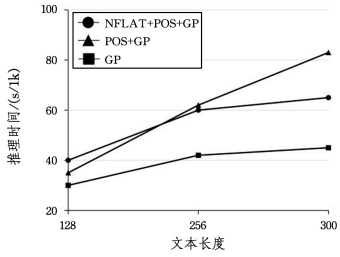

图 6 模型性能

Fig. 6 Model performance

结束语 本文基于语义增强方法、词性标注特征和全局指针网络提出一种深度学习模型，进行话题性话语标记及其类别的自动识别。经实验验证，该模型对训练过的话语标记及其类别的识别精确率可达 93.07%，对未经训练的新话语标记及其类别的识别准确率可达 94.55%(新话语标记几乎不涉及多功能话语标记，故在语料集上的指标数值更高)。对新话语标记获取的高准确率充分证明了该工作应用于挖掘尚未被统计的汉语话语标记、扩充现代汉语话语标记库的可能性，这将极大地提高语言学界对各类话语标记归纳分析的工作效率。所提模型在机器翻译、篇章分析、对话生成等中文自然语言处理任务中发挥重要作用；但在多功能话语标记以及与上下文结构不完全独立的话语标记识别上，精确率还有进一步提升的空间。后续的工作拟从多特征融合、文本分类的数据增强方法等角度继续探索，以期在话题性话语标记之外的其他类别标记词的识别和分类工作上有更出色的表现。

参 考 文 献

[1] XIAO M. Research hotspots and development analysis of discourse markers [J]. Central China Humanities, 2021, 13(3): 160-169.

[2] ZHOU M Q. Research on the system of discourse markers and cognition of modern Chinese[M]. Beijing: China Social Science Press, 2022: 1-23.

[3] LIU L Y. Research on Chinese discourse markers[M]. Beijing: Beijing Language and Culture University Press, 2011: 26-38.

[4] XU J J. The discourse marker RANHOU and its functions in spoken Chinese [J]. Foreign Languages Research, 2009(2): 9-15, 112.

[5] LI Z J. Chinese new function words [M]. Shanghai: Shanghai Education Press, 2011.

[6] ZHOU M Q. An overview of the system of modern Chinese discourse markers[J]. Journal of Zhejiang International Studies University, 2020(1): 80-88, 108.

[7] LI X M. A study on Chinese metalinguistic markers[M]// Beijing: China Social Science Press, 2011: 104-137.

[8] LI Z P. A study of discourse markers in modern Chinese language [M] // Beijing: World Publishing Corporation, 2015: 78-83.

[9] XI J G. Pragmatic markers in English and Chinese: A cognitive study[M]// Hangzhou: Zhejiang University Press, 2009: 52-65.

[10] ZHAO Y Y. Design of discourse marker feature recognition system based on multi-dimensional spectrogram[J]. Modern Elec-

tronics Technique,2021,44(12):83-86.

[11] XIAO M,XIAO Y. Research on interpretability recognition of Chinese discourse markers based on dependency graph[J]. Journal of Central China Normal University(Natural Science),2023,57(4):528-538.

[12] QI P N,LIAO Y L,QIN B. Survey on deep learning for Chinese named entity recognition[J]. Journal of Chinese Computer Systems,2023,44(9):1857-1868.

[13] DONG C,ZHANG J,ZONG C,et al. Character-based LSTM-CRF with radical-level features for Chinese named entity recognition[C]// Proceedings 24 ICCPOL. Springer International Publishing,2016:239-250.

[14] MENG Y X,WU W,WANG F,et al. Glyce:Glyph-vectors for Chinese Character Representations[C]// Proceedings of the 33rd International Conference on Neural Information Processing Systems. 2019:2746-2757.

[15] WU S,SONG X N,FENG Z H. MECT:Multi-Metadata Embedding based Cross-Transformer for Chinese Named Entity Recognition[J]. arXiv:2107.05418,2021.

[16] NIE Y Y,TIAN Y H,WAN X,et al. Named Entity Recognition for Social Media Texts with Semantic Augmentation[J]. arXiv:2010.15458,2020.

[17] LIAO M,JIA Z,LI T R,et al. Chinese Named Entity Recognition Based on Label Information Fusion and Multi-Task Learning[J]. Computer Science,2024,51(3):198-204.

[18] WU S,SONG X N,FENG Z H,et al. Non-flat-lattice transformer for chinese named entity recognition [J]. arXiv:2205.05832,2022.

[19] LI X,YAN H,QIU X,et al. FLAT:Chinese NER Using Flat-Lattice Transformer[C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020.

[20] DAI Z H,YANG Z L,YANG Y M,et al. Transformer-XL:Attentive Language Models beyond a Fixed-Length Context [C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019:2978-2988.

[21] VASWANI A,SHAZEER N,PARMAR N,et al. Attention is all you need[C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017:6000-6010.

[22] YAN H,DENG B C,LI X N,et al. TENER:Adapting Transformer Encoder for Named Entity Recognition[J]. arXiv:1911.04474,2019.

[23] CHE W X,FENG Y L,QIN L B,et al. N-LTP:An Open-source Neural Language Technology Platform for Chinese[C]// Proceedings of Association for Computational Linguistics. 2021:42-49.

[24] SU J L,MURTADHA A,PAN S F,et al. Global Pointer:Novel Efficient Span-based Approach for Named Entity Recognition [J]. arXiv:2208.03054,2022.

[25] ORIOL V,MEIRE F,NAVDEEP J. Pointer networks[J]. arXiv:1506.03134,2015.

[26] DENG L,QI P H,LIU Z P,et al. BGPNER:A BERT-based global pointer network for named entity-relation joint extraction method[J]. Computer Science,2023,50(3):42-48.

[27] SU J L,LU Y,PAN S F,et al. Reformer:Enhanced transformer with rotary position embedding[J]. arXiv:2104.09864,2021.

[28] YANG Z,DAI Z,YANG Y,et al. Xlnet:Generalized autoregressive pretraining for language understanding[C]// Proceedings of the 33rd International Conference on Neural Information Processing Systems. 2019:5753-5763.

[29] BROWN T,MANN B,RYDER N,et al. Language models are few-shot learners[J]. Advances in Neural Information Processing Systems,2020,33:1877-1901.

YANG Jincai, born in 1976, professor, doctoral supervisor, is a member of CCF (No. 35662M). His main research interests include advanced database and information system, Chinese information processing, artificial intelligence and natural language processing.

(责任编辑:柯颖)