

基于WBA模型融合语句特征的语气词识别与分类

杨进才, 郭清创, 吕文杰, 沈显君

(华中师范大学计算机学院, 湖北 武汉 430079)

摘要: 语气词是语气的主要表达方式, 能表现说话人的情感和真实意图, 识别与分类语气词有助于把握句子的真实含义与情感倾向, 从而提升机器翻译、情感分析等自然语言处理任务的性能。语气词用法较为灵活, 同一语气词往往可以表示多种语气, 使得语气词的自动标识面临着挑战。为此, 提出一种融合语句特征的WBA模型对语气词进行识别与分类。首先, 使用WoBERT对语料文本进行预训练以获得具有上下文信息的特征向量; 其次, 在特征向量的基础上融合词性、依存句法关系、语义依存关系等外部特征增强语义信息; 最后, 使用BiLSTM模型捕捉上下文的语义依赖关系, 并使用自注意力机制筛选特征, 以进一步提升模型性能。实验表明, 该方法在语气词识别与十二分类上取得了较好的效果, 宏F1值分别为98.19%、88.74%, 相较于不同模型进一步证明了该方法的有效性。

关键词: 语气词; 序列标注; WoBERT; 特征融合; 深度学习

DOI: 10.11907/tj.241475

开放科学(资源服务)标识码(OSID):

中图分类号: TP18

文献标识码: A

文章编号: 1672-7800(2025)007-0135-08



Modal Particle Recognition and Classification Based on WBA Model Fused With Sentence Features

YANG Jincai, GUO Qingchuang, LYU Wenjie, SHEN Xianjun

(School of Computer Science, Central China Normal University, Wuhan 430079, China)

Abstract: Modal particles are the main way of expressing mood, which can express the speaker's emotions and true intentions. Identifying and classifying modal particles in sentences helps to grasp the true meaning and emotional tendency of the sentence, thereby improving the performance of natural language processing tasks such as machine translation and sentiment analysis. The usage of modal particles is relatively flexible, and the same modal particle can often represent multiple moods, which poses a challenge to the automatic identification of modal particles. Therefore, a WBA model that integrates sentence features is proposed to recognize and classify modal particles. Firstly, WoBERT is used to pre-train the corpus text to obtain feature vectors with contextual information; Secondly, external features such as part of speech, dependency syntactic relationships, and semantic dependency relationships are integrated on the basis of feature vectors to enhance semantic information; Finally, the BiLSTM model is used to capture the semantic dependencies of the context, and the self-attention mechanism is used to filter features to further improve the model performance. The experiment shows that this method has achieved good results in modal particle recognition and twelve classification, with macro F1 values of 98.19% and 88.74%, respectively. Compared with different models, it further proves the effectiveness of this method.

Key Words: modal particles; sequence annotation; WoBERT; feature fusion; deep learning

0 引言

语气作为言语交际的关键要素, 不仅能反映说话者的

态度与情感, 还能展现说话者要达到的交际目的^[1]。在语气表达过程中, 语调和语气词扮演着至关重要的角色, 语气词不仅能帮助语气表达, 还能在语调的基础上为语言增添丰富的色彩^[2]。

收稿日期: 2024-05-22

扫描二维码阅读全文:



作者简介: 杨进才(1967-), 男, 博士, 华中师范大学计算机学院教授、博士生导师, 研究方向为现代数据库与信息系统、中文信息处理; 郭清创(2001-), 男, 华中师范大学计算机学院硕士研究生, 研究方向为中文信息处理; 吕文杰(1998-), 男, 华中师范大学计算机学院硕士研究生, 研究方向为自然语言处理; 沈显君(1973-), 男, 博士, 华中师范大学计算机学院教授, 研究方向为数据挖掘、机器学习。本文通讯作者: 杨进才。

语气词分类主要根据表达语气的不同对句中的语气词进行识别与分类,有助于把握句子的真实含义与情感倾向,从而提升机器翻译、情感分析、问答系统等自然语言处理应用的性能^[3]。然而,语气词识别与分类是一项具有挑战性的任务,同一词语在句中既能充当语气词又能充当其他语法成分。例句1中“的”属于语气词,但在例句2中却属于结构助词。

例1:这种病很难治的。

例2:他是武汉的。

此外,语气词用法较为灵活,同一语气词往往能表达不同的语气,例如语气词“呢”,在例句3、例句4、例句5中均表示不同的语气。

例3:他知道你来了,想看看你呢。

例4:我写的那三个字在哪里呢?

例5:批评得重了呢,她就掉眼泪;批评得轻了呢,她又不当回事。

近年来,预训练模型在自然语言处理领域取得了显著成效,凭借丰富的语言知识和上下文信息为下游任务提供了强力支持^[4]。本文使用预训练模型与深度学习技术相结合的方法对语气词进行识别与分类^[5]。

1 相关工作

从中文信息处理角度而言,语气词的研究主要集中在语气词分类体系研究和语气词识别与分类两个方面。在语气词分类体系研究方面,高名凯^[6]将语气词分为表示询问的询问词“么、吗”,表示命令的命令词“吧、罢、呀、啊、哇”等,表示感叹的感叹词“啊、呵、呀、哩、咧”等。黄伯荣等^[7]将语气词划分为4类,分别为表示陈述的“了、的、吧、呢、着了、嘛、呗、啦、啊”等,表示疑问的“吧、吧、呢、啊”等,表示祈使的“吧、了、啊”、表示感叹的“啊”。何乐士^[8]将古代汉语中的虚词划分为语气词、叹词、连词等八大类,语气词又被划分为判断、陈述、测度、疑问、感叹、祈使等6个次类。王力^[9]将语气词划分为决定、表明、夸张、疑问、反诘、假设、揣测、祈使、催促、忍受、不平和论理等12个类别。

在语气词识别与分类方面,陈火龙^[10]提出 BiLSTM-CRF 模型对古汉语虚词进行词性标注,在包含语气词六分类的二级词性分类上取得了 67.82% 的 F1 值。周溢辉等^[11]研究了现代汉语中常用的 10 个语气词,建立了现代汉语语气词用法词典与语气词用法规则库,基于规则方法自动识别语气词的准确率为 78.43%。然而,该研究存在一定的局限性:①纳入的语气词数量较少,未能覆盖实际表达中常用的各类语气词;②孤立地看待语气词的用法,未对语气词进行系统性分类;③研究所采用的技术手段相对落后,未能充分利用现代深度学习技术的优势,模型效果较差。

在利用深度学习技术处理中文信息,王宇晗等^[12]提出

基于预训练模型与嵌入式主题模型挖掘文本主题的方法。其中,基于中文词的 WoBERT 结合了中文词特性,在获取词嵌入时以词为单位降低了词义的不确定性,效果优于基于 BERT 的模型。杨进才等^[13]提出融合语言特征的 BERT+BiLSTM 模型识别无标复句的关系,在融入词性、极性及依存句法关系等外部特征后,模型的识别性能得到显著提升。臧洁等^[14]提出融合双通道的情感分类模型,使用多头自注意力机制计算 BiLSTM 各时间的字词特征权重分配情况,对分类结果较关键的词汇给予更多关注,提升了模型分类准确率。

本文参考中国语言学家王力教授的十二分类系统,从《现代汉语语气成分用法词典》中选取 27 个现代汉语中常用的语气词作为研究对象,将语句作为有序的词组成的序列,并结合深度学习技术对语句序列进行标注,实现了语气词识别与分类^[15]。

2 特征提取

2.1 文本特征表示

文本特征表示主要包括基于词向量的模型和预训练模型两种^[16]。传统词向量模型通常基于固定大小的上下文窗口生成词嵌入,限制了其对全局上下文信息的捕捉能力;预训练模型一般使用 Transformer 结构,通过自注意力机制捕捉整个句子的上下文信息,从而更准确地表示词语的含义^[17]。常用的 Bert 模型以字为基本单位对文本进行处理,以减少分词歧义问题,但忽略了词语的整体意义,导致处理结果丢失了部分语义信息。

本文为了更完整地处理文本信息,采用以词为基本单位的中文 Bert 模型 WoBERT (Word-based BERT)^[18]。首先, WoBERT 在哈尔滨工业大学开源的 RoBERTa-wwm-ext 基础上进行预训练,将每个词用 Bert 自带的 Tokenizer 切分为字;其次,将平均字嵌入(embedding)作为词嵌入的初始值,还加入了 20 000 个额外的中文词,通过优化词汇表与训练过程来提升模型对中文文本的处理能力;最后, WoBERT 模型默认采用 jieba 工具进行语句分词,支持直接处理已分词的文本序列^[19]。

2.2 词性特征

传统词向量模型对词语进行建模时,往往忽略了词语的词性信息,每个词仅对应一个固定的词向量表示^[20]。深度学习模型能自动从文本中提取特征,但无法捕获对语气词分类标识有影响的特征,需要在模型中融入一些外部特征。然而,在实际表达中同一词语在不同词性下往往承载着不同的语义信息^[21]。为了更准确地把握不同词性下的语义差异,使用哈工大的语言技术平台 LTP (Language Technology Platform) 获取语句的词性信息。首先,将句中的词语与其对应的词性组合为“词语_词性”,得到包含词性信息的语句序列;其次,使用 Word2Vec 对该语料进行训

练,得到含有词性信息的词向量表示,并将其作为词性特征。

2.3 依存句法特征

在研究语句的句法关系时,传统模型通常只关注与当前词语相邻的词,而依存句法分析可跨越多个词语,关联距离较远的词,从而获取更丰富的语义信息。依存句法分析是指通过分析语句成分之间的依存关系来揭示句法结构,依存关系指的是词与词之间存在的支配与被支配的关系,这种关系具有方向性^[22]。在依存句法分析中,谓词被视为句子的核心,其他成分与动词直接或间接地产生联系。

例6:这是一件多么值得骄傲的事啊。

图1为例句6中使用LTP工具进行依存句法分析后得到的依存关系。由此可见,“是”作为中心词连接着整个句子,还揭示了词语之间的依存关系,例如SBV表示主谓关系。

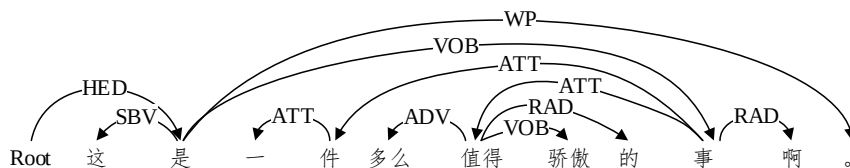


Fig. 1 Dependency syntactic relationship of example 6

图1 例句6的依存句法关系

2.4 语义依存特征

语义依存关系主要关注的是语句成分之间的直接语义关联,与句法依存关系的不同之处在于语义依存关系着重分析句中词语的语义关联和逻辑关联^[23]。语义依存不随语句结构变化而变化,不受句法结构的束缚,直接连接具有直接语义关联的语言单元,分析句子的语义依存关系能明确词语在句中承担的语义角色。

在语义依存关系中,复句有且仅有一个核心词汇,其他词语支配词均有语义依存关系。图2为在例句8中使用LTP工具进行语义依存分析的结果,语气词“啊”与其支配词“是”之间的语义依存关系为“mDEPD(依附标记)”,“骄傲”与其支配词“事”之间关系为“FEAT(修饰关系)”。

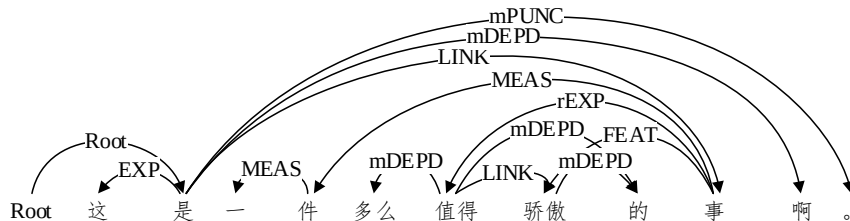


Fig. 2 Semantic dependency relationship of example 6

图2 例句6的语义依存关系

3 融合语句特征的 WoBERT-BiLSTM-Attention 模型

3.1 模型结构

图3为融合语句特征的WoBERT-BiLSTM-Attention模型结构,由输入层、特征融合层、BiLSTM层、注意力机制层、输出层组成。首先,对包含 n 个词语的输入序列 $X=[X_1, X_2, \dots, X_n]$,借助LTP工具分别提取词性、依存句法和语义依存特征,将上述3种特征拼接后通过全连接神经网络转换为语句特征;其次,将语句特征与WoBERT提取的词向量特征拼接为融合特征,并将其输入BiLSTM层以进一步捕捉上下文的语义信息;最后,通过自注意力机制对输出的特征向量进行重要性加权,经过Softmax分类后获得序列标注结果 $[Tag_1, Tag_2, \dots, Tag_3]$ 。

3.2 特征融合层

特征融合层的作用是整合语句特征与预训练的文本特征,处理流程如图4所示。首先,对于输入分词序列 $X=[X_1, X_2, \dots, X_n]$,使用WoBERT作为预训练语言模型进行

编码得到向量序列 e ,计算流程如式(1)所示。其中,WoBERT在处理文本序列时在序列头部、尾部分别加上特殊占位符[CLS]、[SEP],因此向量序列 e 也会添加相应的向量表示。

$$e = [e_{[CLS]}, e_1, \dots, e_i, \dots, e_{[SEP]}] = \text{WoBERT}([X_1, \dots, X_i, \dots, X_n]) \quad (1)$$

对序列 X 进行词性标注后,可通过“词语_词性”信息查找对应的Word2Vec词向量,以获得序列的词性向量表示 p 。由序列 X 的依存句法分析结果可得 $\text{head}=[\text{head}_1, \text{head}_2, \dots, \text{head}_n]$ 和 $\text{label}=[\text{label}_1, \text{label}_2, \dots, \text{label}_n]$, head_i 表示依赖词的位序, label_i 表示依存关系的类型。将两者转换为对应的数值形式并按照式(2)进行拼接,得到序列 X 依存句法特征 d 。

$$d_i = [\text{head}_i, \text{label}_i] \quad (2)$$

同理,还可得到语义依存特征 s 。由于依存句法和语义依存关系都依赖于词语的位序信息,为此将序列的位置编号转换为特征 t 。

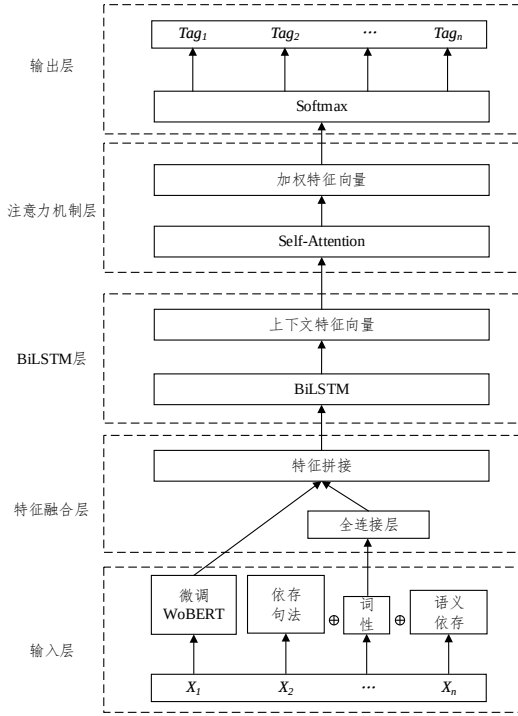


Fig. 3 Model framework

图3 模型框架

为了保证特征序列长度的一致性,在词性特征 p 、词序特征 t 、依存句法 d 、语义依存特征 s 的头部和尾部添加占位符[SEG]对应的特征表示,并将上述特征进行拼接得到整合的外部特征 x 。具体计算方式如式(3)所示。

$$x_i = [p_i, t_i, d_i, s_i] \quad (3)$$

由于外部特征 x 中各部分的重要性不同,通过一层全连接神经网络对其进行加权得到语句特征 h 。具体计算方

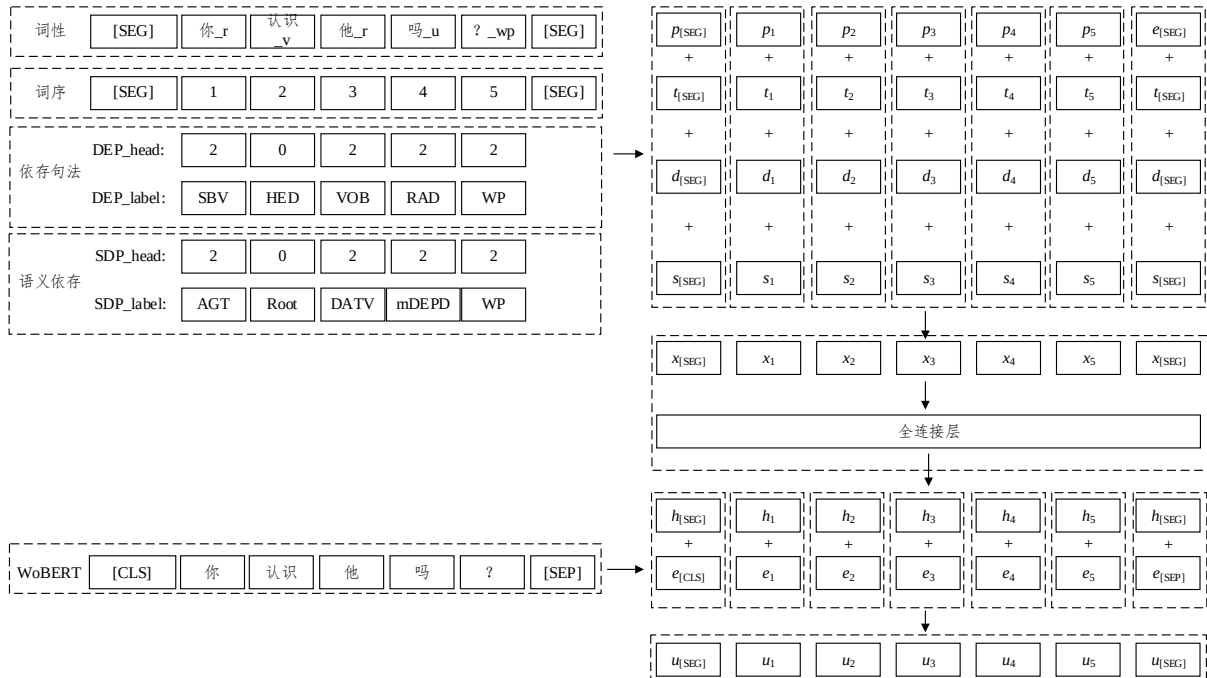


Fig. 4 Feature fusion process

图4 特征融合流程

式如式(4)所示。

$$h = \sum_{i=1}^n w_i x_i + b_i \quad (4)$$

式中: w_i 为权重参数, b_i 为偏置量。

最终,将预训练词向量特征 e 与语句特征 h 进行拼接即可得到融合后的特征向量 u 。

3.3 BiLSTM层

长短期记忆网络(Long short-term memory, LSTM)是一种特殊的循环神经网络(Recurrent Neural Network, RNN),引入门控机制解决长序列训练过程中的梯度消失或梯度爆炸问题,具体结构如图5所示。

图5中 X_t 为 t 时刻输入、 h_t 为 t 时刻输出、 C_t 为 t 时刻的细胞状态。LSTM主要通过 f_t 、 i_t 和 o_t 门结构有选择性地实现信息流动。

(1) f_t 代表遗忘门,用来控制 C_{t-1} 中信息的遗忘程度,计算公式如式(5)所示。

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f) \quad (5)$$

式中: σ 为Sigmoid函数; W_f 为遗忘门权重; b_f 为遗忘门偏置。

(2) i_t 代表输入门,负责控制信息的更新程度,计算公式如式(6)所示。首先,利用tanh函数可得候选细胞信息 C'_t ,计算公式如式(7)所示;其次,通过遗忘门和输入门可得信息细胞信息 C_t ,计算公式如式(8)所示。

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \quad (6)$$

$$C'_t = \tanh(W_c[h_{t-1}, x_t] + b_c) \quad (7)$$

$$C_t = f_t * C_{t-1} + i_t * C'_t \quad (8)$$

式中: W_i 为输入门权重; b_i 为输入门偏置; W_c 为候选细

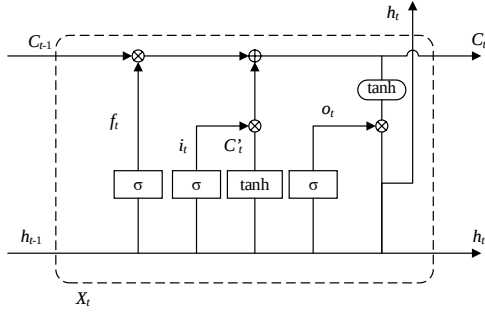


Fig. 5 Internal structure of LSTM

图5 LSTM内部结构

胞信息权重; b_c 为候选细胞信息偏置。

(3) o_t 代表输出门,作用是控制信息输出,计算公式如式(9)所示。最终 t 时刻的隐层输出 h_t 的计算公式如式(10)所示。

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \quad (9)$$

$$h_t = o_t \times \tanh(C_t) \quad (10)$$

式中: W_o 为输出门权重; b_o 为输出门偏置。

在中文信息处理中一个词的含义与上下文都存在一定的联系,因此该层搭建了双向长短期记忆递归神经网络(Bidirectional Long Short Term Memory Recurrent Neural Network, BiLSTM),分别从正反两个方向学习语义信息,具体结构如图6所示。其中,正向与反向 LSTM 的输出序列分别为 $\vec{h}_t, \overleftarrow{h}_t$, 则 t 时刻的 BiLSTM 的输出 H_t 由 $\vec{h}_t, \overleftarrow{h}_t$ 连接而成,计算公式如式(11)所示。

$$H_t = [\vec{h}_t, \overleftarrow{h}_t] \quad (11)$$

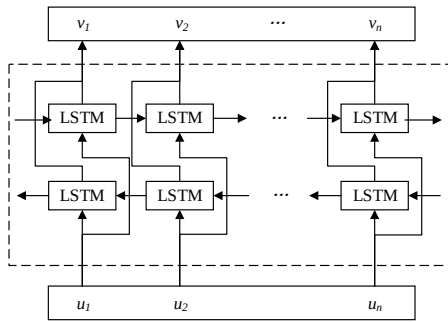


Fig. 6 BiLSTM model structure

图6 BiLSTM模型结构

3.4 注意力机制层

由于 BiLSTM 提取到的上下文特征向量中包含了多种语义信息,但并非所有信息都有助于分类与识别语气词。为此,本文创新性地引入了自注意力机制来筛选特征向量中的重要信息。在注意力机制中,输入的数据序列被看作为 K (Key) 和 V (Value), 首先根据给定目标中的查询值 Q (Query) 计算 K 与 Q 之间的相似度,其次将相似度转换为对应的权重系数并与 V 值加权求和,即可得到注意力输出^[24]。

自注意力机制是注意力机制的变体,相较于注意力机

制而言更关注输入数据的内部相关性^[25]。 Q, K, V 来自同一数据源,由同一矩阵通过不同线性变换而来,具体计算过程如图7所示。首先,通过 Q 与每个 K 计算得到相应的注意力权重;其次,将权重作用于对应的 V 值上得到输出。具体计算公式如式(12)所示。

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (12)$$

式中: d_k 为 K 的维度; $\frac{1}{\sqrt{d_k}}$ 为缩放因子,作用是防止点

积过大影响网络学习; $softmax$ 为权重归一化函数。

在通过自注意力机制对特征进行筛选时,可设置输入数据与输出结果的维度不变。由于自注意力机制中的变换矩阵为可学习参数,因此通过网络学习的方式可获取最佳的分配权重,进而得到最优的特征表示。

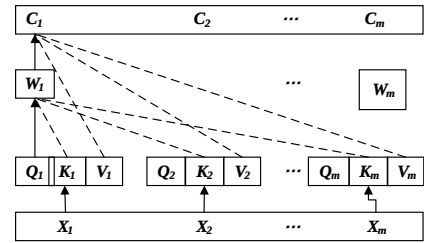


Fig. 7 Self attention mechanism workflow

图7 自注意力机制工作流程

4 实验结果与分析

4.1 语料资源

本文依据语气词与类别的对应关系,采用人工标注方法构建语气词分类标注数据集,语料主要来源于两个部分:①CCL语料库是一个大规模的中文语言语料库,包含了数量大、质量高的中文成语和词汇用例,部分口语语料、影视作品语料、网络语料和书面语语料,报纸语料占绝大比例;②BCC汉语语料库是北京语言大学语言智能研究院研发的通用单语语料库,总字数为95亿字,涵盖报刊、文学、对话、微博、科技、综合和古汉语等领域语料^[26-28]。

首先,从语料库中检索出包含27个语气词的语句,使用LTP工具对句子进行分词;其次,判断语气词类别并将结果添加到标注序列中;最后,对句子中的非语气词成分进行区分和标注,以确保数据的完整性。本文经过以上步骤构建了包含10140条语句的语气词分类标注数据集,各类语气词的数量如表1所示。在实验过程中,分别抽取各类语料中70%的数据作为训练集,剩余30%的数据作为测试集。

4.2 参数设置与评价指标

表2为模型参数设置,评价指标为Accuracy(准确率)、Precision(精确率)、Recall(召回率)和F1值,通过宏平均(macro-average)的计算方法评价各指标。

Table 1 Distribution of language materials for various modal particles

表 1 各类语气词的语料分布情况

类别	语气词	数据量
决定	不成 ₁ ,的 ₁ ,啦 ₁ ,了 ₁ ,嘞 ₁ ,哩 ₁ ,喽 ₁ ,嘛 ₁ ,呐 ₁	1 089
表明	啊 ₁ ,吧 ₁ ,喂 ₁ ,的 ₂ ,啦 ₂ ,嘞 ₂ ,哩 ₂ ,嘛 ₂ ,哪 ₁ ,呐 ₂ ,呢 ₁ ,也罢 ₁ ,也好 ₁	2 103
夸张	啊 ₂ ,罢了 ₁ ,了 ₂ ,嘞 ₃ ,哩 ₃ ,呢 ₂ ,哟 ₁	1 064
疑问	啊 ₃ ,啵 ₁ ,的 ₃ ,啦 ₃ ,来着 ₁ ,了 ₃ ,嘞 ₄ ,哩 ₄ ,吗 ₁ ,么 ₁ ,哪 ₂ ,呐 ₃ ,呢 ₃ ,似的 ₁	1 731
反诘	不成 ₂ ,就是了 ₁ ,吗 ₂ ,呢 ₄	266
假设	吧 ₂ ,的 ₄ ,的话 ₁ ,了 ₄ ,呢 ₅	514
揣测	吧 ₃ ,不成 ₃ ,喽 ₂ ,哪 ₂	418
祈使	啊 ₄ ,吧 ₄ ,了 ₅ ,喽 ₃ ,嘛 ₃	769
催促	啊 ₅ ,了 ₆ ,嘞 ₅ ,哟 ₂	241
忍受	罢了 ₂ ,喂 ₂ ,的话 ₂ ,而已 ₁ ,就是了 ₂ ,了 ₇ ,也罢 ₂ ,也好 ₂	881
不平	就是了 ₃ ,了 ₈ ,似的 ₂	529
论理	啊 ₆ ,来着 ₂ ,么 ₂ ,哪 ₃	535

Table 2 Model parameter settings

表 2 模型参数设置

参数	取值
Optimizer	Adam
Max_learning_rate	0.000 1
Min_learning_rate	0.000 01
Epoch	10
Batch_size	32
Dropout	0.1
WoBERT 词向量维度	768
语句特征维度	305

4.3 比较实验

传统词向量模型在序列标注领域已得了广泛应用,为了验证所提模型的有效性,实验过程中使用传统词向量来替代预训练模型表示文本特征,并结合深度学习中常用的 RNN、LSTM 及 BiLSTM 模型进行语气词分类与识别。

本文使用 Word2Vec 模型在语气词标注语料库上进行词向量训练,词向量训练的算法为 Skip-Gram,滑动窗口的大小为 5,迭代的次数为 100,词向量维度为 300。在训练完成后,利用词向量生成对应的词嵌入,并输入到其他模型中进行处理。此外,BERT、RoBERTa、MacBERT 也是序列标注领域常用的模型,实验中将其应用于语气词识别与分类任务,并与其他模型进行比较,实验结果如表 3 所示。

由表 3 可见,使用 Word2Vec 表示文本特征后,Word2Vec_BiLSTM 语气词识别效果最佳,F1 值为 91.14%,相较于 Word2Vec_RNN、Word2Vec_LSTM 模型分别提升 7.3%、0.12%。在对语气词进行分类时,Word2Vec_RNN、Word2Vec_LSTM 及 Word2Vec_BiLSTM 的 F1 值分别为 24.56%、24.89%、53.40%,整体分类效果较差。

在使用预训练模型表示文本特征时,MacBERT 在语气

Table 3 Experimental results of modal particle recognition and classification

表 3 语气词识别与分类实验结果 (%)

	Algorithm	Accuracy	Precious	Recall	F1
识别	Word2Vec_RNN	96.68	98.25	76.60	83.84
	Word2Vec_LSTM	97.95	98.88	85.59	91.02
	Word2Vec_BiLSTM	97.97	98.88	85.76	91.14
	BERT	99.32	96.09	98.90	97.45
	RoBERTa	99.20	95.26	98.93	97.01
	MacBERT	99.43	96.82	98.88	97.82
分类	本文模型	99.54	99.74	96.75	98.19
	Word2Vec_RNN	45.70	29.93	25.50	24.56
	Word2Vec_LSTM	49.07	31.71	25.84	24.89
	Word2Vec_BiLSTM	70.69	67.17	50.21	53.40
	BERT	88.14	81.81	79.25	80.32
	RoBERTa	87.37	80.51	79.06	79.56
	MacBERT	88.74	81.47	80.59	80.86
	本文模型	91.22	88.80	88.86	88.74

词识别与分类上效果最佳,F1 值分别为 97.82%、80.86%。相较于基于 Word2Vec 的模型,性能得到了明显提升,原因有:①Word2Vec 基于静态的词向量表示,每个单词在语料库中只有一个固定的向量表示,无法充分捕捉单词在不同上下文中的动态含义;②Word2Vec 模型主要关注局部上下文信息,无法捕捉文本中的长距离依赖关系,而在 BERT、RoBERTa、MacBERT、WoBERT 等预训练模型中采用动态方式表示文本特征,因此能够捕捉单词在不同上下文中的细微差别;③Transformer 结构能有效建模长距离依赖关系,产生的效果更好。

综上,本文模型整合了 WoBERT、BiLSTM、自注意力机制及外部语句特征,在语气词识别与分类上 F1 值相较于 MacBERT 分别提升 0.37%、7.88%。

4.4 消融实验

为了进一步验证所提各模块对提升模型性能的作用,将 WoBERT 作为基准模型与其他模型进行比较,实验结果如表 4、表 5 所示。由此可知,在基准模型 WoBERT 上组合 LSTM 与 BiLSTM 后,语气词识别与分类的 F1 值分别提升

Table 4 Experimental results of modal particle recognition

表 4 语气词识别的实验结果 (%)

Algorithm	Accuracy	Precious	Recall	F1
WoBERT	97.33	89.49	90.45	89.96
WoBERT_LSTM	97.43	89.84	90.93	90.38
WoBERT_BiLSTM	97.47	90.18	90.72	90.44
WoBERT_BiLSTM_Attention	97.50	90.05	91.29	90.66
WoBERT_BiLSTM_语句特征	99.32	97.11	97.81	97.46
WoBERT_BiLSTM_POS_Attention	98.28	90.75	97.96	94.00
WoBERT_BiLSTM_DEP_Attention	98.30	92.56	94.94	93.71
WoBERT_BiLSTM_SDP_Attention	97.55	90.29	91.44	90.86
WoBERT_BiLSTM_语句特征_Attention	99.54	99.74	96.75	98.19

Table 5 Experimental results of modal particle classification

表 5 语气词分类的实验结果		(%)			
Algorithm	Accuracy	Precious	Recall	F1	
WoBERT	76.06	77.48	63.76	69.70	
WoBERT_LSTM	76.53	76.74	64.69	70.04	
WoBERT_BiLSTM	77.54	77.75	65.08	70.80	
WoBERT_BiLSTM_Attention	73.74	75.76	61.72	67.39	
WoBERT_BiLSTM_语句特征	90.95	84.27	82.59	83.34	
WoBERT_BiLSTM_POS_Attention	91.37	83.76	83.32	82.90	
WoBERT_BiLSTM_DEP_Attention	88.35	84.54	79.91	82.01	
WoBERT_BiLSTM_SDP_Attention	66.22	81.54	60.86	69.54	
WoBERT_BiLSTM_语句特征_Attention	91.22	88.80	88.86	88.74	

0.42%、0.48% 和 0.34%、1.1%, 因为 BiLSTM 输出的结果包含更丰富的语义信息, 识别效果更好。在 WoBERT_BiLSTM 上添加 Attention 机制后, 语气词识别的 F1 值提升 0.22%, 但分类 F1 值下降 3.36%, 因为 WoBERT 模型已包含自注意力机制组件, 在不添加其他外部特征的情况下, 直接在模型添加自注意力机制可能会导致信息冗余, 使模型难以区分重要的信息, 导致性能下降。在 WoBERT_BiLSTM 上添加语句特征后, 语气词识别与分类的 F1 值分别提升 7.02%、12.54%, 证明了词性、依存句法、语义依存等外部特征对语气词识别与分类具有积极作用。

虽然, 深度学习模型具备了自动挖掘文本语义信息的能力, 但人工识别特征往往能提供更直接、更有针对性的信息。因此, 在构建模型时结合人工特征与深度学习技术往往能产生更优秀的效果。

在 WoBERT_BiLSTM_Attention 上分别添加词性、依存句法、语义依存特征后, 语气词识别与分类的 F1 值分别提升 3.34%、3.05%、0.2% 和 15.51%、14.62%、2.15%, 词性和依存句法特征相较于语义依存特征而言优化效果更好。在融合语句特征的 WoBERT_BiLSTM 模型基础上引入自注意力机制后, 识别与分类的 F1 值分别提高 0.73%、5.4%, 因为融合特征中各部分的重要程度并不相同, 可能存在冗余信息, 但自注意力机制能筛选出特征中的关键信息, 通过学习注意力权重可获得最优的特征表达。

综上, 引入自注意力机制不仅能提升模型对重要特征的敏感度, 还能有效降低冗余信息对识别性能造成的不良影响。

5 结语

本文研究了语气词的识别与分类任务, 基于 WoBERT、BiLSTM 模型和自注意力机制提出了融合语句特征的 WoBERT_BiLSTM_Attention 语气词识别与分类模型。实验表明, 本文方法相较于传统词向量模型识别与分类效果更优, 在引入语句特征与注意力机制后模型性能得到了显著提升。

未来, 将对更多语气词进行识别与分类, 还会扩大语料库的规模, 让模型在更丰富的语料上进行训练, 从而增强模型的泛化能力和分类精度, 以进一步提升模型识别与分类效果。

参考文献:

[1] QI H Y. Mood particles and mood systems[M]. Hefei: Anhui Education Publishing House, 2004.
齐沪扬. 语气词与语气词系统[M]. 合肥: 安徽教育出版社, 2004.

[2] LYU S X. A brief introduction to Chinese grammar[M]. Shanghai: Commercial Press, 1984.
吕叔湘. 中国文法要略[M]. 上海: 商务印书馆, 1984.

[3] NI Y. Research on the modal particles of modern Chinese at the end of a sentence and pragmatic function[J]. Journal of Mudanjiang College of Education, 2013(3): 19-21.
倪宇. 现代汉语句末语气词及其语用功能研究[J]. 牡丹江教育学院学报, 2013(3): 19-21.

[4] RANG Y C, MA J. Image caption-based multimodal alignment sentiment analysis model[J]. Data Analysis and Knowledge Discovery, 2024-04-12. <http://kns.cnki.net/kcms/detail/10.1478.G2.20240408.2010.002.html>.
壤雨辰, 马静. 基于图像字幕的多模态对齐情感分析模型[J/OL]. 数据分析与知识发现, 2024-04-12. <http://kns.cnki.net/kcms/detail/10.1478.G2.20240408.2010.002.html>.

[5] RAO D N, XU Z H, LIANG R S. Research on answer generation based on knowledge base question answering[J]. Computer Engineering, 2024-04-12. <https://doi.org/10.19678/j.issn.1000-3428.0068433>.
饶东宁, 许正辉, 梁瑞仕. 基于知识库问答的回答生成研究[J]. 计算机工程, 2024-04-12. <https://doi.org/10.19678/j.issn.1000-3428.0068433>.

[6] GAO M K. Chinese grammar theory[M]. Shanghai: Commercial Press, 1986.
高名凯. 汉语语法论[M]. 上海: 商务印书馆, 1986.

[7] HUANG B R, LIAO X D. Modern Chinese: revised edition[M]. Beijing: Higher Education Press, 1997.
黄伯荣, 廖序东. 现代汉语: 修订版[M]. 北京: 高等教育出版社, 1997.

[8] HE L S. A dictionary of ancient Chinese function words[M]. Beijing: Chinese Language Publishing House, 2006.
何乐士. 古代汉语虚词词典[M]. 北京: 语文出版社, 2006.

[9] WANG L. Modern Chinese grammar[M]. Shanghai: Commercial Press, 2010.
王力. 中国现代语法[M]. 上海: 商务印书馆, 2010.

[10] CHEN H L. The POS tagging system of ancient Chinese function words based on Bi-LSTM-CRF[D]. Wuhan: Huazhong University of Science and Technology, 2019.
陈火龙. 基于 Bi-LSTM-CRF 的古汉语虚词词性标注系统[D]. 武汉: 华中科技大学, 2019.

[11] ZHOU Y H, MU L L, ZAN H Y, et al. Research on automatic recognition of Chinese modality usage[J]. Computer Engineering, 2010, 36(23): 155-157, 161.
周溢辉, 穆玲玲, 簪红英, 等. 汉语语气词用法的自动识别研究[J]. 计算机工程, 2010, 36(23): 155-157, 161.

[12] WANG Y H, LIN M, LI Y L, et al. Research on embedded text topic model based on BERT[J]. Computer Engineering and Applications,

- 2023, 59(1): 169-179.
- 王宇晗, 林民, 李艳玲, 等. 基于BERT的嵌入式文本主题模型研究[J]. 计算机工程与应用, 2023, 59(1): 169-179.
- [13] YANG J C, MA C, XIAO M. Relation recognition of unmarked complex sentence based on feature fusion[J]. Computer Science, 2023, 50(S2): 69-74.
- 杨进才, 马晨, 肖明. 基于特征融合的无标复句关系识别[J]. 计算机科学, 2023, 50(S2): 69-74.
- [14] ZANG J, LU J T, WANG Y, et al. Chinese short text sentiment classification model integrating dual-channel features[J]. Computer Engineering and Applications, 2024, 60(21): 116-126.
- 臧洁, 鲁锦涛, 王妍, 等. 融合双通道特征的中文短文本情感分类模型[J]. 计算机工程与应用, 2024, 60(21): 116-126.
- [15] QI H Y. A dictionary of the usage of modal elements in modern Chinese [M]. Beijing: Commercial Press, 2011.
- 齐沪扬. 语气成分用法词典[M]. 北京: 商务印书馆, 2011.
- [16] GOROKHOVATSKYI O, VNUKOVA N, OSTAPENKO V, et al. Semantic-based clustering for education-science-business interaction bibliometric analysis[EB/OL]. <https://ceur-ws.org/Vol-3664/paper10.pdf>.
- [17] MONDAL S, BARMAN M, NAG A. Question classification task based on deep learning models with self-attention mechanism[J]. Multimedia Tools and Applications, 2024, 84: 1-30.
- [18] OU H, YU L, TIAN S, et al. An adversarial-example generation method for Chinese sentiment tendency classification based on audiovisual confusion and contextual association[J]. Knowledge and Information Systems, 2023, 65(12): 5231-5258.
- [19] CUI X, SONG C, LI D, et al. RoBGP: a Chinese nested biomedical named entity recognition model based on RoBERTa and global pointer [J]. Computers, Materials & Continua, 2024, 78(3): 3603-3618.
- [20] SHAYEGAN M J, KOJOURI S. A deep learning approach for sarcasm detection on Twitter[EB/OL]. <https://ijict.itrc.ac.ir/article-1-613-en.pdf>.
- [21] YANG J C, CAO Y, HU Q. Automatic recognition of causal complex sentence relationships in Chinese based on DPCNN model and sentence feature fusion[J]. Journal of Chinese Information Processing, 2022, 36(9): 19-27.
- 杨进才, 曹元, 胡泉. 基于DPCNN模型与语句特征融合的汉语因果类复句关系自动识别[J]. 中文信息学报, 2022, 36(9): 19-27.
- [22] LIU H T. Dependency grammar: from theory to practice [M]. Beijing: Science Press, 2009.
- 刘海涛. 依存语法的理论与实践[M]. 北京: 科学出版社, 2009.
- [23] YU J S, SHI J L, YANG L E, et al. Automatic transformation of enhanced dependencies in Chinese[J]. Journal of Chinese Information Processing, 2023, 37(10): 26-33.
- 余靖思, 师佳璐, 杨麟儿, 等. 汉语增强依存句法自动转换研究[J]. 中文信息学报, 2023, 37(10): 26-33.
- [24] LIU Z, TANG H, WANG N Z, et al. A text summarization method combining RNN and sparse self attention[J]. Computer Engineering, 2025, 51(1): 312-320.
- 刘钟, 唐宏, 王宁喆, 等. 融合RNN与稀疏自注意力的文本摘要方法[J]. 计算机工程, 2025, 51(1): 312-320.
- [25] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[DB/OL]. <https://arxiv.org/abs/1706.03762>.
- [26] ZHAN W D, GUO R, CHANG B B, et al. Development of Peking University CCL corpus[J]. Corpus Linguistics, 2019, 6(1): 71-86, 116.
- 詹卫东, 郭锐, 常宝宝, 等. 北京大学CCL语料库的研制[J]. 语料库语言学, 2019, 6(1): 71-86, 116.
- [27] ZHAN W D, GUO R, CHEN Y R. The CCL corpus of Chinese texts, available online at the website of center for Chinese linguistics (abbreviated as CCL) of Peking University[EB/OL]. http://ccl.pku.edu.cn:8080/ccl_corpus.
- [28] XUN E D, RAO G Q, XIAO X Y, et al. Development of BCC Corpus in the context of big data [J]. Corpus Linguistics, 2016, 3(1): 93-109, 118.
- 荀恩东, 饶高琦, 肖晓悦, 等. 大数据背景下BCC语料库的研制[J]. 语料库语言学, 2016, 3(1): 93-109, 118.

(责任编辑: 刘嘉文)