文章编号：

融合反问特征的卷积神经网络的中文反问句识别*

文治[1]，李旸[1]，廖健[1]，陈鑫[1]，王素格[1,2]

（1. 山西大学计算机与信息技术学院，山西 太原 030006；

2. 山西大学计算智能与中文信息处理教育部重点实验室，山西 太原 030006）

摘要：反问是一种带有强烈情感色彩的表达方式，对其进行自动识别将提升隐式情感分析的整体效率。针对汉语反问句识别问题，本文分析了反问句的句式特点，将反问句的句式结构融入到卷积神级网络的构建中，提出一种融合句式结构的卷积神经网络的反问句识别方法。首先利用置信度大于 70%的反问句的特征词、序列模式，对大规模未被标注的微博语料进行初步筛选，获取大量伪反问句。然后通过多个卷积核分别对句子的词向量和反问句的特征进行抽取，获取句子语义特征和反问词特征，将两者共同作用生成句子的表示。最后利用 softmax 分类器实现句子的分类。实验结果表明，利用该方法对微博中反问句的识别准确率、召回率和 F 值分别达到了 89.5%、84.2%和 86.7%。

关键词：反问句；卷积神经网络；特征；情感分析

中图分类号：TP391　　**文献标识码：**A

Rhetorical Questions Identification Based on Combining Convolutional Neural Network with Features

Wen Zhi[1],Li Yang[1], Liao Jian[1],Chen Xin[1],Wang Suge[1,2]

(1. School of Computer & Information Technology, Shanxi University, Taiyuan, Shanxi 030006, China; 2.Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Shanxi University, Taiyuan, Shanxi 030006, China)

Abstract: The Rhetorical Question is a kind of expression with a strong emotional color. Identified Rhetorical Questions automatically will improve the overall efficiency of implicit sentiment analysis. In this paper, Chinese Rhetorical Question identification method is proposed, through analyzing the sentence structure characteristics of Rhetorical Questions, and combining the sentence structure of Rhetorical Questions with the construction of convolutional neural network. Firstly, using the feature words and sequence pattern features of Rhetorical Question sentences with greater than 70% confidence, a large number of unlabeled microblog corpora are initially screened to obtain a large number of pseudo-words. Then, word vectors and the features of the Rhetorical Questions are extracted to generate the expression of sentences by multiple convolution kernels to obtain the semantic characteristics of the sentences and the features of the Rhetorical Questions. Finally, the softmax classifier is used to classify sentences. The experimental results show that the proposed method achieves the identification accuracy, recall rate, and F-measure of the Rhetorical Question in Weibo reach 89.5%, 84.2% and 86.7% respectively.

Key words: rhetorical questions; convolutional neural network; feature; sentiment analysis

1 引言

文本情感分析是指通过对文本的主观性分析发现作者所表达的情感倾向。依据文本中是否包含有显性情感词，将其又分为显示情感分析和隐式情感分析[1]。随着情感分析的研究不断深入，人们开始更多的关注隐式情感分析。我们通过对真实微博语料的标注，发现反问句

* **收稿日期：**　　**定稿日期：**

基金项目：国家自然科学基金资助项目（61632011, 61672331, 61603229, 61573231）

作者简介：文治（1993—），男，硕士研究生，研究方向为自然语言处理；李旸（1988—），女，博士研究生，研究方向为情感分析；廖健（1990—），男，博士，研究方向为情感分析；陈鑫（1992—）女，博士研究生，研究方向为自然语言处理；王素格（1964—），女，教授，研究方向为自然语言处理，情感分析。

在隐式情感句中占有 12.31%，可见对反问句的研究也是隐式情感分析一个不可或缺的问题。反问句作为一种间接言语行为[2]，所表达的真实意图是不同于其字面含义的。如果无法正确的识别反问句就可能对情感分析任务造成一定的偏差甚至错误。因此，如何有效地对反问句进行识别，将具有较大的应用价值，它可以帮助企业正确理解用户对其产品和服务的态度，并做出正确的决策。例如，“售后服务工作人员来电质问我们半天，要知道哪里坏了，怎么弄的话，还要你们干嘛呢？”，这条评论表达了用户对售后服务强烈的不满，如果公司能够正确地了解用户态度并及时做出补救策略，可以有效地避免公司客户的流失。

目前，对于汉语反问句的研究大都集中在语言学领域，包括反问句、疑问句与陈述句的区别、反问句句式结构、反问句的语用价值以及反问句的标记等[3]，并且取得了丰硕的研究成果。吕叔湘[4]先生就反问句和疑问句曾指出“反问和询问是作用的不同，在句子形式上并无差别。”，即反问句与疑问句就表面句式结构而言是非常相似的，若仅从句式结构对反问句识别就会造成混淆。刘钦荣[5]认为反问句有其特有的表达式，这为我们进行自动反问识别研究提供了一定的理论支持。他认为反问与疑问句的不同是在语义不发生变化的情况下，反问句的特指问和是非问是可以互相变换的，甚至特指问内部有的也可以互换。而疑问句则没有这样的特性。从反问句本身出发，黄伯荣等人[6]认为反问是无疑而问；明知故问即作者在已经知道了答案的情况下发出询问，但并不需要从他人那里获取答案。殷树林[7]则结合反问的性质特征得出了反问的定义，即指反问句是对其形式有所否定的无疑而问的问句。从语用的角度，许皓光[8]认为反问中的反问语气是至关重要的，他在大量反问句的研究基础上，对反问句语义语用功能进行了概述,并且提出了构成5种反问语气的主要因素有语言环境的作用、副词的使用、语句重音转移、语序变化和词语意义的虚化。另外，徐思益[9]、苏英霞[10]、史金生[11]和柴森[12]等从探讨反问句的结构特点入手，与一般疑问句的结构相比较，总结了一些反问句特有的表达式,并分析了它们的语义语用特点。此外,有些研究者从反问句标记[13-14]的角度对反问句进行研究，结合反问句特有的句式结构对反问句进行标记，并且总结了四类反问句的标记，分别是副词类、疑问代词类、双音化词和固定短语类。上述语言学家的大量研究成果虽不能直接运用到计算机自动识别反问句,但对反问句识别的任务提供了语言学基础，使得句式结构可以作为自动识别反问句重要特征之一。

反问句的自动识别任务要求计算机能够通过文本判别其是否为反问句。传统的词袋模型（Bag of words, BOW）在对句子表示时，仅仅把句子看作是词的集合，并不考虑句中词序和句法等要素，使得抽象词语的语义表示特征不够充分且存在表示向量高维稀疏等问题。而 Mikolov 在 2013 年提出了 Word2Vec[15]，该方法是通过对大量文本学习，利用上下文信息有效的提取词的语义特征，并将其映射到 n 维特征空间，再利用空间向量的距离，用于度量词语之间的语义相似度。这种方法能够对词语的语义信息进行有效表示，并将这种表示应用于多种自然语言处理任务，例如，分类、聚类、同义词获取和机器翻译等。卷积神经网络[16,17]（Convolutional Neural Network，CNN）具有捕获局部特征的能力，Yoon [18]曾在 2014 年首次将 CNN 应用于句子分类中，它通过利用 CNN 中的卷积层、池化层和全连接层抽取句子特征获得句子表示，最后通过一个分类器对句子进行分类。随后，在此基础上，Kalchbrenner 等人[19]提出了一种动态卷积神经网络（Dynamic Convolutional Neural Network，DCNN），采用 Dynamick-Max Pooling 返回 K 个最大值的子序列,在一定程度上保留了词序信息。与 CNN 不同，循环神经网络 RNN[20-22](Recurrent Neuron Network)的优势是对序列化数据进行建模，且能够捕获数据的长期依赖，但是由于它不能够并行运行，因此，RNN 的训练时间往往要比 CNN 花费的多。

因此，本文将反问句中特定的句式结构特征和 Word2Vec 词的分布式向量的语义信息进行联合表示，再结合卷积神经网络捕获特征的能力，建立了融合反问句特征的卷积神经网络模型，用于中文反问句进行自动识别。

2 反问句识别任务分析

反问句识别就是判断一个句子是反问句还是非反问句，我们将其看作二分类问题。其关键问题是考察反问句与非反问句存在哪些不同的特征。由文献[3]可知，反问句的句式结构、语义有鲜明的特点。在句式结构上，反问句有其特定的句式结构，语义上，反问句是对语义前提的判断，且反问句表达的真实意图与字面意思相反。因此，我们从结构和语义两个方面对反问句识别任务开展研究。

2.1 反问句的结构分析

本文从两方面对反问句的结构进行分析，即显式词语角度（反问句中包含的关键词）和特定句式结构。

2.1.1 反问包含的显式词语与符号特征

反问句中的显式词语与符号特征是指那些能够表明句子类型或是对反问句识别能够起到一定作用的句子成分，如“难道、怎么”等词语。语言学家曾从反问句标记的角度对反问的显式特征开展研究[9][15]，并取得了一定的成果。我们搜集、归纳和整理了常用的反问句显式特征，如表 1 所示。

表 1 反问的显式词语与符号特征表

显式词语与符号特征类型	显式符号特征
“不”	不都是, 不也是, 不就是, 不还是,
“何”	何妨, 何苦, 何不, 何尝, 何必, 何苦, 何曾,
“岂”	岂不是, 岂不, 岂止, 岂可,
“还”	还能, 还不, 还能不, 还不是
“怎么”	怎么能, 怎么会, 怎么不, 怎么,
“什么”	有什么, 凭什么, 为什么, 算什么,
“谁”	谁能, 谁不, 谁叫, 谁说,
“哪”	哪能, 哪里, 哪会, 哪还,
能愿动词	能, 会, 敢, 肯, 愿意,
句末语气	吗, 呢, 啊, 呀, 嘛,
标点符号	？ ！

2.1.2 反问的句式结构

反问句中存在一些特定的句式结构，语言学上也称之为反问句特有的句式结构。殷树林[23]等人曾通过对大量语料整理得出 17 种特定的反问句式结构。但他们的语料主要是来源于文学作品，本文结合对新浪微博中的语料中反问句的标注结果，又搜集整理出一些反问特有的句式结构，共得到 28 种特定句式结构。我们选取部分句式结构及对应的例句在表 2 中列出。

我们对搜集得到的 28 种反问特定的句式结构，在标注语料上进行置信度和支持度的计算。每条句式结构的置信度和支持度的计算方式如下公式（1）和（2），其中 *yt* 表示满足句式结构且是反问句；*nt* 表示不满足句式结构是反问句；*yf* 表示满足句式结构但不是反问句；*nf* 表示不满足句式结构也不是反问句。

$$confidence = \frac{yt}{yt + yf} \quad (1)$$

$$\sup port = \frac{yt}{yt + nt + yf + nf} \quad (2)$$

表 2 反问句特有的句式结构及其例句

反问特有的句式结构	例句
难道……不成？	难道你还怕他个外乡人不成？
能（会|肯|）+不+动词短语	如果那么便宜他会不买？
不就是……吗？	你们这些商人不就是想从客户那多挣点钱吗？
不就（也|还|又|正）……？	你这样不正遂了她的想法？
谁说（让|叫）……？	谁说女子不如男的？
……（岂|岂不|岂非）……？	鸟则择木，木岂能择鸟？
哪儿(哪里)+ 动词短语 + …… ？	大千世界，哪里没有野花的倩影？
怎么 + （能|会|肯） + …… ？	#匆匆那年#感触挺多的我们怎么会变成这样？
……还有[什么]……？	今年的春晚还有看头吗？
……[不怎么]……就[能不]……？	不是自家人，难道就不该互相关心吗？

本文对上述 2.1.2 中提及的 28 种特定句式结构进行了置信度和支持度计算，选取其中置信度大于 70%的特定句式结构，如表 3 所示。

表 3 特定句式结构的置信度和支持度

反问句的特定句式结构	置信度（confidence）	支持度（support）
……难道……[呢吗么呀]？	99.31%	4.13%
……能不……[呢吗么呀]？	88.33%	1.59%
……还有[什么]……？	75.00%	0.09%
……[不怎么]……就[能不]……？	93.75%	0.43%

从表 3 中可以看出，反问句的特定句式结构具有置信度较高且支持度较低的特点。如果我们利用这些句式结构进行反问句的识别任务，虽然能够保证识别句子的正确率，但是由于句式结构的低支持度，可能出现仅识别少部分的固定句式结构的反问句。

2.1.3 基于特定句式结构的反问句的数据获取

由于 CNN 模型虽然能够提取有效的特征，但对数据数量要求比较高，为了获取足够多的句子用于模型训练，若手工进行大量的语料标注，则需要耗费大量的时间和精力。因此，我们利用 3.1.2 节的句式结构，用于获取置信度较高的句子。受文献[24-25]利用表情符对情感语料自动标注的启发，本文利用具有高置信度反问句的特定句式结构，对未标注的语料进行自动标注，可以获取大量的伪数据。

2.2 反问句的语义分析

反问句是通过问句的形式进行确定性的陈述。根据文献[6]，反问句的特点是无疑而问、表示否定，即反问句并不表示真实的疑问，且反问句想要表达的真实意图是与其字面意思恰好相反。例如，下面的例句。

例句 1：这样缺斤少两的怎么让消费者去信任你的商品质量呢？

例句 2：难道在确保已出线时，就不能保留体力休养伤体吗？

上述例句 1 中，旨在说明由于缺斤少两是消费者无法信任产品质量，但句中并没有出现

任何对信任的否定修饰，而是通过缺斤少两的语义前提推断出对产品质量的不信任。整个句子通过肯定的字面含义表达其否定的真实意图。例句 2 中，作者想要表达在确保出线的前提下，应该保留体力休养，但他没有直接说明通过一种否定的形式来表达。通过上述两个例句，能够看出反问句存在语义前提，且反问句包含有字面意义和隐含意义，而隐含意义与字面意义恰好相反。

3 融合反问特征的卷积神经网络模型

CNN 已经被广泛的应用在自然语言的句子建模中。首先将词通过上下分布表示为固定维度的词向量。然后通过利用不同大小的过滤窗口对句子的矩阵进行卷积，获取能代表句子语义的局部特征，再通过池化层选取包含信息最丰富的特征来完成句子的建模。为了高效识别反问句，根据第 3 节介绍的反问句的特征，本文将反问特征融合到卷积神经网络结构中，用于识别反问句。该方法是通过卷积核对句子包含的反问词进行卷积操作，引入反问词的特征，然后再通过词向量卷积生成的语义特征共同作用生成句子的表示，最后通过 softmax 分类器对句子分类。下面将详细介绍融合反问特征的卷积神经网络模型。网络结构如图 1 所示。

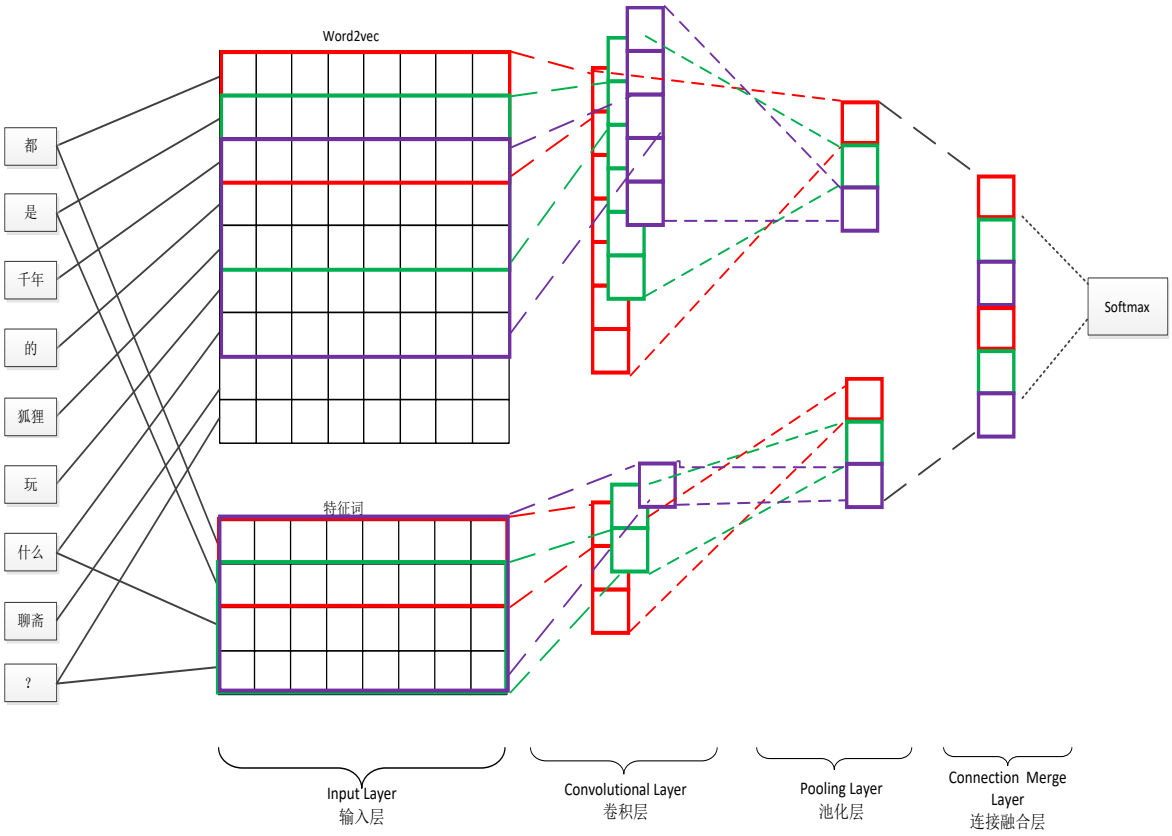

图 1 融合反问特征的卷积神经网络结构

设 S=(w_1,w_2,…, w_n)为由 n 个词组成，其中 w_i 表示第 i 个词。利用每个 w_i （i=1,2,…,n）生成 K 维向量，将 w_i 的向量进行拼接得到句子 S 的向量矩阵，其表现形式见公式（3）。

$$S=(w_1,w_2,\mathrm{L}\ \ w_n)=\begin{bmatrix} w_{11},w_{12},\mathrm{L}\ \ w_{1k} \\ w_{21},w_{22},\mathrm{L}\ \ w_{2k} \\ \mathrm{M} \\ w_{n1},w_{n2},\mathrm{L}\ \ w_{nk} \end{bmatrix} \quad (3)$$

利用公式（1），句子可以得到的结构化表示的句子向量矩阵 S，将其作为卷积神经网络

模型的输入。

（1）卷积层：基于卷积核的反问句语义特征提取。

利用不同大小的卷积窗口，卷积层对反问句的句子向量矩阵 S 进行卷积，将句子的局部语义信息进行融合。设卷积窗口为 c，窗口大小为 $l\times k$，w_j 为窗口向量，其形式见公式（4）。利用卷积窗口 c 对句子 S 进行卷积，卷积过程见公式（5）。利用 g 表示一个卷积窗口卷积生成新的语义融合向量。

$$W_j=(w_j,w_{j+1}\mathrm{L}\ w_{j+l-1})\qquad j\in(1,n-l+1) \tag{4}$$

$$\mathrm{g}=f\left(\sum_{j=1}^{n-l+1}c\cdot W_j+b\right) \tag{5}$$

其中，公式（4）中 n 表示句子的长度，公式（5）中表示向量对应元素的乘积，b 表示卷积核的偏置。$\mathrm{g}\in R^{n-l+1}$ 表示 $n-l+1$ 的列向量。f 表示激活函数，这里我们采用 $\mathrm{Re}LU$ 函数，使用 t 个卷积窗口对句子进行卷积，生成 t 个新的语义向量，将其拼接得到矩阵 $G_{(n-l+1)\times t}=(g_1,g_2,\Lambda\ g_t)$，矩阵 G 的转置变换 G^T 的展现形式见公式（6）。

$$G^{T}_{t\times(n-l+1)}=\begin{pmatrix}g_1\\g_2\\\mathrm{M}\\g_i\\\mathrm{M}\\g_t\end{pmatrix} \tag{6}$$

（2）池化层：反问句中有效语义信息特征提取。

该层通过池化操作选取最具代表反问句全部语义信息作为有效特征。这里采用最大池化操作，即从生成的语义向量中选取出最大值，认为它所包含的语义信息最为丰富，最能够代表反问句 S 的特征。用 M 表示池化操作后产生的语义特征向量具体操作见公式（7）。

$$M_{t\times(n-l+1)}=\begin{bmatrix}\max(g_1)\\\max(g_2)\\\mathrm{M}\\\max(g_i)\\\mathrm{M}\\\max(g_t)\end{bmatrix} \tag{7}$$

（3）连接层：融合语义特征向量提取。

连接层是将不同大小卷积窗口卷积和池化后得到的反问的语义特征向量进行融合，共同作用得到反问句 S 的表示向量。于是将上一步池化后得到的所有反问特征向量进行全连接操作，最终得到句子表示向量，此处用 S 表示，表现形式如公式（8）所示。

$$S=\sum_{i=1}^{t}W_i\cdot M_i+b \tag{8}$$

其中，M_i 表示第 i 个卷积窗口卷积池化操作后产生的向量，W_i 表示 $m\times d$ 的系数矩阵。

（4）反问句特征的卷积池化：反问句中的特征提取。

对反问句中存在特定的句式结构、标记及符号等特征。在对反问句的识别过程中，如果能够有效正确的识别这些特征，将有助于准确的判断句子的类别。于是本文利用上述 3.1.1 提及的反问显式词语与标记，抽取出句子中的反问标记，对抽取出的特征词按上述进行卷积、池化和全连接操作，得到新的特征向量，将其融入到句子表示中，以加强这些特征词对句子表示影响。对反问句的特征进行如上述（2）、（3）和（4）描述的卷积操作、池化操作和连接操作，最后得到特征表示 f。

（5）融合层：语义特征与反问特征进行融合。

该层是将上述经过操作（4）得到句子表示 S 以及经过操作（5）得到的反问特征表示 f 进行拼接，得到基于特征融合的句子表示 S_f，见公式（9），$\oplus$ 表示向量的拼接。

$$S_f = S \oplus f \tag{9}$$

（6）*softmax* 层：对句子表示进行反问句识别

该层经过操作（6）步骤，获取基于特征融合的句子表示 S_f，并将其输入到 *softmax* 层判断句子是否为反问句，其判断依据见公式（10）。

$$y = soft\max(WS_f + b) \tag{10}$$

4 实验设置

（1）数据集及评价指标

本文实验数据来源于手工标注的真实微博语料，经过人工标注的反问句 1732 条，非反问句 1732 条，共计 3464 条，其中非反问句也来源于真实的微博语料。所有的实验数据均经过去噪、去重和分词等预处理操作。使用 Word2Vec 来对词向量进行训练，训练过程采用交叉熵的损失函数。$y_i \in \{0,1\}$ 表示训练样本 x_i 的真实标签，$y_i \in [0,1]$ 表示预测标签 i 的概率，为防止过拟合采取 L_2 正则化。

实验过程中，我们随机将数据集的 80%作为训练集，10%作为验证集，10%作为测试集，反复实验 10 次，求取各个衡量指标的平均值。实验结果采用的评价指标为精确率（precision）、召回率（recall）和 F1 值。

（2）模型参数设置

本文利用的是卷积神经网络（CNN），其参数是在验证集上进行了大量实验，并进行参数调整。实验过程中的参数更新采用 Kingma[26]提出的 Adam Optimzation 进行梯度下降更新模型参数。通过多次实验，选取最优的模型参数：词向量维度为 300；过滤器的滑动窗口大小为 3, 4, 5；过滤器数量为 128；批处理大小（batch size）为 100；迭代次数为 50；学习率为 0.001。在 CNN[W]、CNN[F]、CNN[W+F]_1、CNN[W+F]_2 实验中，模型都采用了相同的参数。

5 实验比较与结果分析

5.1 伪数据对反问句判别影响实验

为了验证伪数据对实验的影响，本文在原有手工标注的数据集中分别尝试添加了 500、1000、1500、2000 条伪反问数据，为了保证数据的平衡，又添加了与之相应的非反问句。将这些真实数据和伪数据共同输入到本文提出的模型训练中，实验结果见表 4 所示。

表 4：不同伪数据的数量对比实验结果

伪数据数量	Precision	Recall	F1
0	**89.5%**	**84.2%**	**86.7%**
500	57.5%	89.2%	70.0%
1000	57.9%	91.0%	70.8%
1500	57.4%	91.7%	70.6%
2000	58.6%	91.0%	71.3%

从表 4 可以看出，若添加了伪数据，能够得到较高的召回率，是因为伪数据有特定结构；精确率却较低，是由于将真实数据和伪数据共同用于模型训练，而且随着伪数据的增多可能引入错误的信息也会增多。因此，在本文实验的数据中并没有使用伪数据。

5.2 八种方法对比实验

为了验证本文提出方法的有效性，选取如下的不同方法进行比较实验。

（1）基于规则的方法（RB）。该方法是利用上述 2.1.2 中所提及到 28 种的反问句的特定句式结构，首先将这些句式结构改写成正则表达式，然后利用正则表达式对大量文本进行匹配，若句子能够与任何一条正则表达式相匹配，则认为该句子是反问句，否则，不是反问句。

（2）基于 SVM 文本分类方法（SVM）。该方法主要是通过文本特征提取、文本特征表示、特征归一化处理将文本转化为特征向量，最后实现句子的反问句识别。即首先对文本进行分词、去停用词等预处理，所有的词看作是特征集。然后从这些特征集中选取具有代表性的词。本文采用卡方检验进行特征选取。对选取的特征进行归一化处理，计算得到特征权重。对于给定文本按照选取的特征转化成向量，在使用 SVM 进行分类。SVM 是通过结构分险最小化来找到一个最优超平面，使其达到最优分类效果。本方法的特征利用卡方检验选取 1000 个词作为特征集，利用 tf-idf 对选取的特征进行权重计算，实验采用的是由台湾大学开发的 LIBSVM[1]工具包，选用的 SVM 核函数是 RBF 核函数。

（3）基于 CRF 的反问句的识别。该方法将反问句的自动识别问题转化为序列标注问题，采用 IOB2 的标注集合来标记功能块。每个标记由两部分组成，第一部分表示功能块的位置，第二部分表示功能块的类型标记，两部分之间通过“-”连接。实验中采用的是 CRF++[2]。

（4）基于词语的双向循环神经网络（BiLSTM）[27]。该方法中模型超参数根据文献[28]，设置词向量的维度为 300，隐藏状态的维度为 1000，batch size 为 64，dropout 算法的比率设置为 0.5，模型的学习率设置为 0.001。

（5）基于词语的卷积神经网络（CNN[W]）[18]。该方法利用预训练的 Word2vec 将句子转化成连续的词向量，将其作为传统的 CNN 模型的输入。

（6）基于反问句的显示特征词、符号作为卷积神经网络（CNN[F]）输入特征。该方法利用第 3.1.1 节中提及的反问显式词语与标记，作为特征对句子进行表示。即将每个句子表示成基于特征词的词向量，然后输入到传统的 CNN 模型中。

（7）融合反问特征的卷积神经网络（CNN[W+F]_1）。该方法利用预训练的 Word2vec 转化成连续的词向量，再利用 3.1.1 节中的反问的显式词语和标记作为抽取反问句的特征，将词向量和特征在表示层进行拼接，得到融合反问句特征的词向量，然后将拼接后的向量作为 CNN 模型的输入，再进行卷积、池化和全连接操作。最后通过 softmax 分类器对句子进行分类预测。

（8）融合反问特征的卷积神经网络（CNN[W+F]_2）。该方法是利用本文 4.2.2 节提出

[1] https://www.csie.ntu.edu.tw/~cjlin/libsvm/
[2] https://sourceforge.net/projects/crfpp/

的融合反问特征的卷积神经网络的反问句识别。首先将句子利用预训练的 Word2vec 转化成连续的词向量，然后利用 3.1.1 节的反问显式词语与标记抽取出句子中反问句的特征，分别利用卷积核对词向量和特指特征的词向量进行卷积，将卷积生成的新的特征向量进行拼接，最后将拼接后的向量输入到 softmax 分类器中对句子进行分类预测。

利用上述设置的 8 个方法的实验，获得的实验结果见表 5

表 5：八种方法识别反问句的实验结果

实验方法	Precision	Recall	F1
RB	71.1%	47.5%	56.9%
SVM	53.1%	69.2%	60.1%
CRF	80.1%	77.6%	78.8%
BiLSTM	77.2%	77.0%	77.1%
CNN[W]	86.1%	81.8%	83.9%
CNN[F]	83.8%	84.3%	84.1%
CNN[W+F]_1	85.7%	84.7%	85.2%
CNN[W+F]_2	**89.5%**	**84.2%**	**86.7%**

从表 5 中，我们可以得出：

①SVM、BiLSTM、CRF 和 CNN[W]都是利用句子中的词语特征进行反问句识别，CNN[W]的实验效果明显优于 SVM、BiLSTM 和 CRF，这是因为基于 SVM 的反问句识别仅仅考虑了特征是否存在，并没有考虑句子的语义信息，BiLSTM 中虽然也考虑了句子语义信息，但由于微博语料中反问常常存在于一个分句中，若句子较长时，一些词语会随句子长度的增长对句子表示的影响较小，这样就会对句子的识别造成错误，而 CNN[W]是对含有语义信息的词向量进行操作，在卷积过程中将每个子句的语义信息充分融入到句子表示中。因此，实验效果明显优于 SVM、BiLSTM 和 CRF。

②RB 和 CNN[F]都利用了反问句的特征，但 RB 的实验效果较差，主要原因 RB 是基于正则表达式的匹配，仅考虑了句式结构且比较苛刻不够灵活，造成了反问句识别的召回率较低。通过比较 RB 和 CNN[F]可以看出，在对反问句进行识别时，使用语义信息可以提升反问句识别的效率。

③从 CNN 算法来看，四个 CNN 的对比实验结果在 F1 值指标下由优到劣的顺序为 CNN[W+F]_2 > CNN[W+F]_1 > CNN[F] > CNN[W]。对于同时考虑反问特征和词向量，CNN[W+F]_1 的精确率和召回率分别达到 85.7%、84.7%。CNN[W+F]_2 的精确率高达 89.5%，而召回率与 CNN[W+F]_1 基本持平，达到 84.2%。CNN[W+F]_2 算法的召回率低于 CNN[F]算法的召回率，这种结果可能是因为 CNN[W+F]_2 是先通过卷积抽取出词特征和反问特征共同用作句子表示，当较长句子包含的反问特征较少时，反问特征对句子表示影响较小，使得模型可能造成分类错误。

由此可知，在反问句识别时中，分别考虑句子的语义和反问的特征，可以对反问句进行有效的表示，从而提升反问句的识别效果。

5.3 错误分析

为了对模型分类错误进行分析，从模型分类错误的结果中随机选取了 50 个句子。根据造成模型分类错误原因的不同将其分为两类。结合例句对其进行说明。

例句 1：这样的三星手机的功效，怎么能不让人喜欢呢？

例句 2：伦敦奥运会在基本没怎么看的情况下，就要闭幕了？

例句 3：尊称路上头等舱的 w12 ，前脸很霸气有木有 ！

上述例句中，例句 1 是反问句，例句 2 明显不是反问句。但是由于在例句 1 和例句 2 中都存在有 3.1.1 节提及的反问的显式词语符号特征“怎么”，这些词语特征对句子的表示会产生一定的影响，使得模型将其都分类为反问句。造成这种错误的原因应是模型仅考虑了显式词语特征是否存在，没有考虑这些特征在句子中充当的成分，导致识别错误。如果能够在模型中考虑到反问句的句法结构的成分特征可能会对分类性能有所提升。

另一个明显的错误是对包含在例句 3 中“有木有”和“是不是”等这一类词的反问句中，使得模型不能对其进行识别。造成这一错误的原因可能是因为模型在一定程度上要依赖搜集的反问的显示特征，特征不足造成了模型识别的错误。因此，可以通过扩充反问的特征来解决这一类问题。

6 结束语

本文针对社交媒体中反问句，依据其特有的句式结构、上下文语义相关的特点，设计了融合反问特征的卷积神经网络的方法进行中文反问句的识别。对于反问语料较少的情况，可以利用反问句特有的特征搜集语料对模型进行训练。既能解决模型训练中语料不足的问题，又能缓解语料标注过程中大量时间和人力的浪费。将本文所提方法应用到社会媒体语料中进行反问句的识别，取得了不错的效果，进一步证明了本文所提出的方法的有效性。但本文仍有不足，在模型中仅仅使用了包含显示词语和符号特征的反问的句式结构，并没有考虑到反问句的句法结构特征。在今后研究中可以考虑将反问的句法结构特征和句子语义特征共同作用于句子的表示，并充分利用伪数据进行模型的训练，以提高反问句的识别性能。

参考文献：

[1] 赵妍妍, 秦兵, 刘挺. 文本情感分析[J]. 软件学报, 2010, 21(8):1834-1848.

[2] 冯江鸿. 反问句的语用研究[D]. 上海外国语大学, 2003.

[3] 刘钦荣. 反问句的句法、语义、语用分析[J]. 河南师范大学学报(哲学社会科学版), 2004, 31(4):107-110.

[4] 吕叔湘. 吕叔湘文集. 第一卷, 中国文法要略[M]. 商务印书馆, 1990.

[5] 刘钦荣. 反问句和询问句句法结构间的关系[J]. 沈阳师范大学学报:社会科学版, 1995(4):85-88.

[6] 黄伯荣, 廖序东. 现代汉语[J]. 1983.

[7] 殷树林. 现代汉语反问句研究[D]. 福建师范大学, 2006.

[8] 许皓光. 试谈反问句语义形成的诸因素[J]. 辽宁大学学报(哲学社会科学版), 1985(3):66-68.

[9] 徐思益. 反问句特有的表达式[J]. 渤海大学学报(哲学社会科学版), 1986(4):57-64.

[10] 苏英霞. “难道”句都是反问句吗?[J]. 语文研究, 2000(1):56-60.

[11] 史金生. 表反问的“不是”[J]. 中国语文, 1997(1):25-28.

[12] 柴森. 谈强调反问的“又”和“还”[J]. 世界汉语教学, 1999(3):65-69.

[13] 刘芳. 现代汉语反问句的标记研究[D]. 沈阳师范大学, 2011.

[14] 梁冠华. 现代汉语有标记反问句研究[D]. 曲阜师范大学, 2015.

[15] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality[J]. 2013, 26:3111-3119.

[16] 周飞燕, 金林鹏, 董军. 卷积神经网络研究综述[J]. 计算机学报, 2017,6(40):1-23.

[17] Gu C, Wu M, Zhang C. Chinese Sentence Classification Based on Convolutional Neural Network[J]. 2017, 261(1):012008.

[18] Kim Y. Convolutional Neural Networks for Sentence Classification[J]. Eprint Arxiv, 2014.

[19] Kalchbrenner N, Grefenstette E, Blunsom P. A Convolutional Neural Network for Modelling Sentences[J]. Eprint Arxiv, 2014, 1.

[20] Mikolov T, Karafiát M, Burget L, et al. Recurrent neural network based language model[C]// INTERSPEECH 2010, Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September. DBLP, 2010:1045-1048.

[21] Qian Q, Huang M, Lei J, et al. Linguistically Regularized LSTMs for Sentiment Classification[J]. 2016:1679-1689.

[22] Li F, Zhang M, Fu G, et al. A Bi-LSTM-RNN Model for Relation Classification Using Low-Cost Sequence Features[J]. 2016.

[23] 殷树林. 现代汉语反问句特有的句法结构[J]. 湖南科技大学学报(社会科学版), 2007, 10(3):101-105.

[24] Go A, Bhayani R, Huang L. Twitter sentiment classification using distant supervision[J]. Cs224n Project Report, 2009.

[25] Pak A, Paroubek P. Twitter for Sentiment Analysis: When Language Resources are Not Available[C]// International Workshop on Database and Expert Systems Applications. IEEE, 2011:111-115.

[26] Kingma D P, Ba J. Adam: A Method for Stochastic Optimization[J]. Computer Science, 2014.

[27] Ding Z, Xia R, Yu J, et al. Densely Connected Bidirectional LSTM with Applications to Sentence Classification[J]. 2018.

[28] 黄磊, 杜昌顺. 基于递归神经网络的文本分类研究[J]. 北京化工大学学报(自然科学版), 2017, 44(1):98-104.

作者联系方式：

作者一：文治，山西省太原市小店区坞城路 92 号山西大学计算机与信息技术学院，030006，18234091599，529093584@qq.com；

作者二：李旸，山西省太原市小店区坞城路 92 号山西大学计算机与信息技术学院，030006,15735170958，770624917@qq.com；

作者三：廖健，山西省太原市小店区坞城路 92 号山西大学计算机与信息技术学院，030006,15935619813，liaojian_iter@163.com；

作者四：陈鑫，山西省太原市小店区坞城路 92 号山西大学计算机与信息技术学院，030006,18734838988，1315614497@qq.com；

作者五：王素格，山西省太原市小店区坞城路 92 号山西大学计算机与信息技术学院，030006,13934649855，wsg@sxu.edu.cn。