基于特征融合的汉语被动句自动识别研究

胡康[1] ， 曲维光[1*] ， 魏庭新[2] ， 周俊生[1] ， 顾彦慧[1] ， 李斌[3]
(1.南京师范大学 计算机与电子信息学院/人工智能学院，江苏省 南京市 210023;
2.南京师范大学 国际文化教育学院，江苏省 南京市 210097;
3.南京师范大学 文学院，江苏省 南京市 210097;
*通讯作者，Email：wgqu_nj@163.com)

摘要

汉语中的被动句根据有无被动标记词可分为有标记被动句和无标记被动句。由于其形态构成复杂多样，给自然语言理解带来很大困难，因此实现汉语被动句的自动识别对自然语言处理下游任务具有重要意义。本文构建了一个被动句语料库，提出了一个融合词性和动词论元框架信息的PC-BERT-CNN模型，对汉语被动句进行自动识别。实验结果表明，本文提出的模型能够准确地识别汉语被动句，其中有标记被动句识别F1值达到98.77%，无标记被动句识别F1值达到96.72%。

关键词： 汉语被动句； 自动识别； 特征融合； 语料库

Automatic Recognition of Chinese Passive Sentences Based on Feature Fusion

HU Kang[1], QU Weiguang[1*], WEI Tingxin[2], ZHOU Junsheng[1], GU Yanhui[1], LI Bin[3]
(1.School of Computer and Electronic Information/School of Artificial Intelligence,
Nanjing Normal University, Nanjing, Jiangsu 210023, China;
2.International College for Chinese Studies, Nanjing Normal University, Nanjing, Jiangsu 210097, China;
3.School of Chinese Language and Literature, Nanjing Normal University, Nanjing, Jiangsu 210097, China;
*Corresponding, Email: wgqu_nj@163.com)

Abstract

Chinese passive sentences can be categorized as marked and unmarked passive sentences according to the presence or absence of markers. Due to its diverse morphological variations, it brings great difficulties to Natural Language Understanding (NLU), and the automatic recognition of it is of great significance to the downstream tasks of Natural Language Processing (NLP). In this paper, we firstly construct a passive sentence corpus, then propose a PC-BERT-CNN model incorporating part-of-speech features and verb's argument frame information to automatically identify Chinese passive sentences. The experimental results show that our model can identify Chinese passive sentences more accurately than the existing automatic parsing tool LTP, the F1 values of marked and unmarked passive sentences recognition task reach 98.77% and 96.72% respectively.

Keywords: Chinese passive sentences , automatic recognition , feature fusion , corpus

基金项目：国家社会科学基金重大项目(21&ZD288)

1 引言

作为一种用于说明主语和谓语之间关系的语法形式，被动句广泛存在于各种自然语言中。与主动句中主语做动作的施事不同，被动句的主语由动作的受事充当。由于语言的类型和特征存在差异，不同语言中的被动句在表示方法上也不尽相同。例如在英语这种重形合的语言中，其被动句大多数是结构被动句(syntactic passive)(蒋坚松, 2002)，即谓语含有“be+v-ed”形式标记的句子，因此在机器理解时，容易通过动词形态变化以及助动词来识别句子语态。而汉语是一种重意合的语言，缺乏动词形态上的变化，虽然也存在一部分有标记被动句（如“被”字句），但与英语不同的是，汉语中还存在大量不含标记词的无标记被动句，这类句子的被动意隐含在词汇或语境中，这给汉语被动句的自动识别造成困难。在以下五个例句中，例1、5为非被动句，例2~4为被动句。其中，例1和例2从句法结构上看，都是主谓补结构，但例1是主动句而例2为无标记被动句，因此在判断句子中各成分语义关系时容易混淆；例3和例4分别是含标记词“被”和“让”的有标记被动句；例5虽然含“让”，但它是该句的动词而非作为被动标记的介词，被动标记词的这种同形多义现象给被动句的识别带来很大挑战。

例1 我吃完了。
例2 饭吃完了。
例3 我的钱包被偷了。
例4 我让门槛绊倒了。
例5 他让这位老人先走。

语体	会话	小说	新闻	学术	均值
有标记被动句	3.90	6.69	9.30	4.50	6.10
无标记被动句	3.55	9.19	3.65	2.56	4.70
总计	7.45	15.88	12.95	7.06	10.8

Table 1: 汉语被动句每万字出现次数

目前对被动句的研究主要集中在语言学层面。在自然语言处理领域，目前主流的自动解析器可以对汉语句子，也包括被动句，进行句法、语义层面的解析。然而，通过这些解析器得到的句子解析结果并不能直接判定被动句，而需要人为根据标注规范制定相应的判断规则或条件，再进行被动句判定。这导致被动句识别，尤其是无标记被动句识别任务的性能大大降低。

宋文辉等人 (2007)统计了被动句在会话、小说、新闻及学术四种语体中的分布情况如表 1所示，可见在这四种语体中，被动句在小说和新闻语体中所占比例最高，因此本文选取《人民日报》新闻语料作为被动句识别的基础语料进行人工标注，并提出一种基于深度学习的方法实现汉语被动句的自动识别，可以在大规模的语料中迅速、准确地筛选出被动句，以期对句法解析、AMR解析等下游任务提供支持。

本文主要贡献如下：第一，构建了一个包含13530条句子的被动句语料库，用于深度学习模型训练；第二，提出了一个PC-BERT-CNN(POS and CPB enhanced BERT-CNN Model)模型，实现了汉语被动句的自动识别，该模型对BERT模型进行词性增强并融入了中文命题语料库(Chinese Proposition Bank, CPB)中的动词论元框架信息；第三，实验结果表明，本文提出的模型在被动句自动识别任务上取得了较好的性能，其中有标记被动句识别的F1值达到98.77%，无标记被动句识别的F1值达到96.72%。

2 相关工作

2.1 被动句本体研究现状

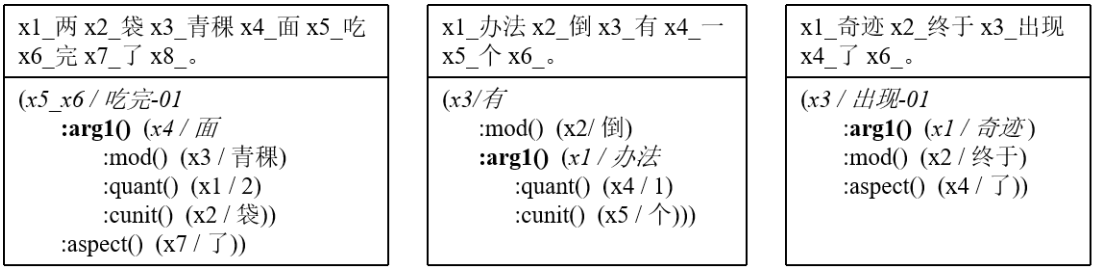
被动句在汉语中随处可见。近年来关于被动句的研究大都集中在语言学领域，研究对象主要涉及被动句的构式、语义和语用等方面。邹丽玲 (2016)从英译汉的视角解析了汉语无标记被动句表达被动含义时，不加“被”等被动标记词的原因。王灿龙 (1998)根据无标记被动句中被动语态主要用于及物动词的特点，从单音节及物动词和双音节及物动词两个方面分析能够进入无标记被动句的词语。王芸华 (2014)从不同角度考察了被动句主语的语义角色，认为被动句的主语除受事之外，也能由与事、主事、结果等其他语义角色充当。汤敬安 (2016)对比有标记被动句和无标记被动句之间的构式差异，并对二者的认知过程进行研究，指出二者的差异主要体现在语言结构中的受事注意范围、详略度、扫描方式等方面。对于有标记被动句标记词的界定问题，不同学者的看法也不尽相同，李珊 (1994)认为可以将标记词划分为“被、叫、让、给、为、被/让/叫...给、被/为...所”7类，而乔莎莎 (2015)则将标记词的数量进一步扩充到11种，分别为“被、叫、让、给、蒙、由、被...给、叫...给、让...给、被...所、为...所”，鞠彩萍 (2007)通过语料证实，“遭”在某些条件下也能作为标记词使用。

第二十一届中国计算语言学大会论文集，第384页-第394页，南昌，中国，2022年10月14日至16日。

2.2 被动句语料库构建现状

现有的中文语料库通常面向句子整体的句法、语义结构进行标注，而没有针对被动句的标注语料。被动句在这些语料中的数量较少，而且被动句在这些语料中并没有被显式标注。

例如中文抽象语义表示（Chinese Abstract Meaning Representation, CAMR）标注了动词与论元之间的关系 (李斌等, 2017)，它对被动句的标注方式是把位于主语位置的受事标注为动词的arg1，如图 1 中(a)所示，“面”标注为动词“吃”的受事arg1。不过这一标注方式并非被动句独有，如(b)和(c)中的论元结构虽然与(a)一致，但句子(b)是一个话题做主语的非被动句，句子(c)中的动词“出现”不具有被动用法，也是一个非被动句。因此CAMR语料库不能直接作为被动句数据集使用，本文将采用人工标注的方法构建一个被动句语料库，用于模型训练。



(a) “两袋青稞面吃完了。” (b) “办法倒有一个。” (c) “奇迹终于出现了。”

Figure 1: CAMR语料库标注的句子

2.3 被动句识别研究现状

目前针对被动句解析的专项研究较少，但在句法或语义解析任务中，可以通过判断解析图中语义角色或论元关系结构完成对被动句的识别。为探究现有自动解析器在被动句识别上的效果，本文使用哈工大发布的语言技术平台（LTP）[1](Che et al., 2010)对句子的句法结构和语义角色分别进行解析。对于依存句法解析结果，若解析图中存

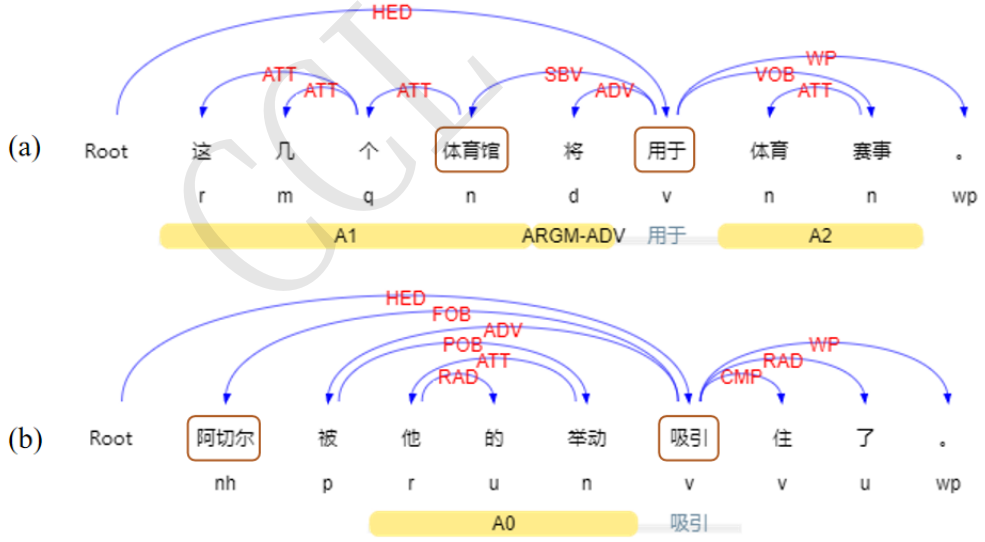


Figure 2: LTP对句子的自动解析

在“FOB←V”或“FOB←V→ADV(→POB)”[2]关系链时，把该句子记为正确解析，否则记为错

[1] http://ltp.ai/index.html

[2] FOB代表前置宾语，在被动句中通常指向受事；ADV代表状中结构，在有标记被动句中通常指向标记词；POB代表介宾关系，在有标记被动句中通常指向施事，可不出现在句子中。

误解析；对于语义角色解析结果，若动词及其受事论元A1、施事论元A0都判断正确，则记为正确解析，若有至少一个论元错误或没有解析出论元则记为错误解析。通过分析发现，LTP对被动句的解析性能并不高（具体统计数据见5.3节）。图 2 展示了LTP对两个被动句的自动解析结果，句(a)是一个无标记被动句，动词“用于”的A1解析正确，即LTP语义角色解析正确。但“体育场”应是“用于”的前置宾语(FOB)而非主谓关系(SBV)。句(b)中存在”FOB←V”关系链，即“阿切尔”是“吸引”的前置宾语，句法解析正确，但语义角色解析结果中，A1“阿切尔”却未被解析。这说明LTP在被动句自动解析方面仍存在一定的提升空间。

本文将汉语被动句的自动识别任务视为一个文本分类问题。近年来，深度学习在文本分类中的运用日渐成熟。Kim (2014)最先在卷积神经网络（Convolutional Neural Networks，CNN）的基础上提出TextCNN用于文本分类。Devlin等人 (2019)提出的BERT基于Transformer的双向深度语言模型，可以捕获双向上下文语义，在机器翻译、文本分类、文本相似性等多个自然语言处理任务中取得优异的表现。Qin等人 (2020)提出了一种特征投影的方法，将现有特征投影到共同特征的正交空间中，生成的投影垂直于共同特征并且对分类更具辨别力。朱向其等人 (2021)提出了一种基于改进词性信息和ACBiLSTM的中文短文本分类模型，在THUCNews数据集上取得良好的分类性能，准确率和F1值分别达到97.43%和95.24%。Nguyen等人 (2021)提出了两种利用外部知识的方法来丰富句子的语义表示，进而识别与恶意软件相关的句子，使得SVM分类器的F1值提升了9%。受此启发，对于本文的被动句识别任务，也可引入相应的外部知识并进行特征融合，以提升分类模型的性能。

3 被动句语料库构建

3.1 被动句语料来源

本文选取1998年1至3月《人民日报》新闻语料作为被动句标注的基础语料。由于《人民日报》语料中每一行文本代表一个自然段或标题，所以需要把提取出的纯文本语料切分为单个句子，去除新闻标题等不符合完整句式特点的文本，构成待标数据集共5万句，约200万字。

3.2 被动句的标注规范

定义“NP1+Mark+[NP2]+V”和“NP1+V”分别为有标记被动句和无标记被动句的一般构式，其中V为及物动词，N1代表动作的广义受事，N2代表对应的广义施事，Mark代表标记词，[ ]中的成分可不出现。被动句标注的重点在于确定主语和谓语之间的被动关系，进而根据有无标记词分为有标记被动句或无标记被动句。

3.2.1 有标记被动句

根据前人研究，本文把被动标记词的范围限定在“被、叫、让、给、蒙、由、为、遭”这八大类。当句子中动词对主语造成“强影响”或“不如意”的结果时，我们认为该句子很可能是被动句，具体来说，当句子主语充当谓语动词的以下几种论元角色时，可将句子标注为有标记被动句。

（1）**受事、与事论元** 受事代表直接受动词支配或影响的人或事物；与事是动作非主动的参与者。如例6中，

例6 (a)他的钱包被人偷走了。 (b)她被骗去不少钱。

“钱包”是动词“偷”的受事，“她”是“骗”的与事，两个句子均表达了“不如意”的结果，因此将这类句子标注为有标记被动句。

（2）**感事、主事论元** 感事是非自主的感知性事件的主体；主事指性质、状态或变化性事件的主体，与系事相对。如例7中，

例7 (a)他被一阵枪声吓醒了。 (b)埃雷迪亚被称为“花城”。

感事“他”由表示消极或突发反应的词——“吓”支配，表达一种“不如意”的状态，因此认为该句子是有标记被动句。(b)中“埃雷迪亚”是主事，“花城”是系事，二者通过“称为”这一评定、认同性的动词联系起来，这类句子也被视为有标记被动句。

（3）**工具、材料论元** 工具和材料都在动作发生过程中会受到控制，同时会对它们本身造成“强影响”。如例8中，

例8 (a)刺刀都被他捅弯了。 (b)乳胶漆被他涂了墙。

第二十一届中国计算语言学大会论文集，第384页-第394页，南昌，中国，2022年10月14日至16日。

“刺刀弯了”是在陈述“捅”这一动作过程中给工具造成的强影响。同理，“乳胶漆”作为“涂”的材料论元，被附加在别的物体上，也受到了“强影响”，因此把它视为被动句。

3.2.2 无标记被动句

在无标记被动句中，由于缺少被动标记，导致句子对动词有严格的语义选择限制，因此在标注无标记被动句时，除了观察句子主语充当的语义角色外，还需考虑动词本身的语义特点。当句中动词具有以下两种特点时，则把该句子标注为无标记被动句。

（1）**可控性** 表示动作可以由动作发出者根据自己的意愿进行控制。如例9中，

例9 (a)桌子已经擦了。 (b)这棵树已经死了。

动词“擦”可控而“死”不可控，因此(b)不属于无标记被动句。

（2）**强动作性** 表示的是人、其他生命体的动作或某种自然力造成的动作，动作性比较明显。如例10中，

例10 (a)作业做完了。 (b)脚冻伤了。 (c)人们的安全意识提高了。

“做”和“冻”都是动作性比较强的动词，(a)和(b)均属于无标记被动句；而“提高”的动作性弱，因此(c)不属于无标记被动句。

3.2.3 特殊情况

当某动词具有多种义项时，需要根据语境判断该动词在当前句子中的语义，进一步判定句子是否属于被动句。如例11中，

例11 (a)比赛中他被对手摔倒了。 (b)他走路时不小心摔倒了。

虽然两个句子都表达“不如意”的结果，但(a)中“摔倒”属于自主性动词，具有可控性；而(b)中的“摔倒”则属于非自主动词，不具有可控性；因此前者属于被动句而后者不是。

此外，若一个句子同时符合有标记被动句和无标记被动句的构式，则将其拆分或改写为多个句子，使得每个句子都是单标签样本。如例12中既包含有标记被动语态“心+被+打动”，又包含无标记被动语态“生源+积聚”，因此将它拆分为有标记被动句(a)和无标记被动句(b)。

例12 久而久之,下岗人的心被打动了,生源也就慢慢积聚起来了。

(a) 久而久之,下岗人的心被打动了。

(b) 生源也就慢慢积聚起来了。

3.3 被动句语料库统计

标注人员为2名具有语言学专业背景的志愿者，经过培训后根据上述标注规范展开标注工作。标注过程分为两个阶段，第一阶段为试标阶段，2人同时标注100条句子，完成后进行一致性统计的结果为85%，二者的标注结果存在的差异主要体现在动词在不同语境下的语义理解上。在经过适当调整后，进行第二阶段的正式标注，共13530条句，整体标注的一致性达到91%。语料库中约5%的句子由拆分或改写得到，各类别样本数据如表 2 所示。

句子类别	样本数量(条)	样本示例
有标记被动句	4495	大火已于傍晚六时被扑灭。
无标记被动句	4570	检查情况已在当地曝光。
非被动句	4465	她将参加钢琴组的比赛。

Table 2: 被动句语料库构成

此外，对有标记被动句的八大类标记词进行了统计，如表 3 所示，可知，“被”是最典型的被动标记词，“由”字句也比较常见，而其他几类标记词在新闻领域文本中出现的频率较低。

标记词	被	由	为	给	遭	让	叫	蒙	总计
数量(条)	3064	1238	145	34	7	4	3	0	4495
占比(%)	68.16	27.54	3.23	0.76	0.16	0.09	0.07	0.00	100

Table 3: 被动句语料库中的标记词分布

4 被动句识别模型设计

本文将汉语被动句识别任务建模为一个三分类问题，即把句子分为有标记被动句、无标记被动句和非被动句三种类别，提出一个融合词性信息和论元框架信息的PC-BERT-CNN模型，从而对汉语句子所属类别进行预测。模型的输入为单个句子，输出为句子的类别标签。模型由4个模块组成：①词性增强的词嵌入模块，得到融合了词性信息的BERT向量表示；②CPB论元框架信息提取模块，得到句中所有动词的论元框架信息的静态词向量表示；③特征提取和拼接模块，得到融合词性特征和动词论元特征的句子表征；④标签预测模块，得到模型的输出，即句子的预测类别标签。模型整体架构如图 3 所示。

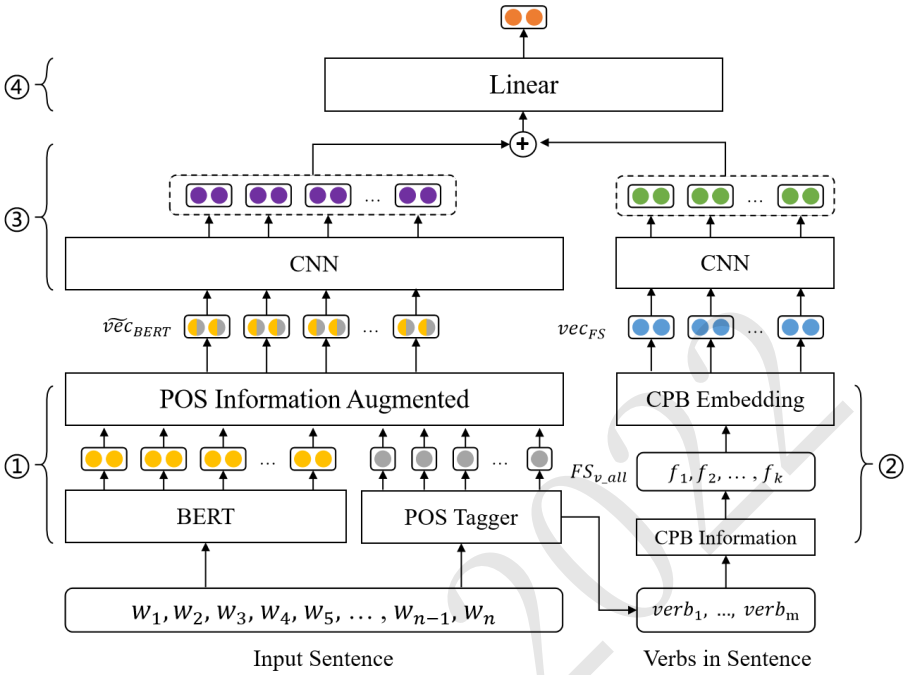


Figure 3: 被动句识别模型结构图

4.1 词性增强的词嵌入

被动句的核心是谓语动词，不同词性的词语具有不同的表征能力。在被动句识别任务中，谓语动词、施受事名词和标记介词是体现被动语态的重要组成部分。BERT在获取词嵌入时，利用自注意力机制充分考虑了每个字符的上下文信息，从而动态调整其向量表示，因此句子中每个词语的属性对模型的训练也具有重要作用。在得到BERT词向量后，利用结巴分词工具[3] 对句子进行分词和词性标注，根据本文研究对象把词性划分为动词(V)、名词(N)、介词(P)和其他(O)四类，并给各个词类分配权重，记为词性因子，如表 4 所示。

词类	说明	词性因子	取值范围
V	谓语动词	α	[0,1]
N	代表施事、受事的名词/代词	β	[0,1]
P	被动标记词	γ	[0,1]
O	其他修饰性的词语	δ	[0,1]

Table 4: 各类词性因子说明

各词性因子值为实验的超参数，设它们满足约束条件$\alpha + \beta + \gamma + \delta = 1$，各词性因子的最优取值通过参数分析得到。对于输入样本$S = [w_1, w_2, w_3, \ldots, w_n]$，首先经过BERT得到样本的表征向量$vec_{BERT}$，同时进行词性标注并分配权重，得到样本的词性因子序列$weight_{POS}$，然

[3] https://github.com/fxsjy/jieba

后将vec_{BERT}与$weight_{POS}$进行权重计算，得到融合词性信息的表征向量$\widetilde{vec}_{BERT}$，计算公式如(1)-(3)所示。

$$vec_{BERT} = BERT_{Embedding}(S) = [\vec{x}_0, \vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n] \tag{1}$$

$$weight_{POS} = [wp_0, wp_1, wp_2, \ldots, wp_n], wp_i \in [0, 1] \tag{2}$$

$$\widetilde{vec}_{BERT} = vec_{BERT} \times weight_{POS} = [\vec{y}_0, \vec{y}_1, \vec{y}_2, \ldots, \vec{y}_n] \tag{3}$$

其中$\vec{y}_i = \vec{x}_i \times wp_i$，“×”代表向量与标量的乘法运算。

4.2 CPB论元框架信息提取

CPB语料库对动词的论元框架，即动词义项进行了描述(Xue and Palmer, 2005)，每个Frameset对应动词的一种义项。图 4 列出了“评为”、“殉职”和“处理”三个动词在CPB中的论元框架信息，其中f1为该动词的第一种义项，f2为第二种义项。由于能够进入被动语态的动词应至少具备arg0和arg1这两个论元，因此论元框架中只含一个论元的动词是无法在被动句中使用的，如“殉职”。此外CPB动词论元框架中各个论元语义的描述对动词本身的语义也能起到很好的补充作用。故本文将动词的论元框架信息，即所有的Frameset从CPB中抽取出来，编码为词向量后进行特征融合，以期对模型性能的提升有积极的影响。

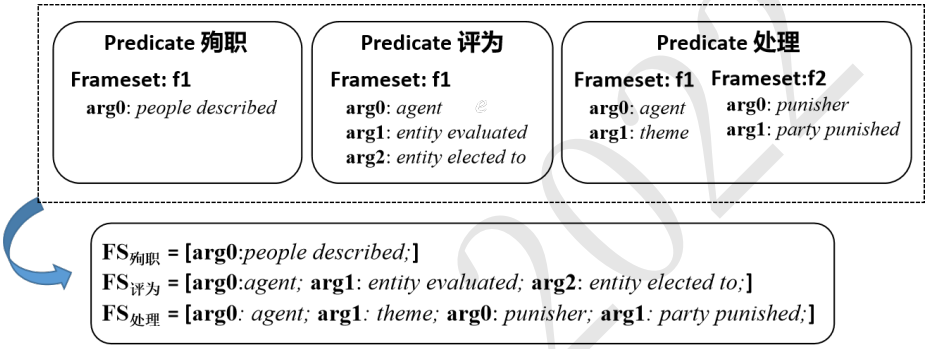


Figure 4: 从CPB语料库抽取论元框架信息

对于输入样本S，从词性标注结果中提取出当前样本句子中的所有动词$[verb_1, \ldots, verb_m]$，再从CPB语料库中抽取出每个动词的论元框架信息FS_i并进行字符串拼接，若在CPB语料库中未匹配到该动词的论元框架，则FS_i用空字符串替代，最终得到句子中所有动词的论元框架信息FS_{v_all}，由k个CPB动词论元框架描述信息的符号序列组成。再利用随机初始化的静态词嵌入层将FS_{v_all}转为向量表示，记为vec_{FS}，计算公式如(4)-(5)所示：

$$FS_{v\_all} = FS_1 + \ldots + FS_m = [f_1, f_2, \ldots, f_k] \tag{4}$$

$$vec_{FS} = Embedding(FS_{v\_all}) = [z_1, z_2, \ldots, z_k] \tag{5}$$

其中$k = \sum_{j=1}^{m} str_len(FS_j)$，“+”代表字符串加法。

4.3 特征提取和拼接

对于样本句子的词性增强BERT向量表示$\widetilde{vec}_{BERT}$和句中动词论元框架信息的向量表示vec_{FS}，利用卷积神经网络(CNN)对其进行特征提取，得到两个输出$\widetilde{F}_1$和$\widetilde{F}_2$，再将二者进行拼接，得到输入样本最终的特征表示$\widetilde{F}_s$，计算公式如(6)-(8)所示。

$$\widetilde{F}_1 = CNN(\widetilde{vec}_{BERT}) \tag{6}$$

$$\widetilde{F}_2 = CNN(vec_{FS}) \tag{7}$$

$$\widetilde{F}_s = \widetilde{F}_1 \oplus \widetilde{F}_2 \tag{8}$$

4.4 标签预测

将样本最终的特征$\widetilde{F}_s$送入全连接层，之后对网络进行dropout处理，以防止出现过拟合现象，最终经过softmax函数得到的结果即为样本预测的类别标签。

5 实验

5.1 数据集划分及实验设置

数据集划分 本文的实验语料包含三个类别的数据，按照6:2:2的比例将数据集进行切分并随机打乱，分为训练集、验证集和测试集，各类别样本个数如表 5 所示。

类别	训练集(条)	验证集(条)	测试集(条)
有标记被动句	2697	899	899
无标记被动句	2742	914	914
非被动句	2679	893	893

Table 5: 数据集划分

超参数设置 实验使用的BERT[4]模型版本为Chinese L-12 H-768 A-12，学习率设置为1e-5，其他超参数设置如表 6 所示。

超参数	含义	值
epochs	数据集迭代次数	3
batch_size	单批次样本数量	32
pad_size	每个样本最大token数量	128
filter_sizes	卷积核尺寸	(2,3,4)
num_filters	卷积核数量	256
dropout	丢弃概率	0.1
cpb_embed_dim	cpb信息词嵌入维度	16
cpb_pad_size	cpb信息最大token数量	12
cpb_filter_sizes	cpb卷积核尺寸	(2,3)
cpb_num_filters	cpb卷积核数量	256
pos_weight	词性因子取值	(0.3,0.3,0.3,0.1)

Table 6: 超参数设置

评价指标 本文实验采用准确率P(Precision)、召回率R(Recall)和F1值(F1-measure)作为模型性能的评价指标。对比实验则使用正确率Acc(Accuracy)作为评价指标。

5.2 实验结果及分析

在文本分类领域，基于CNN和BERT预训练模型的分类方法是比较常用且高效的方法。为了验证模型的有效性，将本文提出的PC-BERT-CNN模型与TextCNN、BERT和BERT-CNN三种模型进行实验对比，对每个类别句子的识别性能分别计算其P、R和F1值，得到的各模型在各类句子自动识别任务上的实验结果如表 7 所示。

可以看出，本文提出的模型在三类句子上的性能均达到最佳。观察TextCNN模型与BERT-CNN模型的实验结果，容易发现，在引入BERT预训练模型之后，各类句子识别的F1值都有了明显的提高，原因在于TextCNN使用的是静态词向量，不能解决文本中一词多义的问题，而BERT在动态生成词向量的过程中考虑了更多的上下文信息，使得在CNN捕获局部特征时，能够更好地理解句子语义。被动句识别任务与普通句子文本分类任务不同，它重点关注的对象是动词以及施受事等，相比BERT-CNN模型，本文模型在融入词性信息和动词论元框架信息后，丰富了句子中动词及其相关成分的语义表示，因此分类性能明显提升。

[4] https://github.com/google-research/bert

模型	有标记被动句识别			无标记被动句识别			非被动句识别		
	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)
TextCNN	95.85	**98.87**	97.34	91.38	93.39	92.37	**96.01**	90.89	93.38
BERT	**99.21**	97.55	98.37	94.64	96.50	95.56	95.73	95.41	95.57
BERT-CNN	99.10	98.33	98.72	**97.40**	94.20	95.77	93.66	**97.54**	95.56
本文模型	**99.21**	98.33	**98.77**	96.51	**96.94**	**96.72**	95.99	96.42	**96.20**

Table 7: 各类句子自动识别的实验结果

容易发现，在相同模型的实验条件下，有标记被动句的识别性能更高，这是因为有标记被动句本身就存在特征明显的被动标记词，因此通用的文本分类模型就已经能取得较好的识别性能，本文模型通过融合外部特征后，F1值稍有提高，达到98.77%。对于无标记被动句，由于其缺乏明显的标记词，而通过引入词性信息和论元框架信息可以很好地丰富句子的特征表示，因此本文模型与另外三个通用的分类模型相比，在无标记被动句识别任务上取得了最好的性能，其F1值达到96.72%，且在召回率显著提升的情况下，同时保证准确率维持在较高的水平。同样地，对于非被动句而言，由于模型能够较好地学习到被动句的结构和语义特征，那么对于非被动句的特征也应当能较好地捕获，因此本文模型在非被动句识别任务上也获得了较好的性能。

5.3 对比实验

语言技术平台(Language Technology Platform, LTP)是哈工大社会计算与信息检索研究中心(HIT-SCIR) 研发的一套中文自然语言处理开源基础技术平台，其最新发布的LTP 4.0在词法分析、句法分析和语义分析等六项自然语言处理任务上都取得了SOTA或具有竞争力的表现(Che et al., 2021)。其中LTP 4.0(Base2)[5]是LTP 4.0系列各版本模型中，在语义角色标注和句法依存分析任务上都表现最佳的模型，因此本文利用该模型的语义角色标注和句法依存分析这两项功能，对测试集中的句子进行自动解析。由于LTP得到的句子解析结果不能直接判定被动句，我们统计了各类句子的正确解析数目和正确率Acc[6]，并与本文模型的实验结果进行对比。对比实验所用的测试集包含899条有标记被动句、914条无标记被动句和893条非被动句。对比实验结果如表 8 所示。

模型	有标记被动句		无标记被动句		非被动句	
	正确数(条)	Acc(%)	正确数(条)	Acc(%)	正确数(条)	Acc(%)
LTP语义角色标注	487	54.17	814	89.06	850	95.18
LTP句法依存分析	840	93.44	693	75.90	808	90.48
本文模型	**884**	**98.33**	**886**	**96.94**	**861**	**96.42**

Table 8: 本文模型与LTP的对比实验结果

可以看出：首先，在有标记被动句识别方面，LTP的解析性能并不高，尤其语义角色标注功能的解析性能较差，正确率仅有54.17%，而本文模型的性能远远超过LTP，正确率达到98.33%；其次，对于无标记被动句而言，LTP的解析性能也不高，其中句法依存的解析性能只有75.90%，这也说明无标记被动句识别任务是提升句法解析性能所必须克服的一个重要任务，而本文模型效果远远优于LTP，正确率达到96.94%。最后，本文模型在非被动句识别任务上的性能也明显优于LTP自动解析器，正确率达到96.42%。

5.4 消融实验

本文提出的模型融合了词性信息和CPB中的动词论元框架信息，在被动句自动识别任务上取得了良好的性能。为探究二者对模型性能的影响，本文进行了消融实验，实验结果如表 9 所示，其中Ours代表本文模型，Ours-cpb和Ours-pos分别表示在本文模型基础上移除论元框架信息和词性信息的模型。

[5] https://github.com/HIT-SCIR/ltp
[6] 判断标准参照2.3节

模型	有标记被动句识别			无标记被动句识别			非被动句识别		
	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)
Ours	99.21	98.33	98.77	**96.51**	**96.94**	**96.72**	**95.99**	**96.42**	**96.20**
Ours-cpb	**99.22**	**98.78**	**99.00**	95.45	96.50	95.97	95.94	95.30	95.62
Ours-pos	98.99	98.33	98.66	95.96	96.17	96.07	95.54	95.97	95.75

Table 9: 被动句识别消融实验结果

可见在模型移除论元框架信息之后，无标记被动句和非被动句的识别性能均有下降，但有标记被动句识别的F1值反而提升到99.00%，这是由于有标记被动句本身已经具有明显的标记词特征，在引入论元框架信息后，过多的特征表示反而可能会对其造成干扰甚至产生负作用。反观无标记被动句，由于缺少标记词，自动识别的难度较大。在移除论元信息后，无标记被动句识别的F1值降低了0.75%，P值降低了1.06%，这说明论元信息能很好地补充无标记被动句中的特征表示。此外，在移除词性信息后，三类句子的识别性能均有所下降，这说词性信息有利于句子的特征提取和识别。

5.5 词性因子对实验的影响

α	β	γ	δ	Weighted F1
0.25	0.25	0.25	0.25	96.34
0.4	0.3	0.3	0	96.93
0.4	0.4	0.2	0	96.71
0.4	0.3	0.2	0.1	96.86
0.3	0.4	0.2	0.1	96.82
0.3	**0.3**	**0.3**	**0.1**	**96.97**
0.3	0.3	0.1	0.3	96.60
0.3	0.1	0.3	0.3	96.16
0.1	0.3	0.3	0.3	96.32
0.3	0.3	0.2	0.2	96.49

Table 10: 不同词性因子取值下本文模型的性能

为探究词性因子的不同取值对实验的影响，本文通过调整各词性因子的取值进行多次实验，取模型性能最佳时对应的参数值为词性因子的最优取值。参数调整的过程包含三个步骤：①各因子的初始值设为0.25，将其结果作为对照组；②在保证四个词性因子的和为1的条件下，调整或交换其中两个词性因子的值，进行多次实验对比；③如果模型性能有提升，则更新对照组为当前最佳性能对应的参数值，重复步骤②和③。表 10 展示了部分参数设定情况下对应的实验结果，其中Weighted F1为三类句子识别F1的加权平均值。结果表明当四个词性因子分别为0.3、0.3、0.3和0.1 时，模型性能最优，这是因为动词是句子结构的核心，且被动句中的施受事一般是名词或代词，有标记被动句中的标记词为介词，三者对被动句识别都非常重要，因此所占权重较大，而其他词性的词语通常只作为修饰成分，对被动句识别的影响不大，因此权重较小。

6 结语

本文首先手工标注了一个被动句语料库，涵盖三种类型的句子——有标记被动句、无标记被动句和非被动句，然后根据被动句的句式特点，将被动句识别任务建模为一个三分类任务，进而提出了一个融合词性信息和动词论元框架信息的PC-BERT-CNN模型，实现了汉语被动句的自动识别。实验结果表明该模型取得了较好的识别效果，且比现有主流自动解析器能更加准确地识别被动句，有标记被动句和无标记被动句识别的F1值分别达到98.77%和96.72%。

我们将在后续工作中，一方面继续扩大语料规模并进行细粒度标注，为更深入的被动句研究提供支持；另一方面，尝试对细粒度被动句语料进行解析，进一步提升被动句解析的性能。

第二十一届中国计算语言学大会论文集，第384页-第394页，南昌，中国，2022年10月14日至16日。

参考文献

Che W, Feng Y, Qin L, and Liu T. 2021. *N-LTP: An Open-source Neural Language Technology Platform for Chinese. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 42-49.

Che W, Li Z, and Liu T. 2010. *Ltp: A chinese language technology platform. Proceedings of the 23rd international conference on computational linguistics*, 13-16.

Devlin J, Chang M W, Lee K, and Toutanova K. 2019. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the North American Chapter of the Association for Computational Linguistics*, 4171-4186.

Kim Y. 2014. *Convolutional Neural Networks for Sentence Classification. Proceedings of the Association for Computational Linguistics*, 1746-1751.

Nguyen C, Tran V, and Le Nguyen M. 2021. *Enrichment of Features for Malware-Related Sentence Classification using External Knowledge. Proceedings of the 33rd IEEE International Conference on Tools with Artificial Intelligence*, 1144-1148.

Qin Q, Hu W, and Liu B. 2020. *Feature projection for improved text classification. Proceedings of the Association for Computational Linguistics*, 8161-8171.

Xue N, and Palmer M. 2005. *Automatic semantic role labeling for Chinese verbs. Proceedings of the 19th International Joint Conference on Artificial Intelligence*, 1160-1165.

蒋坚松. 2002. 英汉对比与汉译英研究. 湖南人民出版社.

鞠彩萍. 2007. “遭”字句——兼论被动标记词的界定与优胜劣汰. 贵州大学学报（社会科学版）, 2007(01):117-121.

李斌, 闻媛, 宋丽, 卜丽君, 曲维光, 薛念文. 2017. 融合概念对齐信息的中文*AMR*语料库的构建. 中文信息学报, 31(06):93-102.

李珊. 1994. 现代汉语被字句研究. 北京大学出版社.

乔莎莎. 2015. 有标记被动句研究. 黑龙江大学.

宋文辉, 罗政静, 于景超. 2007. 现代汉语被动句施事隐现的计量分析. 中国语文, 2007(02):113-124.

汤敬安. 2016. 汉语无标记被动句与有标记被动句的认知辨析. 云梦学刊, 37(06):110-114.

王灿龙. 1998. 无标记被动句和动词的类. 汉语学习, 1998(05):15-19.

王芸华. 2014. 被动句主语的语义角色考察. 贺州学院学报, 30(02):18-22.

朱向其,张忠林,李林川,马海云. 2021. 基于改进词性信息和*ACBiLSTM*的短文本分类. 计算机应用与软件, 38(12):179-186.

邹丽玲. 2016. 英译汉视角下解析汉语无标记被动句的句法结构. 外语学界, 2016(00):272-281.