汉语被动结构解析及其在CAMR中的应用研究

胡康[1] ， 曲维光[1,2,4*] ， 魏庭新[3] ， 周俊生[1] ， 李斌[2] ， 顾彦慧[1]
(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 南京 210023;
2.南京师范大学 文学院，江苏 南京 210097;
3.南京师范大学 国际文化教育学院，江苏 南京 210097;
4.南京师范大学 中北学院，江苏 丹阳 212334;
*通讯作者，Email：wgqu_nj@163.com)

摘要

汉语被动句是一种重要的语言现象。本文采用BIO结合索引的标注方法，对被动句中的被动结构进行了细粒度标注，提出了一种基于BERT-wwm-ext预训练模型和双仿射注意力机制的CRF序列标注模型，实现对汉语被动句中内部结构的自动解析，F1值达到97.31%。本文提出的模型具有良好的泛化性，实验证明，利用本文模型的被动结构解析结果对CAMR图后处理，能有效提高CAMR被动句解析任务的性能。

关键词： 被动结构解析 ； 双仿射注意力 ； CRF ； CAMR ； 后处理

Parsing of Passive Structure in Chinese and Its Application in CAMR

HU Kang[1], QU Weiguang[1,2,4*], WEI Tingxin[3], ZHOU Junsheng[1], LI Bin[2], GU Yanhui[1]
(1.School of Computer and Electronic Information/School of Artificial Intelligence,
Nanjing Normal University, Nanjing, Jiangsu 210023, China;
2.School of Chinese Language and Literature, Nanjing Normal University, Nanjing, Jiangsu 210097, China;
3.International College for Chinese Studies, Nanjing Normal University, Nanjing, Jiangsu 210097, China;
4.ZhongbeiColleg, Nanjing Normal University, Danyang, Jiangsu 212334, China;
*Corresponding, Email: wgqu_nj@163.com)

Abstract

Chinese passive sentences is an important linguistic phenomenon. In this paper, we use the BIO combined with indexing annotation method to annotate the passive structures in passive sentences at a fine-grained level. We propose a CRF sequence labeling model for passive structure parsing in Chinese based on the BERT-wwm-ext pre-training model and the biaffine attention mechanism, and the model achieves a significant F1 value of 97.31%. The proposed model exhibits excellent generalization capabilities. The experimental results have demonstrated that incorporating the parsing results of passive structures obtained from our model for post-processing the CAMR graph, can effectively improve the performance of passive sentence parsing in CAMR.

Keywords: passive structure parsing , biaffine attention , CRF , CAMR , post-processing

1 引言

被动句是一种常见的语法现象，它强调动作的承受者，将动作执行者放在句子中的其他位置或省略不表达，而一般的主动句则强调动作执行者。被动句的使用可以改变句子的重心，

基金项目：国家社会科学基金重大项目(21&ZD288); 国家自然科学基金(62277031)。

使得动作的承受者成为句子的核心，更加突出其在事件中的地位。被动句的使用十分广泛，表 1展示了被动句在不同语体中的出现频次(宋文辉等, 2007)。人们对被动句的广泛运用丰富了语言的多样性，但被动句构式的复杂多样性也给句法分析、语义解析等任务带来较大困难。

例1 这个问题已经被他解决了。
例2 检查情况已在当地曝光。
例3 这些建议虽然未被采纳，但其价值不可低估。

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

被动句自动解析任务的难点主要在于被动结构的复杂多样。被动结构即被动句中表达被动语态的内部结构，符号表示为A1+[M]+[A0]+V，其中A1表示广义受事，A0表示广义施事，M表示有标记被动结构中的标记词，V表示被动行为的谓语动词，[]表示可省略的成分。一个句子可能包含一个或多个被动结构，如例1中存在一个有标记被动结构“问题$_{A1}$+被$_{M}$+他$_{A0}$+解决$_{V}$”，例2存在一个无标记被动结构“检查情况$_{A1}$+曝光$_{V}$”，而例3存在两种被动结构，分别是“建议$_{A1}$+被$_{M}$+采纳$_{V}$”和”价值$_{A1}$+低估$_{V}$”。因此，能准确解析出句子中的被动结构可以为自然语言处理技术的发展提供技术支持，尤其是机器翻译、自动问答和自动摘要等领域。

本文把被动结构成分识别视为一种特殊的语义角色标注（Semantic Role Labeling, SRL）任务。首先对被动句进行细粒度标注，构建了一个用于模型训练的数据集。然后提出了一个针对汉语被动结构的自动解析模型PS-CRF(Passive Sentence CRF)，F1值达到97.31%。该模型利用融合整词掩码技术的预训练模型BERT-wwm-ext获取被动结构上下文语义信息，使用依存分析方法结合双仿射注意力(Biaffine Attention)评分机制进行特征学习，并通过TreeCRF模型预测输出。最后，为验证本文模型的泛化性和实用性，利用该模型对小学语文语料中被动句的解析结果对CAMR图进行后处理操作，实验证明在Smatch和Align-Smatch两个性能指标下，性能均有提升，尤其在Align-Smatch评价指标下的提升更为显著。

2 相关工作

被动句研究一直是语言学领域的重要课题，学者们除了对句子层面的句法、语义研究，还对被动句内具体成分尤其是谓语动词和标记词进行了研究。赵元任 (1979)提出能用于”被”字句的主要是处置动词，且动词必须前带或后带成分。兰宾汉 (2002)提出被字句的谓语中心语必须是表动作行为的及物动词、不及物动词、能愿动词、趋向动词、判断动词，而表示肯定或否定的”有、没有”等不能用来构成被字句。王振来 (2004)分析了自主动词比非自主动词更容易进入被动表述的原因，阐述被动表述式对动词选择具有制约作用。蚁坤 (2000)通过考察1000个被动句，验证了最常用于被动句中的动词是及物动词，不仅具有及物性特征，还具有高级及物性，其中高级及物性(王惠, 1997)指的是谓语动词具有[动作]、[完成]、[瞬时]、[自主]和[肯定]等特征。王一平 (1994)提出当”遭””挨””受”等遭受类动词后面紧跟一个及物动词，这时遭受类动词可以用”被”来替换，即在某些情况下，遭受类动词可以视为一种特殊的被动标记词。

在自然语言处理领域，有研究针对被动句的句式特点，提出了一种融合词性信息和动词论元框架信息的被动句自动识别模型(Hu et al., 2022)，可以实现从大规模语料中快速筛选被动句。但该模型只能从句子层面判断一个汉语句子是否是被动句，而无法对句中具体的成分进行解析。本文旨在实现对被动句内部结构成分的识别，并且把这个任务视为一种特殊的SRL任务，SRL是自然语言处理中的一项重要任务，其目的是识别一个句子中的每个单词或短语在句子中所扮演的语义角色，例如主语、宾语、谓语、时间状语等，SRL相关的前沿技术和方法对本文模型设计有重要参考意义。Li等人 (2021)分析了句法信息对基于序列、树和图的三种SRL基线模型的影响，提出句法信息能在一定程度上有助于模型学习，但这种帮助随着预训练模型的引入而受到限制，且句法信息的作用大小取决于模型集成句法信息的方式。Zhang等人 (2021)分析了词嵌入方法和不同的标注方法对SRL模型性能的影响，验证了预训练模型带来的性能提升优于静态词嵌入模型。Li等人 (2019)提出了一种可以同时解决span-based和dependency-based的端到端的SRL模型，该模型引入了双仿射注意力机制，能够对SRL两种表示方式进行统一有效的处理，有助于探索二者之间的联系。Zhang等

人 (2022)提出一种用依存句法分析解决SRL任务的方法，该方法的主要思想是：先通过一定的规则将SRL 结构转化为依存句法树，然后基于给定的依存句法树学习一个解析器，最后通过CRF模型将预测出的依存句法树恢复为SRL结构。

3 数据集构建

本文数据集的基础语料来源于被动句语料库(Hu et al., 2022)，该语料库包含4495条有标记被动句、4570条无标记被动句以及4465条非被动句。本文将在此基础上，针对该语料库中的被动句展开细粒度的标注，即对被动句中的被动结构进行成分标注。

3.1 数据集标注方法

被动结构可形式化表示为A1+[M]+[A0]+V，实现被动结构成分的识别，即定位句中包含的所有被动结构，并提取出每个被动结构的具体成分，可以用一个四元组(V, A1, A0, M)表示。该任务与语义角色标注任务类似，都是识别出句中谓语动词及其相关的论元角色，因此本文借鉴SRL任务数据集的标注方法对被动结构进行标注。首先对语料进行分词处理，然后使用词语级别的BIO序列标注方法结合索引的方式对语料中的汉语被动结构进行标注。图 1给出了一个被动结构标注示例，其中“B”表示被动结构中某成分的起始边界词语，“I”表示该成分的后续词语，“-”用于连接成分所属类别，“O”表示非被动结构的其他成分，“：”前的数字表示该成分指向的所属被动结构中动词在句中的索引，若当前成分是动词V，为便于在模型训练过程中对数据进行处理，冒号前的数字用不含实际意义的数字“0”代替。

索引	1	2	3	4	5	**6**	7	8
词语	这个	问题	已经	被	他	解决	了	。
标注	**6**:B-A1	**6**:I-A1	O	**6**:B-M	**6**:B-A0	0:V	O	O

Figure 1: 被动结构标注示例

3.2 数据集统计分析

本文从被动句语料库中随机抽取了3839个被动句进行细粒度标注，共标注被动结构4814个，平均每个句子含1.25个被动结构。数据集中的被动结构主要有以下四种类型：

一、普通有标记被动结构：A1+M+[A0]+V。其中A1指的是广义受事，包括受事、与事、感事、主事、材料、工具等论元角色，标记词M可由”被、由、为、给、让”等介词充当，也可是”受到、遭到”等遭受类动词。如例4中的标记词”被”是介词，而例5中”遭到”则是遭受类动词作为标记词。此外，二者省略了广义施事A0。

例4 裁判员$_{A1}$ 被$_{M}$ 袭击$_{V}$。

例5 裁判员$_{A1}$ 遭到$_{M}$ 袭击$_{V}$。

二、特殊有标记被动结构：M+[A0]+V+的+A1。当一个有标记被动结构作为一个定中结构时，由于不是句子主要成分，与普通有标记被动结构存在一定的差异，我们将之标注为特殊有标记被动结构。如例6中”全乡被洪水吞没的土地”是一个定中结构，其中隐含了一个被动表述”土地+被+洪水+吞没”。

例6 仅一冬一春，全乡被$_{M}$ 洪水$_{A0}$ 吞没$_{V}$ 的土地$_{A1}$ 全部修复。

三、普通无标记被动结构：A1+V。无标记被动结构与有标记被动结构的区别主要在于其不含标记词M和施事A0，前者直接强调了动作的完成，而后者因为标记词的存在更加强调动作的被动性，但二者均表示被动语态，如例7和例8所示。

例7 只有精通国家战略，君主的愿望$_{A1}$ 才可能实现$_{V}$。

例8 一座6000多平方米的晒谷场$_{A1}$ 也已建成$_{V}$。

四、特殊无标记被动结构：V+的+A1。与特殊有标记被动结构类似，也有少数无标记被动结构会出现在定中结构中，如例9所示。

例9 本次集中行动中列为$_{V}$ 全国重点的6大案件$_{A1}$ ，现已审结4件。

表 2列出了数据集中各种被动结构的数目以及占比，其中有标记的被动结构共2711个，无标记的被动结构共2103个。在有标记被动结构中，普通有标记被动结构有2655个，其中介词作

标记词的有2465个，遭受类动词作标记词的有190个。定中结构中的被动结构总体来说占比较少，有标记和无标记的被动结构分别占比1.16%和0.27%。

被动结构类别	细分类别	数量（个）	占比（%）	小计（个）
有标记被动结构	介词	2465	51.20	2711
	遭受类动词	190	3.95	
	定中结构	56	1.16	
无标记被动结构	普通无标记	2090	43.42	2103
	定中结构	13	0.27	
		总计	100	4814

Table 2: 被动结构数据集

4 模型设计

本文将被动结构成分识别任务建模为一个语义角色标注任务，提出一个使用高效的中文预训练模型并结合双仿射注意力机制的CRF序列标注模型，实现对汉语被动句中内部结构的自动解析，将该模型记为PS-CRF。对一个输入样本句子$S = w_1, w_2, \ldots, w_n$，其中$w_i$表示第$i$个词语，$i \in 1, 2, \ldots, n$，n为句子词语总数。模型输出为对应词语个数的标签序列$Y = o_1, o_2, \ldots, o_n$，其中$o_i$表示第$i$个词语对应的成分预测标签。模型由三大模块组成，分别是基于BERT-wwm-ext的词嵌入模块、基于双仿射注意力机制的评分模块以及基于TreeCRF的模型预测输出模块。模型整体架构如图 2 所示。

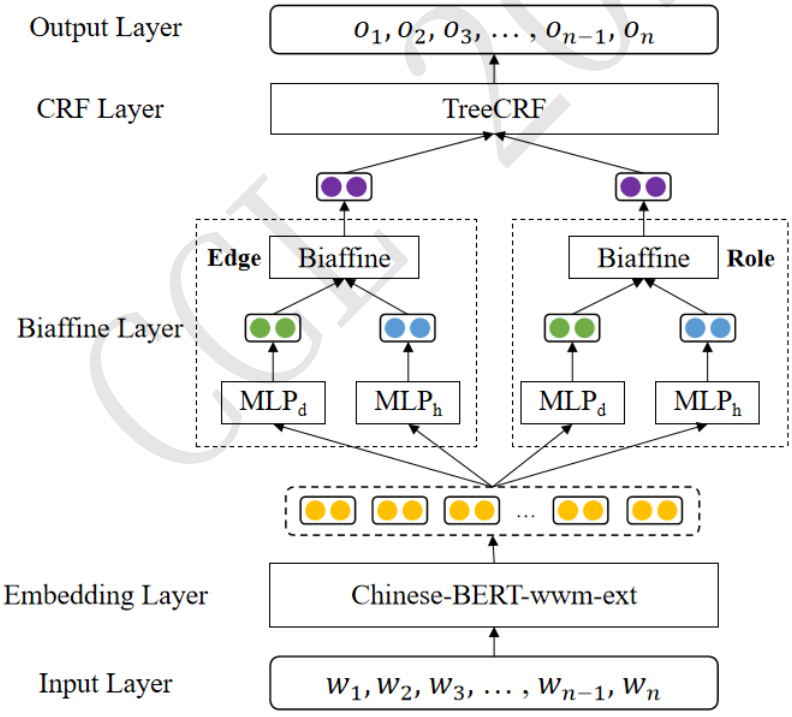

Figure 2: 被动结构成分识别模型图

4.1 基于BERT-wwm-ext的词嵌入层

本研究使用的BERT-wwm-ext预训练语言模型，是基于BERT模型和整词掩码技术(Whole Word Masking, WWM)技术扩展而来的。被动结构的成分大多数是多音节词语，普通BERT模

型在MLM(Masked Language Model)阶段是针对单字进行掩码操作，这容易导致词语被分割，影响句子的语义表示。Cui Y等人 (2021)把整词掩码(Whole Word Masking, WWM)技术引入BERT模型，在训练过程中，更加关注整个词语的语义，从而更好地捕捉语言的上下文信息。为进一步提升中文自然语言处理任务性能，该团队又往BERT模型训练语料增加了大量维基百科、新闻、问答等通用数据，同时对语料进行了严格的数据清洗和预处理，以确保训练得到的模型更加准确和有效，提出了BERT-wwm-ext模型。该模型一方面对中文词语的语义表示性能更好，另一方面丰富了对新闻领域文本的解析能力，因此它相比其他预训练模型更利于被动结构各成分的识别。

4.2 基于Biaffine Attention的评分模块

本文将基于依存分析的SRL方法应用到被动结构成分识别上，把一个句子中的所有被动结构转化为依存树结构，如图3所示，句中的一个被动结构可以转化为一棵多叉树，第一层为树的根节点，第二层为动词，第三层为被动结构除动词外的其他成分。被动结构转化为树结构后，模型的训练目标是从句子中解析出最佳子树，实现这一过程的关键步骤是对依存树进行评分，本文使用双仿射注意力(Bi-affine Attention)来实现这一过程。

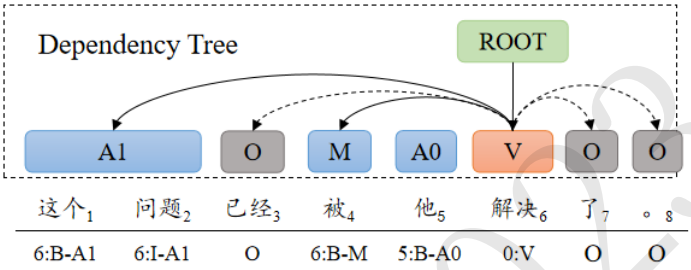

Figure 3: 被动结构的依存树形式

双仿射注意力机制 (Timothy et al., 2017)是一种用于解决依存句法分析问题的神经网络结构，具体来说，双仿射机制首先将每个词的词向量作为输入，经过一些线性变换和激活函数处理之后，得到一个隐向量表示。然后，使用双仿射函数来计算每对词之间的相关性得分，这个得分可以表示两个词之间的依存关系强度。对于两个词i和j，双仿射函数可以表示为公式(1):

$$Biaffine(i,j) = h_i^T W_1 U W_2^T h_j \tag{1}$$

其中，h_i和h_j分别表示词i和j的隐向量表示，W_1和W_2是可学习的权重矩阵，U是一个对称的得分矩阵，可以表示两个词之间的依存关系强度。

依存树的评分过程由结点预测和标签预测两个子任务构成。结点预测子任务是预测两个结点之间是否存在依存关系，标签预测子任务则是预测两个结点之间存在的被动关系属于何种具体关系。在本任务中有三种关系，第一种是动词与根节点之间的关系，记为ROOT-V；第二种是被动结构除动词外的其他成分与该动词之间的关系，记为V-A1、V-A0和V-M；第三种是对句中非被动结构的成分，记为V-O。

对于输入样本$S = w_1, w_2, \ldots, w_n$，经过预训练模型得到词嵌入$X = x_1, x_2, \ldots, x_n$。对句中任意两个结点$a$和$b$，判断二者之间是否存在依存关系，即依存树中是否存在$a \rightarrow b$的依存弧，先定义两个多层感知机MLP^h和MLP^t分别计算依存弧首尾两个结点的隐向量，然后把两个结点的隐向量代入双仿射函数中计算依存强度得分，如公式(2)-(4)。

$$r_a = MLP^h(x_a) \tag{2}$$

$$r_b = MLP^t(x_b) \tag{3}$$

$$Score(a \rightarrow b) = Biaffine(r_a, r_b) \tag{4}$$

标签预测子任务的算法与结点预测子任务类似，也是通过两个MLP结合一个Biaffine函数计算得分，最终遍历完所有可能的结点对，就可得到一棵依存树的得分。

4.3 基于TreeCRF的预测输出层

TreeCRF模型(McDonald and Pereira, 2006)是一种条件随机场模型，用于处理树形结构的数据。在TreeCRF模型中，每个节点表示一个观测变量，每条边表示一个潜在变量，即节点之间的关系。给定一个树和观测变量序列，TreeCRF模型的概率分布可以表示为一组特征函数的乘积，其中特征函数描述了每条边标注的条件概率。模型的学习是通过训练特征函数的权重来实现的。在预测时，使用维特比算法解码，得到最可能的标注序列。

给定一个树T和观测变量序列$x = (x_1, x_2, \ldots, x_n)$，TreeCRF模型的概率分布可以表示为：

$$P(y|x,T) = \frac{1}{Z(x,T)} \prod_{(i,j)\in E} \psi(y_{i,j}, x_i, x_j) \tag{5}$$

其中，E表示任意两个点之间构成边的集合，$y = (y_{1,2}, y_{1,3}, \ldots, y_{n-1,n})$ 表示给定观测变量x 对应的树T 的边上的潜在变量，$\psi(y_{i,j}, x_i, x_j)$表示边(i,j)的特征函数，$Z(x,T)$是归一化常数，用于保证模型的概率分布性质成立，即

$$P(x,T) = \sum_{y} \prod_{(i,j)\in E} \psi(y_{i,j}, x_i, x_j) \tag{6}$$

通过学习特征函数$\psi(y_{i,j}, x_i, x_j)$的权重，可以得到TreeCRF模型。在预测时，可以使用维特比算法进行解码，得到最可能的边的标注序列$\hat{y}$，即

$$\hat{y} = argmax_y P(y|x,T) \tag{7}$$

5 被动结构成分识别实验

5.1 参数设置及评价指标

本文的实验数据集按照6:2:2划分为训练集、验证集和测试集并随机打乱，实验使用的预训练模型基本参数为L-12_H-768_A-12，具体数据集划分和模型超参数设置如表 3和表 4所示。

数据集	句子数量（条）
训练集	2303
验证集	768
测试集	768

Table 3: 数据集划分

超参数	含义	值
epochs	数据集迭代次数	10
batch_size	单批次样本数量	128
pad_size	每个样本最大token数量	128
learning_rate	学习率	5e-5
dropout	丢弃概率	0.1

Table 4: 超参数设置

本文实验按准确率P、召回率R和F1得分进行评价，公式如(8)-(10)所示。

$$\mathrm{P} = \frac{TP}{TP + FP} \tag{8}$$

$$\mathrm{R} = \frac{TP}{TP + FN} \tag{9}$$

$$F1 = \frac{2 * P * R}{P + R} \tag{10}$$

其中，TP 表示模型正确预测的语义角色的数量，FP 表示模型错误预测的语义角色的数量，FN 表示模型未能正确预测的语义角色的数量。

5.2 实验结果与分析

本文采用CRF模型作为基线模型，在此基础上分别采用多种方法进行模型的构建并进行实验对比。首先对比了在相同预训练模型(GloVe)情况下使用句法依存分析方法（Biaffine +TreeCRF）和传统CRF序列预测的方法解决被动句解析的性能差异；然后对比了在相同序列预测模型下不同预训练模型对性能的影响；最后将这些模型与本文提出的模型，即BERT-wwm-ext+Biaffine+TreeCRF进行实验结果对比和分析。实验结果如表 5所示。 由实验结果

模型	P(%)	R(%)	F1(%)
CRF	78.53	76.68	77.59
Biaffine+TreeCRF	81.34	79.63	80.47
BERT+CRF	94.34	92.63	93.48
BERT+Biaffine+TreeCRF	97.18	95.72	96.44
本文模型	**98.46**	**96.18**	**97.31**

Table 5: 被动结构成分识别实验结果

可知，首先，Biaffine+TreeCRF模型比CRF的性能提高近3个百分点，说明将被动结构建模为一种依存句法树结构更有利于解析。这是由于一个句子中可能含有多个被动结构，且多个被动结构之间又可能存在嵌套，但CRF这种传统的序列标注方法无法应用于嵌套识别任务，而Biaffine+TreeCRF采用句法依存分析的方法，将每个被动结构解析为独立的依存树，不会相互影响，所以性能得到了明显的提升。

其次，相比于静态词向量GloVe，不论是CRF还是Biaffine +TreeCRF模型，在使用动态词向量预训练模型BERT之后，模型的性能都得到了较大的提升。这是因为静态词向量模型是基于全局统计信息，无法很好地处理不同语境中的上下文信息。而BERT采用了双向Transformer结构，可以同时考虑前后文信息，使得生成的词向量更具有上下文的代表性。

最后，相比上述四种模型，本文提出的模型取得了最好的解析性能，其F1值达到了97.31%。一方面是由于BERT-wwm-ext预训练模型采用了整词掩码技术，在MASK操作时能更好地学习到被动结构中每个词语完整的语义，从而提高模型的泛化能力和语义表示能力。另一方面该预训练模型是专门针对中文训练的模型，在预训练过程中学习了额外的新闻领域的文本知识，而本文数据基础也来源于新闻语料，因此能更好地捕捉中文句子的语义特征，进而提升了被动结构成分识别的性能。

通过对测试集中预测错误的被动结构进行分析，发现错误主要有两种情况：一是定中结构中的被动结构识别较差。如“.....群众互助互济活动的广泛开展。”中的“开展”一词是具有制作义的自主动词，能进入被动语态，但由于数据集中涉及到状中结构的句子占比只有不到2%，因此模型在训练时难以学习这种特殊结构的语义和句法特征。二是含多种义项的词语被错误识别为被动结构中的动词。例如“车辆乱停放等问题十分突出。”中的“突出”是动词兼形容词，而在该句中是作形容词，造成这一问题可能是由于句中的主语是“问题”，而这个词语在数据集中充当被动结构中的A1频次较高，使得模型对这个词语较为敏感，进而导致模型识别错误。

6 CAMR后处理实验

AMR是一种领域无关的句子语义表示方法，它将一个句子的语义抽象为一个单根有向无环图，其中句子中的实词抽象为概念节点，实词之间的关系抽象为带有语义关系标签的有向弧(曲维光等, 2017)。中文AMR也称作CAMR，它在AMR的基础上对汉语中常见的和特殊的语言现象作了细致的定义。但CAMR现有的自动解析模型对被动句的解析还存在一定的不足。本节实验通过利用被动结构成分的解析结果对CAMR解析图进行后处理，以期提升CAMR被动句解析任务的性能。

6.1 SPRING模型对被动句的自动解析

SPRING模型是BevilacquaM等人(2017)提出的一种AMR自动解析架构，此架构可以完成文本到AMR的解析和AMR到文本的生成两种任务，即利用BART的迁移学习能力完成这两个任务。本文首先利用被动句自动识别模型(Hu et al., 2022)从CAMR小学语文语料中识别被动句，经过自动识别和人工校对，筛选出有标记和无标记被动句各80条，共计160条被动句，对每个句子中的被动结构进行人工标注。然后利用SPRING模型中的Text-to-AMR任务框架进行AMR自动解析，考察了SPRING模型对被动句的解析性能。

Figure 4: CAMR黄金标注与自动解析示例

图4是句子“一天早晨，巨人被喧闹声吵醒了。”的黄金标注和自动解析图，句中包含一个被动结构“巨人$_{A1}$+被$_{M}$+喧闹声$_{A0}$+吵醒$_{V}$”。观察图(a)，句子的谓语动词解析为“x7_吵醒”，其中“x7”是一个概念对齐信息，它表示“吵醒”一词是句子中的第7个词语。同理，受事是“x4_巨人”，作为动词的arg1子结点；施事是“x6_喧闹声”，作为动词的arg0子结点。标记词“被”作为关系对齐包含在关系有向弧arg0中。图(c)是SPRING自动解析图，一方面，该CAMR图中概念节点“巨人”和“喧闹声”对应的语义角色标签解析错误了，正确的SPRING解析图应如(b)所示；另一方面，SPRING模型的解析结果不含概念对齐和关系对齐信息，即缺乏概念节点与句中词语间的索引信息和语义关系有向弧上的虚词信息，导致CAMR解析图中丢失了有标记被动结构中十分重要的标记词信息。

因此，为更加全面地了解SPRING模型对被动句的解析情况，本文基于CAMR图的结构设计了一组被动结构解析正确与否的判定规则，用于实现该模型对被动句解析正确率的统计分析。具体规则如下：

（1）**动词是否正确解析。**在以该动词为中心的被动关系对应的CAMR图中的某个子树，动词抽象而来的概念节点应当是这个子树的根节点。

（2）**受事主语是否解析为动词的arg1。**在当前子树中，受事对应的结点应当是动词概念结点的孩子节点，且关系标签是arg1。

（3）**施事是否解析为动词的arg0。**与arg1同理，若当前被动关系中出现了施事，则其对应的概念结点也应当是动词概念结点的孩子结点，关系标签是arg0。

（4）**不考虑概念和关系对齐信息。**由于SPRING模型的解析结果不具有概念和关系对齐信息，因此在判定的时候仅关注前三条规则。

利用上述判定规则对SPRING模型的CAMR解析图进行判定，如图4(b)可判定为解析正确，而图4(c)解析错误。同时利用本文提出的PS-CRF模型对160条被动句进行解析，统计两类被动句的自动解析正确率，实验结果如表 6所示。

类别	模型	黄金标注(条)	正确解析(条)	正确率(%)
有标记被动句	SPRING	80	36	45.00
	PS-CRF	80	76	**95.00**
无标记被动句	SPRING	80	42	52.50
	PS-CRF	80	73	**91.75**

Table 6: 两种模型解析被动句的正确率

可以看出，SPRING模型在被动句解析上性能较差，而本文提出的PS-CRF模型取得了良好的性能，两种被动句的解析正确率均达到90%以上。为提升CAMR对被动句的解析性能，本文尝试利用PS-CRF模型的识别结果对CAMR图进行后处理。

6.2 后处理算法设计

针对CAMR对被动句解析存在不足的问题，本节设计了一个CAMR后处理算法，来纠正CAMR图中错误解析的被动关系。CAMR后处理算法分为三个步骤：第一，把AMR图和被动结构成分都转化为多元组形式；其次，补充或修改被动结构中的概念节点和关系；最后，把AMR多元组还原成AMR解析图。算法流程图如图 5所示。

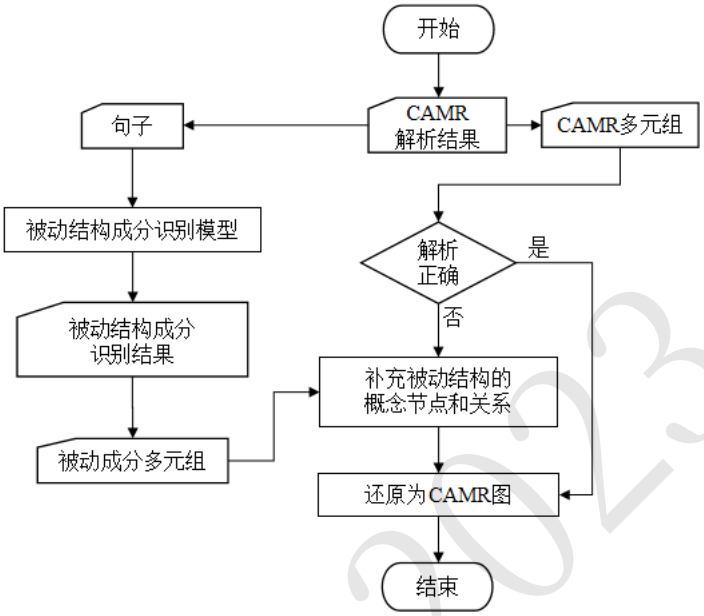

Figure 5: CAMR后处理算法流程图

由于SPRING解析结果不含概念对齐和关系对齐信息，而被动结构中的标记词对应CAMR中的关系对齐，且概念对齐信息有利于从CAMR解析图中定位相关的词语，便于后处理操作。因此本文还设置了两组对照实验，即先通过人工补充SPRING解析图中的概念对齐信息，再进行后处理操作。图 6是句子“许多人被火围困在山顶上。”的SPRING解析图在人工添加概念对齐信息前后的对比图，如(a)和(b)。在此基础上，分别利用被动解析结果对其进行后处理，得到图(c)和(d)。

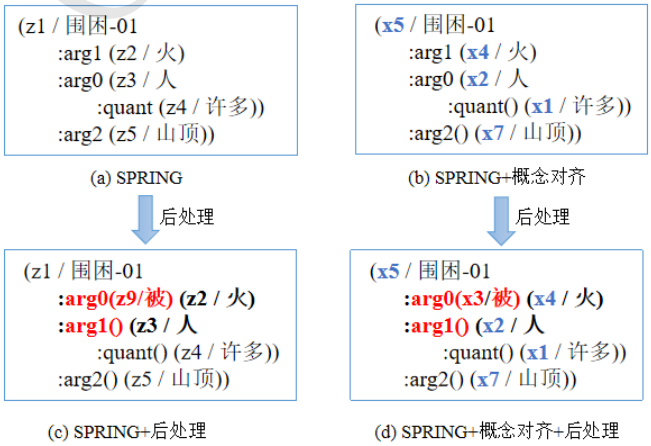

Figure 6: CAMR后处理的不同实验设置

6.3 实验结果与分析

本文采用了Align-Smatch和Smatch两种指标对CAMR解析图进行评价，其中Smatch是基于英文AMR设计的评测指标，Align-Smatch则是一种基于CAMR设计的评价指标，它兼容了中文AMR所特有的概念和关系对齐信息，还可以对有向弧上的虚词进行评估(肖力铭等, 2022)。实验结果如表 7所示。

模型	Align-Smatch(%)	Smatch(%)
SPRING	45.2	69.8
SPRING+后处理	47.6↑	70.7↑
SPRING+人工对齐	62.3	—
SPRING+人工对齐+后处理	65.8↑	—

Table 7: CAMR后处理实验结果

由实验结果可知，在不添加概念对齐信息时，CAMR解析图在利用被动结构成分进行后处理之后，两种评价指标得分均有所提高，其中Align-Smatch提升较大而Smatch值提升略低。这是因为Smatch指标是针对英文设计的，它在对两个AMR图进行匹配评分时，只关注概念节点和边的标签，且标签不含附加成分，实验使用的后处理数据是被动结构，其中施受事的修改、补充是概念节点层面的后处理，而有标记被动结构中的标记词作为论元关系标签中的附加成分，不作为独立的概念节点，因此对于Smatch评价指标而言，后处理带来的性能提升仅仅是由于动词及其施受事成分而不包含标记词，而Align-Smatch把所有的被动结构成分都利用了，所以提升效果明显。此外，后处理之前的Align-Smatch值比Smatch值低，这是由于Align-Smatch在计算得分的时候，不仅仅关注概念节点的匹配程度，更重要的是概念对齐信息和关系对齐信息，而现有的CARM解析器包括SPRING，生成的解析图都不包含这两种对齐信息，所以导致Align-Smatch得分较低。

而在对SPRING解析图人工添加概念对齐信息后，Align-Smatch得分由45.2提升到62.3，这验证了对齐信息对于Align-Smatch指标的重要性，在此基础上再利用被动结构成分进行后处理，分值达到了65.8，提升了3.5个百分点。而在人工添加概念对齐信息之前，后处理操作带来的性能提升为2.4个百分点，由此可见本文提出的被动结构成分识别模型对中文AMR的解析性能有一定的提升效果，尤其是对于包含对齐信息的CAMR图。

7 结语

本文把被动结构成分识别任务建模为一种语义角色标注任务。首先对被动句中的具体结构成分进行了细粒度标注；然后提出了一种BERT-wwm-ext预训练模型结合双仿射注意力机制的CRF序列标注模型，该模型取得了较好的解析性能，F1值达到了97.31%；最后基于CAMR小学语文语料，将本文模型应用到CAMR解析后处理任务中，提升了CAMR对被动结构的解析性能。

在后续工作当中，一方面我们将进一步完善标注规范，尤其是针对特殊被动结构和动词的标注，提升数据标注的一致性、平衡性。另一方面，考虑对被动句自动解析模型进一步优化，尝试融入更多语言学知识，以增强模型的可解释性。

参考文献

Bevilacqua M, Blloshmi R, and Navigli R. 2021. *One SPRING to rule them both: Symmetric AMR semantic parsing and generation without a complex pipeline. Proceedings of the AAAI*, 12564-12573.

Cui Y, Che W, Liu T, Qin B, and Yang Z. 2021. *Pre-training with whole word masking for chinese bert. IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 3504-3514.

Hu K, Qu W, Wei T, Zhou J, Gu Y, and Li B. 2022. *Automatic Recognition of Chinese Passive Sentences Based on Feature Fusion. Proceedings of the CCL*, 384-394.

Li Z, Zhao H, He S, Zhang Y, Zhang Z, Zhou X, and Zhou X. 2019. *Dependency or span, end-to-end uniform semantic role labeling. Proceedings of the AAAI*, 33(01): 6730-6737.

Li Z, Zhao H, He S, and Cai J. 2021. *Syntax role for neural semantic role labeling. Computational Linguistics*, 47(3): 529–574.

McDonald R, and Pereira F. 2006. *Online learning of approximate dependency parsing algorithms. Proceedings of the EACL*, 81-88.

Timothy D, and Christopher D. M. 2017. *Deep biaffine attention for neural dependency parsing. Proceedings of ICLR*, 2017.

Zhang Y, Xia Q, Zhou S, Jiang Y, Li Z, Fu G, and Zhang M. 2022. *Semantic Role Labeling as Dependency Parsing: Exploring Latent Tree Structures inside Arguments. Proceedings of the COLING*, 4212-4227.

Zhang Z, Emma S, and Eduard H. 2021. *Comparing span extraction methods for semantic role labeling. Proceedings of the SPNLP*, 67-77.

兰宾汉. 2002. 汉语语法知识与应用. 北京：石油工业出版社, 2002:123.

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.

马庆株. 1988. 自主动词与非自主动词. 中国语言学报, 1998(03):157-180.

曲维光, 周俊生, 吴晓东, 戴茹冰, 顾敏, 顾彦慧. 2017. 自然语言句子抽象语义表示*AMR*研究综述. 数据采集与处理, 32(01):26-36.

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.

王惠. 1997. 从及物性系统看现代汉语的句式. 语言学论丛, 1997:19.

王一平. 1994. 从遭受类动词所带宾语的情况看遭受类动词的特点. 语文研究, 1994(04):28-34.

王振来. 2004. 被动表述对自主动词和非自主动词的选择. 汉语学习, 2004(06):17-22.

肖力铭, 李斌, 许智星, 霍凯蕊, 冯敏萱, 周俊生, 曲维光. 2022. 基于概念关系对齐的中文抽象语义表示解析评测方法. 中文信息学报, 36(1): 21-30.

蚁坤. 2000. 汉语被动句的句法语义特征和使用条件. 北京语言文化大学.

赵元任. 1979. 汉语口语语法. 北京：商务印书馆, 1979:134-176.