中文词语离合现象识别研究

周露[1] ， **曲维光**[1,2,*] ， **魏庭新**[2,3] ， **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，江苏省 南京市 210023;
2.南京师范大学 文学院，江苏省 南京市 210097;
3.南京师范大学 国际文化教育学院，江苏省 南京市 210097;
*通讯作者，Email：wgqu_nj@163.com)

摘要

汉语词语的离合现象是汉语中一种词语可分可合的特殊现象。本文采用字符级序列标注方法解决二字动词离合现象的自动识别问题，以避免中文分词及词性标注的错误传递，节省制定匹配规则与特征模板的人工开支。在训练过程中微调BERT中文预训练模型，获取面向目标任务的字符向量表示，并引入掩码机制对模型隐藏离用法中分离的词语，减轻词语本身对识别结果的影响，强化中间插入成分的学习，并对前后语素采用不同的掩码以强调其出现顺序，进而使模型具备了识别复杂及偶发性离用法的能力。为获得含有上下文信息的句子表达，将原始的句子表达与采用掩码的句子表达分别输入两个不同参数的BiLSTM层进行训练，最后采用CRF算法捕捉句子标签序列的依赖关系。本文提出的BERT_MASK + 2BiLSTMs + CRF模型比现有最优的离合词识别模型提高了2.85%的F1 值。

关键词： 离合词 ； 自动识别 ； 掩码机制 ； 神经网络

Research on Recognition of the Separation and Reunion Phenomena of Words in Chinese

ZHOU Lu[1], **QU Weiguang**[1,2], **WEI Tingxin**[2,3], **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;
*Corresponding, Email: wgqu_nj@163.com)

Abstract

The Separation of words where characters of a word can be used separately is the special phenomena in Chinese. This paper proposes a character-level sequence tagging method for automatic recognition of separable two-character verbs, which avoids the error propagation of Chinese word segmentation and part of speech tagging, and saves the labor cost of making rules and feature templates. In the process of training, the Chinese pre-trained model BERT is fine-tuned to obtain the vector representations of characters for the target task. Meanwhile, the MASK mechanism is introduced, where the separable words within the separable structures is hidden from the model, to reduce the influence of words themselves on the recognition results and strengthen the learning of the middle insertions in separable structures, and further, different masks for the

基金项目：国家自然科学基金“汉语抽象意义表示关键技术研究”(61772278)；国家自然科学基金“向量组合学习框架下基于依存混合树的中文语义解析研究”(61472191)；江苏省高校哲学社会科学基金“面向机器学习的汉语复句语料建设研究”(2019JSA0220)；国家社会科学基金“中文抽象语义库的构建及自动分析研究”（18BYY127）。

front and back morphemes are used to emphasize their occurrence complicated and occurrence order. Therefore, such implements give the model the ability to recognize complicated and occasional separation forms. To obtain the sentence representation with context information, the original sentence representation and the mask sentence representation are input into two BiLSTM modules with two different parameters for training, respectively. Finally, the CRF algorithm is used to capture the dependency of the tag sequence of the sentence of the sentence. The proposed BERT_MASK + 2BiLSTMs + CRF model improves the F1 value by 2.85% compared with the existing optimal separable word recognition model.

Keywords: Separable Words , Automatic Recognition , MASK Mechanism , Neural Network

1 引言

词语的离合现象是汉语中的一种特殊现象，指词语的前后语素之间可以插入其他成分而被分离，但在分离后所表达的意思仍然是一个整体，词语既有“合用法”，又有“离用法”。这类语法结构尤以二字动词居多，因此本文着重研究二字动词离合现象的自动识别，解决自然文本中词语离用法的识别任务。例如在句子“他帮了我一个忙”中存在词语“帮忙”两个字分开使用的离用法“帮了我一个忙”。

词语的离合现象普遍出现在汉语语料，尤其是日常的口语化表达中。根据语法的“浮现观”，语法是动态的，在交际中使用的语法的结构单位在不停地变化。对词语离合现象这类语法结构，传统研究中主要有对离合词的研究。离合词在组成上是由双音节语素构成的，在使用中呈现可离可合的语法特点，在表现形式上符合本文任务中对词语离合现象的识别情况，因此本文的工作主要借鉴汉语离合词的识别。王海峰 (2010) 发现离合词越是表现日常行为，其离析频率越高，离析种类越丰富。通过对离合词现象的考察和分析，王海峰在2012年 (2012) 提出离析句，即离用法所在句，受元语用意识支配，表现出说话人的主观意图。王蕙淼 (2013)从离合词的成因角度分析，发现离合词口语性很强，人们在生活交际中追求“经济原则”，离用法可以随着语用需求而灵活变化，满足了表达简洁而丰富的追求；加上类化作用将离用法套用到原本不是离合词的词中，这类词语产生的离用法随着使用者增多和时间发展，成为约定俗成的用法，最终进入到了离合词的范畴中。根据这一趋势，我们可以合理认为，在汉语的使用中，尤其是在随意化的日常表达中，词语的离用法现象会越来越多，并在其出现伊始呈现偶发性。

本文提出了基于字粒度的神经网络模型，即BERT_MASK + 2BiLSTMs + CRF模型，节省了大量规则设计及标注工作，避免了标注、分词及词性标注的误差传播；直接面向离用法边界和句中离用法成分位置信息的识别，对中间插入成分不做限制，赋予了模型识别长而复杂的离用法的能力；为减轻模型对高频出现的词语关注程度更高而忽视了对离用法结构的关注，出现过拟合的情况，在离用法识别中设计了掩码(MASK)机制，即对模型隐藏发生分离的两个字，从而强化离用法构成规律、弱化词语本身使用频率，让模型可以从对结构的学习归纳中获得捕捉偶发性词语离用法的能力。对比依存句法分析器的句法分析结果，展示了模型可以达到通过识别词语离用法为依存句法分析提供底层支持的效果，进一步增强句义理解能力，对自然语言处理技术的基础研究具有实际意义。实验结果显示，本文提出的模型在赵聿夕 (2019)构建的数据集上，相比该文提出的最优模型，F1值提高了2.85%，能够有效识别自然文本中二字动词的离用法现象。

2 相关工作

2.1 对离合词的标注

离合词语料的标注规范不统一，语料资源匮乏。周卫华等人 (2010)在中文信息处理系统中对离合词扩展形式做出专门的符号标注。在目前的公开语料库中，北京大学现代汉语研究语料库（CCL语料库） (2002)对双音节离合词做了标注。例如句子“出/v 过/u 两/m 天/q 差/Ng”中，词语“出差”分开使用，因此将“差”标注为语素，使用其子类标记“Ng”。戴茹冰等人 (2020)在中文抽象语义表示（CAMR）标注体系 (李斌等, 2017)中对动宾式离合词的标注

方式做出了组合离合式的规定，即统一将表示前后语素的两个概念节点做合并处理。如例句“警卫员向毛主席敬了个礼。”中存在离合词“敬礼”的离用法“敬了个礼”，在CAMR中被合并处理为一个概念“敬礼”，中间插入成分根据语义在下一层进行标注。正确识别离合词和解析离合词的离用法，可以辅助CAMR的解析，对句子的语义解析及其他下游任务具有重要意义。

2.2 离合词的自动识别

2.2.1 基于规则匹配

以往的研究工作主要着眼于利用大规模语料库对离合词的离用法构成规则进行人工归纳，总结其构成规律并构建离合词词表，从而采用基于规则匹配的离合词自动识别方法。周卫华和胡家全 (2010)细致描写了动宾式和并列式离合词的离用法结构，并分析了这两类离合词离用法构成的特点，提出在中文信息处理系统中建立离合词词库，对离合词离用法做出专门的符号标注。刘博 (2015)基于北京语言大学汉语语言学研究中心语料库，设计了根据词表和规则匹配对现代汉语离合词离用法的自动识别系统，但是该系统无法识别未登录词、中间插入成分多于三个的情况，在使用中需要不断更新离合词离用法结构模板。同样通过对大规模语料，北京语言大学BBC中的综合频道语料中离合词离用法的分析，藏娇娇和荀恩东 (2017)对其做了形式化表示，总结了自动识别规则，设计了基于规则的自动识别算法。但由于插入成分复杂，该方法只选取140个常用的动宾形式离合词进行研究，不在句中标注离合词，仅以“0/1”标注句子中是否存在离合词的离用法。

此类针对离合词离用法总结的规则匹配算法主要依赖人工依据词性制定的特征模板，直接受到分词及词性标注效果的制约，且插入成分限制在四个以内。

2.2.2 基于机器学习和神经网络

针对离合词应保持词义不变的性质，张振景等人 (2016)采用句子的词特征、词性特征、词及词性特征设计特征模板，将句子转换为one-hot特征向量，采用SVM模型建立句子分类器，将词义消歧的方法应用于存在多义的离合词。为保证样本容量，该方法在北京大学的CCL语料库上只选取了5个高频多义离合词进行研究。赵聿夕 (2019)以新华社十四年新闻语料作为原始语料，根据人工制定的离用法规则动态生成离合词词表，并根据词表、基于规则匹配进行离合词自动识别方法研究，对于规则匹配结果错误的句子通过构建特征模板获取句子表示，分别采用机器学习和深度学习方法进行句子二分类任务，最终构建了一个离合词自动识别级联模型。

此类识别模型的效果除了受到分词及词性标注、特征模板只能匹配至多三个中间插入成分的制约外，利用大规模标注数据来进行监督学习，对出现频率较高的词语学习效果更好，而无法识别在训练集中低频、甚至没有出现过的离合词。

为解决这些传统识别模型存在的问题，本文提出了基于字符级序列标注方法的神经网络模型，并且可以在一定程度上避免人为判断步骤和模型构建中管道式的系统错误传播。

3 离合词识别任务研究

3.1 任务定义

本文为避免分词以及词性标注系统造成误差传播的潜在问题，采用字符级序列化标注方法进行词语离用法识别任务，即输入一个句子$S = c_1, c_2, ..., c_n$，其中c_i 表示句中第i个字符，$i \in \{1, 2, ..., n\}$，n为句子长度。模型预测输出为对应句子中每个字符的标签序列$Y = y_1, y_2, ..., y_n$，其中$y_i \in \{B, I, E, O\}$。B标签表示该字符是离用法的起始字符，即词语前语素，I标签表示该字符属于离用法的中间插入成分，E标签表示该字符是离用法的结尾字符，即词语后语素，O标签标注了句子中离用法外的其他成分。我们将存在离用法的句子称为离合句，表 1 中给出“帮忙”所在的一句离合句样本的标注示例。

文本	他	帮	了	我	一	个	忙
标签	O	B	I	I	I	I	E

Table 1: 离合句标注示例

一个正确的离用法应该包括完整的BIE标签，其中I标签至少有1个，对应的二字词语由标

注为B和E标签的语素组成。字符级序列化标注方法对前后语素之间的插入成分的数量没有限制，并且直接面向离用法的识别，主要依靠B、E标签识别离用法边界，进而识别出对应词语。

3.2 数据集构建

本文采用的数据集源自赵聿夕 (2019)工作中构建的离合词语料库，本文中称之为原语料库。原语料库中离合词与其离用法所在离合句为一一对应关系，每一条数据的格式为“离合词 离合句”，如例1所示，无法说明离合句中是否存在对应离合词的多次离用法，而一句离合句只对应一个离合词，显然无法表明一句中有多个不同的离合词离用法的现象，因此这两种情况在原语料中都无法标注出。而采用本文提出的模型可以根据多组BIE标签的识别情况，识别出单句中存在多个离合词及其离用法，并将标注精确到语素在句子中的位置索引，避免了同一字符在句子中出现多次而造成不确定应该选取哪一个组成词语合式的问题。单个样本句的标注形式为“$S|p_1|p_2|.....|p_m|$”，其中p_j表示句中存在的第j个离用法标注，p_j=“B_j E_j”,B_j表示第j个离用法的前语素位置索引，E_j表示第j个离用法的后语素位置索引，$j \in \{0,1,...,m\}$，m为离用法的数量，用“ | ”符号隔开多个离用法标注，每个离用法标注中用空格隔开前后语素索引。例如，在例2中，存在同一个离合词“鼓劲”的两次离用法；在例3中存在两个不同的词语及其离用法，分别是“洗澡　洗完澡”和“下班　下了班”。

例1 帮忙 他帮了我一个忙

例2 鼓实劲而不鼓虚劲| 0　2 | 5　7 |

例3 可等洗完澡下了班| 2　4 | 5　7 |

本文根据句前离合词，在离合句中找到相应的离用法，按照 3.1 节中表 1所示的字符级序列化标注方法进行标注。数据集中共得到32507个句子，随机打乱顺序后按照6:1的比例划分训练集和测试集。数据集划分结果如表 2所示。

数据集	训练集		开发集	测试集
句子数量	正例	负例	7734	4597
	23420	4490		

Table 2: 数据集划分结果

4 模型

以往的离合词自动识别模型主要采用两类方法，一类是主要依据词性、基于规则匹配的方法，另一类是提取词语级别的句子特征来进行分类的方法，因此模型首先需要对文本进行分词和词性标注。因为中文的词与词之间没有确切的分隔符号，现有分词工具容易产生分割错误，而采用不同的分词和词性标注器会产生不同的分词和词性标注结果，对识别结果产生严重干扰。尤其是面向多领域语料时，跨领域的专用词语和生僻词语等的存在也会影响中文分词和词性标注的正确性，其误差将直接导致下游离合词识别错误。

例4 26篇稿件获佳作奖

(a) 26/m 篇/m 稿件/n 获/v 佳作奖/n

(b) 26/m 篇/q 稿件/n 获/v 佳作/n 奖/n

如例4中存在词语“获奖”的离用法，我们应用常见的分词工具jieba对该句进行分词和词性标注，结果为例4(a)。其中，对量词“篇”的词性标注有误，直接影响了根据词性制定的规则匹配结果。同时“佳作奖”被组合，现有的识别模型无法分别获得后语素及中间插入成分。当句子被正确切分标注为例4(b)时，识别模型才能通过字符匹配获得分割开的后语素及中间插入成分，从而可能识别出句子中的离用法。而本文提出的模型输入为自然文本，对文本中的每一个

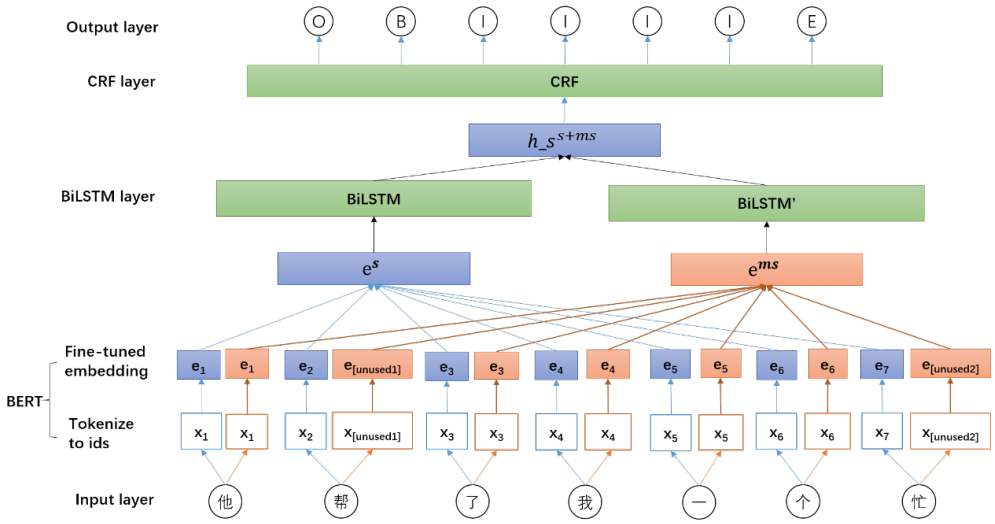


Figure 1: 模型结构图

字符进行标注，因此不需要考虑词边界和词性，省略了分词及词性标注步骤，避免了分词及词性标注错误。

由于句子的可扩展性和离用法的结构多样 (周卫华, 2010)，依据前人总结的规则模板而言，中间插入成分数量不定，每个插入成分的长度也不定，依靠人工总结离用法规则显然是无法穷尽的。本文直接面向识别离用法边界，标注内容表明句中离合词离用法成分位置信息，对中间插入成分的数量和长度不做限制，不仅赋予了模型识别长而复杂的离用法的能力，而且为模型从离用法的学习中识别新的具有离用法的词语提供了可能。

对于提取出了候选离合词但无法用规则以及特征模板表示的句子，现有离合词自动识别模型均将离合词识别任务视为句子分类任务，即用词级的特征模板获取句子表示向量，训练神经网络模型进行分类，判断该句是否为离合句。此类模型对高频离合词及其离用法的识别效果远比对低频离合词及其离用法要好，但同时也容易导致过拟合的问题。例如高频离合词“进球”在训练集正例中出现546次，现有最优模型 (赵聿夕, 2019)将例5划分为正例，但实际上该句中不存在离用法，应为负例。对于低频离合词、尤其是未在训练集中出现的离合词，现有最优模型无法识别出它的离用法。例如“圆梦”在训练集中出现次数为0，现有最优模型将例6划分为负例。为了解决这个问题，本文在获取句子表示时对中文预训练BERT模型进行微调，使其能学习到更多当前语料的信息，并引进了MASK机制，对词语分开使用的前后语素进行遮蔽，减轻词语本身出现频率对识别效果的影响，强化模型对中间插入成分结构的学习。此外，在具体实现中，对前后语素采用不同的遮蔽字符，强调被遮蔽位置的前后顺序关系。

例5 他/r 一共/d 只/d 攻/v 进/v 了/u 5/m 个/q 球/n

例6 中国/ns 帮/v 我们/r 圆/v 了/u 地铁/n 梦/n

基于以上三点改进方向，受经典的BiLSTM-CRF序列标注模型 (Huang et al., 2015)启发，本文提出了一个引入MASK机制的词语离合现象识别模型，即BERT_MASK + 2BiLSTMs + CRF模型，主要由三个模块组成，分别是字符嵌入层、句子编码层和预测输出层。模型整体结构如图 1 所示，下文将对每个部分依次展开介绍。

4.1 基于中文BERT模型的字符嵌入层

在处理中文语料时，字嵌入相较于词嵌入更适合中文没有词边界的特性。为解决离用法结构复杂多样的问题，节省人工总结离用法形式规则、制定特征提取模板的工序，本文提出基于字符的神经网络模型。模型的输入为一个自然句子S，将其转化为字符级序列$\{c_1, c_2, ..., c_n\}$，采用BERT (Devlin et al., 2018)模型得到每个字符的向量表示。为了使字嵌入包含更多的领域含义、更适应当前任务的语料，本文在训练过程中对预训练的中文BERT模型进行微调。首

先根据中文BERT模型的字典得到句子中每个字符$c_i\,(i \in \{1,2,...n\})$的对应序号$x_i$，再调用微调的BERT模型获得字符$c_i$ 的嵌入式表示向量e_i。句子的表示e^s 由字符的表示拼接得到。

$$e^s = e_1 \bigoplus e_2 \bigoplus ... \bigoplus e_n \tag{1}$$

在训练过程中，为减少词语本身出现频率对识别效果的影响，强化对中间插入成分的学习，在BERT层对词语离用法中的前后语素用预设的掩码替换，引入MASK机制 (Liu et al., 2020)，用词表中无意义的字符替换离合词，实现对模型隐藏出现离用法的词语，并做了变体：为了强调词语分开使用的前后语素的出现顺序，用“$[unused1]$” 替换词语的前语素字符，用“$[unused2]$” 替换词语的后语素字符，由此得到新的句子字符序列$S' = \{c_1, ..., '[unused1]', ..., '[unused2]', ..., c_n\}$。将此时通过BERT模型获得的带掩码的句子字符级表示序列记为e^{ms}，该表示中没有词语本身的信息。

$$e^{ms} = e_1 \bigoplus ... \bigoplus e_{[unused1]} \bigoplus ... e_{[unused2]} \bigoplus e_n \tag{2}$$

4.2 基于BiLSTM的句子编码层

经过字符嵌入层，直接拼接字嵌入得到的句子表示缺乏了对字符在句子位置中的考虑。为了获取句子的上下文信息，本文采用BiLSTM模型捕捉字符间的双向语义依赖，以每个BiLSTM单元输出的标签分数来编码句子表示，从而在句子表示中融合了字符的标签信息。将e^s和e^{ms}分别输入两个参数不共享的BiLSTM层模块，记为BiLSTM和BiLSTM’，分别得到输出${h_s}^s$和${h_t}^{ms}$，并分别通过一层线性层以整合经过深层网络学习到的抽象句子特征，并将两个线性层输出拼接得到$hidden^{s+ms}$。此时，句子的表示既包含句子原始的信息，也包含隐藏了词语本身的句子信息，即我们既关注出现离用法的词语，又关注词语离用法的构成规律。

$${h_s}^s = BiLSTM\,(e^s) \tag{3}$$

$${h_t}^{ms} = BiLSTM'\,(e^{ms}) \tag{4}$$

$${h_s}^{s+ms} = {h_s}^s \bigoplus {h_t}^{ms} \tag{5}$$

4.3 基于CRF算法的预测结果输出层

词语离用法中标签之间存在较强的依赖性，比如离用法的起始标签应该是“B”，中间插入成分标签前应该存在起始标签，需要存在完整的BIE三类标签才能算作一个离用法存在等。BiLSTM层获取每个标签的分数只能单独对某个字符的标签进行决策而无法考虑这些约束条件。本文利用CRF算法的得分转移矩阵来强调字符级标签序列的顺序，用维特比算法解码 (2005)得到可能性最大的句子标签预测序列。在训练过程中，使用CRF损失函数，对整个模型不断优化。在预测过程中，仅使用学习未引入MASK机制的字符序列e^s的BiLSTM模块的输出作为CRF层的输入。

5 实验

5.1 模型参数设置

本文采用BERT的中文预训练模型Chinese_L-12_H-768_A-12，并在训练过程中进行微调，获取字嵌入向量。通过BiLSTM层获取句子向量表示，然后输入CRF层获得预测标签序列。使用Adam优化器 (Kingma and Ba, 2014)，BERT模型初始学习率为$2e^{-5}$，其余各层的初始学习率为$1e^{-4}$，每一次循环结束进行学习率衰减，学习率衰减率为0.7，共进行40次循环训练。为减轻过拟合现象，采用dropout方法 (Hinton et al., 2012)。训练时保存在开发集上性能最优的模型并用该模型进行测试。其他超参数如表 3所示。

超参数	参数值
输入句子最大长度	100
句编码向量维度	200
字嵌入向量维度	754
Batch Size	30
Dropout Rate	0.5

Table 3: 超参数设置

5.2 评价方法

赵聿夕 (2019)的工作中样本单句只存在一个离合词,因此评价方法为以句子分类评价PRF指标。在本文的实验中，单句可以出现多个离合词的离用法，无法按照句子正负例评价，因此我们对应地将一个离用法视作一个样本数据。假设模型预测结果R中有m个离用法，测试集正确标注G中有r个离用法，模型识别为负例的正例数FN，模型识别为正例的负例数FP，模型正确识别的负例数TN，模型正确识别的正例数TP。算法 1为识别结果分类算法。

Algorithm 1 识别结果分类

Require: 模型预测结果R,测试集黄金标注G
Ensure: TP，TN，FP，FN
while R中存在下一个句子S' **do**
  $S' \leftarrow S$
  记S的预测结果m个离用法$\{p_1, p_2, \ldots, p_m\}$
  在G中找到S，获取黄金标注的r个离用法$\{g_1, g_2, \ldots, g_r\}$
  if $m = 0, r > 0$ **then**
    $FN \leftarrow FN + 1$
  end if
  if $m = 0, r = 0$ **then**
    $TN \leftarrow TN + 1$
  end if
  if $m > 0, r = 0$ **then**
    $FP \leftarrow FP + 1$
  end if
  if $m > 0, r > 0$ **then**
    if $\{p_1, p_2, \ldots, p_m\} = \{g_1, g_2, \ldots, g_r\}$ **then**
      $TP \leftarrow TP + 1$
    else
      抛出S和$\{p_1, p_2, \ldots, p_m\}$ 进行人工判断分类
    end if
  end if
end while

实验中使用准确率P、召回率R和F1值对模型的性能进行评价，计算公式如下：

$$P = \frac{TP}{TP + FP} \tag{6}$$

$$R = \frac{TP}{TP + FN} \tag{7}$$

$$F1 = \frac{2 * P * R}{P + R} \tag{8}$$

5.3 实验结果

本文在相同的数据集上复现了目前最优模型 (赵聿夕, 2019)的离合词自动识别级联模型，即首先经过规则匹配，其结果与正确标注校对，识别错误的句子再用KNN模型、SVM模型和CNN + LSTM + Attention模型做句子分类，以三个模型的投票结果作为最后的输出，与本文提出的模型的预测结果进行对比，实验结果如表 4所示。根据实验结果对比，本文提出的模型在准确率、召回率和F1值评价上都有了提高，尤其是准确率提高了3.89%，F1值提高了2.85%。

模型	P	R	F1
级联模型	94.59%	97.26%	95.91%
本文模型	**98.48%**	**99.04%**	**98.76%**

Table 4: 实验结果评价

5.4 实验结果分析

5.4.1 模型识别出离合词词表外的离合词

赵聿夕 (2019)的工作提供了一个基于语料自动构建并经过人工校对的离合词词表，包含1554个离合词。据统计，训练集的正例中含有1339个离合词。在模型识别结果中，有67个未在训练集中出现的词语，其中45个是原离合词词表中的离合词，即已经在测试集中标注出来的离合词；22个不在离合词词表中。提取出这22个词语以及其所在句，通过人工判断，有2个正确的离合词，并且它们对应的所有样本都是正例。如表 5 所示，可以将“拄拐”和“失球”加入离合词词表进行扩充。至此，我们得到了一个新的离合词词表，包含离合词1556个。

离合词	离用法	离合句
拄拐	拄 着 拐	现在她拄着拐能走一里多路在运河边散步
失球	失 7 球	斯洛文尼亚队以进18球失7球列第二

Table 5: 模型识别出测试集中存在离合词词表外的离合词及其离用法、离合句

5.4.2 模型对低频离合词的识别情况

我们以出现频数为1作为低频的情况作为考量，在训练集中共出现了416个频数为1的离合词，其中68个不重复的离合词出现在了测试集中。在模型识别结果中，共识别出65个不重复的离合词，对词频为1的离合词识别率达到95.65%，由此可以看出本文提出的模型对低频离合词有良好的识别效果。例如，离合词“加班”分别在训练集中和模型在测试集上的识别结果中出现离用法的情况如表 6所示，该词语在在训练集中仅出现了一次，本文模型在测试集上可以识别出该词语的离用法，并且并不与训练集中出现的离用法完全相同，而是学习到了离用法的组成规律。我们将模型的这种强化离用法构成规律、弱化词语本身使用频率的能力归功于MASK机制的引入。此外，在 5.6节中，我们对MASK机制的效果进行了更深入的探究。

训练集	他曾**加过班**
测试集	晚上**加个班**就可以基本收完
	人民出版社经济编辑部负责人李春生今天**加了班**

Table 6: 离合词“加班”在训练集中和模型识别结果中出现的情况

5.4.3 单句中存在多个离用法

观察模型识别结果，本文模型共识别出了10句含有多个离用法的句子，共计20个离用法，由于原语料库中无法标注出含有多个离用法的句子，如 3.2节中例2和例3。我们对模型识别结果进行人工判断，其中16个离用法识别正确，占比80%。

5.5 对比实验

句法分析是自然语言处理的一个关键问题，对句子内部的语言单位成分进行分析，通过分析它们的依存关系揭示句法结构。本文认为，存在离用法的句子中，出现离用法的词语其前后语素之间必然存在依存关系，也就是体现出句法结构上的关联性。正确识别离用法能够为句法分析提供帮助，为句法分析器的性能提高提供帮助。

为减轻中文依存句法分析器首先对语料进行分词和词性标注处理的干扰，本文选用百度基于深度学习平台飞桨和大规模标注数据研发的DDParser(Baidu Dependency Parser) (Zhang et al., 2020)，在主要来源于新闻和杂志语料的CTB5语料 (Palmer et al., 2005) 上UAS评价指标达到90.31%，该语料与本文所用语料领域相符，并有一定重合，如新华社日报语料。

本节实验所用的实验语料来自实验测试集正例语料，共3824句。我们将模型识别出离用法视为模型识别出前后语素间存在关系。对比实验结果如表 7 所示。

	正确识别	未识别出前后语素间关系	识别正确率
DDParser	3659	165	95.69%
本文模型	3759	65	**98.30%**

Table 7: 对比实验结果评价

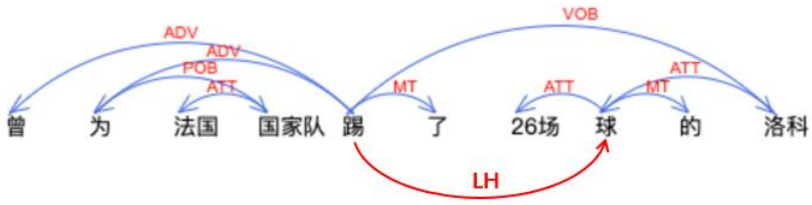
例如，对例7和例8，DDParser未识别出词语分开使用的前后语素之间的依存关系，分析结果输出如图 2 所示，我们认为能够识别出该词语的离用法现象，将语素合并成词语处理更易理解句子的句法结构及句义。为方便展示句法分析结果，我们将DDParser 的分析结果输出其对应的句法分析树，并加入模型的离用法识别结果弧（LH），该弧由前语素指向后语素。

例7 没有一个好的身体就做不好工作

例8 曾为法国国家队踢了26场球的洛科

例7的DDParser分析结果输出如图 2(a) 所示，本文在“做”和“好”之间加上类型为“LH”的关系弧。根据依存关系分析结果，结点“不好”是结点“工作”的定语（ATT），但根据实际句义我们知道“不好”是动词“做”的补语，“不”作为副词在“好”前表示否定。而参考离合现象的识别结果“做不好”是“做好”的离用法，进而可能得到更好的句法分析结果并解读出正确的句义。例8的结果输出如图 2(b) 所示，本文在“踢”和“球”之间加上类型为“LH”的关系弧。根据依存关系分析结果，结点“踢”与结点“洛科”存在动宾关系（VOB），“球”与“洛科”存在定中关系（ATT），结合实际句义显然存在分析错误。而参考本文模型识别出“踢球”的离用法，结点“踢”和“球”具有直接的依存关系，可以将两个结点做概念合并分析，进而得到正确的句义。

(a) 例7 “没有一个好的身体就做不好工作”的依存句法分析树



(b) 例8 “曾为法国国家队踢了26场球的洛科”的依存句法分析树

Figure 2: 依存句法分析树示例

5.6 消融实验

为探究引入MASK机制对离合词识别任务的有效性，我们以不掩盖字符的BERT + 2BiLSTMs + CRF为基线模型（Baseline）做了消融实验，在同一数据集上保持相同的参数设置。实验结果对比如表 8 所示，Baseline + MASK 即为引入MASK机制的模型。

模型	Baseline	Baseline+MASK
TP	3809	3829
FP	56	59
FN	55	37
TN	690	687
P	**98.55%**	98.48%
R	98.58%	**99.04%**
F1	98.56%	**98.76%**

Table 8: 消融实验结果

对比结果分类数，本文提出的模型可识别的正例多20句，不可识别的负例也有明显减少，少了18句，不可识别的正例仅多3句，可识别的负例仅少3句。引入MASK机制的模型在召回率和F1 值上取得了提高，但准确率上有所下降。经过分析模型错误的识别结果，我们发现准确率下降的原因主要来自语料本身标注不准确、有遗漏的未标注离合词及其离用法，即模式识别结果在人工校对中实则是正确的，以及对离用法中间插入成分的结构识别的过拟合问题。

在识别新的具有离用法的词语的效果方面，我们通过模型能否识别未在训练集中出现过的离合词来衡量，如表 9 所示。相较Baseline模型，引入MASK机制的模型在未在训练集中出现的词语数量和不在事先构建的离合词词表中的词语数量上都有所增多，我们可以认为引入MASK机制更有可能获得离合词词表外的离合词，如 5.4.1 节中提及的“拄拐”和“失球”。

模型	Baseline	Baseline+MASK
未在训练集中出现的词语数	60	67
不在离合词词表中的词语数	16	22

Table 9: 模型识别未学习过的离合词的结果

6 总结与展望

本文将二字动词离合现象的自动识别任务视作字符级序列化标注任务，避免了中文分词和词性标注系统带来的误差传播。相较主要基于规则匹配的识别模型，字符级神经网络模型节省了从大规模语料中人工总结规则以及制定特征模板的工作，并且直接面向词语离用法的识别，避免了由词表及规则匹配错误带来的干扰，跳出了模板规则有限性的制约，从而使模型具备了识别结构长而复杂的离用法的能力。同时，引入了掩码机制的模型强化离用法规则的学习，可以有效识别出低频出现的离用法及词语，同时减轻词语本身高频率出现导致的过拟合现象，提高了模型的识别效果。

目前可供研究的关于词语离合现象语料较少，在未来，我们希望借助本文提出的模型进一步扩大面向通用领域的词语离合现象语料库，扩充存在离用法的词语词表，与此同时借助新的标注语料让模型学习到更多、更复杂的离用法，提升模型识别效果。此外，本文提出的模型将离用法识别结果精确到词语分离语素在句子中的位置，并且能够标注出单句中多个离用法，我们希望能通过模型的识别，便于CAMR体系进行词语离合式组合，帮助进一步扩大CAMR可以表示的范围、提高标注效果。

同时，我们希望通过对词语离用法的识别，合并词语概念，为分词系统、依存句法分析等自然语言处理技术提供语言单位成分层面的分析支持，提高句义理解能力。面对语言发展中涌现的词语离用法，我们希望通过本文提出的模型及时捕捉到这类偶发现象，发现具有离用法的词语以及新的词语离用法形式，及时掌握汉语语言发展趋势。

参考文献

Devlin J, Chang M W, Lee K, and Toutanova K. 2018. *Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.*

Graves A and Schmidhuber J. 2005. *Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Networks*, 18(5):602-610.

Hinton G E, Srivastava N, Krizhevsky A, Sutskever I and Salakhutdinov R R. 2012. *Improving neural networks by preventing co-adaptation of feature detectors. arXiv preprint arXiv:1207.0580.*

Huang Z H, Xu W, and Yu K. 2015. *Bidirectional LSTM-CRF Models for Sequence Tagging. arXiv preprint arXiv:1508.01991.*

Kingma D and Ba J. 2014. *Adam: A Method for Stochastic Optimization. arXiv preprint arXiv:1412.6980.*

Liu J L, Chen Y B, and Zhao J. 2020. *Knowledge Enhanced Event Causality Identification with Mention Masking Generalizations. Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI*,3608-3614.

Palmer M, Chiou F D, Xue N W and Lee T K. 2005. *Chinese Treebank 5.0 LDC2005T01. Philadelphia: Linguistic Data Consortium, DOI: https://doi.org/10.35111/j3yz-jw79.*

Zhang S, Wang L J, Sun K and Xiao X Y. 2020. *A Practical Chinese Dependency Parser Based on A Large-scale Dataset. arXiv preprint arXiv: 1611.01734.*

戴茹冰, 侍冰清, 李斌, 曲维光. 2020. 基于*AMR*语料库的汉语省略与论元共享现象考察. 外语研究,37(02):16–23.

李斌, 闻媛, 宋丽, 卜丽君, 曲维光, 薛念文. 2015. 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.

刘博. 2015. 基于语料库的离合词扩展形式自动识别研究. 河北大学.

俞士汶, 段慧明, 朱学锋, 孙斌. 2002. 北京大学现代汉语语料库基本加工规范. 中文信息学报,16(5):49–64.

王海峰. 2010. 基于语料库的现代汉语离合词语义特征考察. 河北师范大学学报*(*哲学社会科学版*)*,33(01):96-100.

王海峰. 2012. 离合词离析结构句的元语用功能考察. 汉字文化,2012(06):10-15.

王薏淼. 2013. 试论汉语词汇中的离合词现象. 黑龙江大学.

臧娇娇, 荀恩东. 2017. 基于*BCC*的离合词离析形式自动识别研究. 中文信息学报,31(01):75-83+93.

张振景, 李新福, 田学东, 王凯. 2010. 基于*SVM*的离合词词义消歧. 计算机科学,43(02):239-244.

赵聿夕. 2010. 面向应用的汉语离合词识别. 南京师范大学.

周卫华. 2010. 现代汉语离合词的扩展形式及特点. 三峡论坛*(*三峡文学.理论版*)*,2010(06):123-127+150.

周卫华, 胡家全. 2010. 中文信息处理中离合词的处理策略. 三峡大学学报*(*人文社会科学版*)*,32(06):39–41.