

文章编号: 1003-0077(2023)06-0025-08

汉语词语离合现象识别研究

周露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)

摘要: 离合现象是指汉语中一种词语的前后语素之间可以插入其他成分, 但分离后表达的意思仍然是一个整体的现象。该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, 以避免自动分词及词性标注的错误传递; 引入掩码机制, 遮蔽句中离合词, 以强化对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; 设计双编码模型, 对原始句子与掩码后的句子分别进行编码。实验结果表明, 该文提出的 BERT_MASK + 2BiLSTMs + CRF 模型比当前性能最优的离合词识别模型提高了 2.85% 的 F_1 值。

关键词: 离合词; 自动识别; 掩码机制; 神经网络

中图分类号: TP391

文献标识码: A

Separated Word Recognition in Chinese

ZHOU Lu¹, QU Weiguang^{1,2,4}, WEI Tingxin³, ZHOU Junsheng¹, LI Bin², GU Yanhui¹

(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. Zhongbei College, Nanjing Normal University, Danyang, Jiangsu 212334, China)

Abstract: The separated words, as a unique grammatical phenomenon in Chinese, refers to the components of words can be divided by other grammatical constituents without changing the word meaning. This paper proposes a character-level sequence tagging method for automatic recognition of separated two-character verbs via the MASK mechanism. We mask the separated words, and assign different embeddings to the first and second components of a word to emphasize the morpheme sequence order. A double coding model is designed to encode the original sentence and the masked sentence, respectively. Experimental results show that the proposed BERT_MASK + 2BiLSTMs + CRF model improves the F_1 -value by 2.85% compared with the state-of-the-art model.

Keywords: separated words; automatic recognition; MASK mechanism; neural network

0 引言

汉语词语的离合现象指的是词语的前后语素之间可以插入其他成分, 但在分离后所表达的意思仍然是一个整体的语言现象。如例 1 所示, 词语“帮忙”以分离的形式出现在句子中, 可以插入多种句法

成分。

例 1 帮大忙→帮了大忙→帮了一个大忙→帮了我一个大忙

虽然词语语素彼此不相邻, 中间可以插入多种成分, 但在理解句子语义时仍需要将其视为一个整体概念。这种语义整合、空间分离的现象对中文分词、句法解析等任务提出了挑战: 一是单字词语与

收稿日期: 2022-01-16 定稿日期: 2022-04-28

基金项目: 国家社会科学基金(21&ZD288)

语素之间的划界；二是词语与短语的划界^[1]，因此离合词识别对于中文自然语言处理与理解有着非常重要的意义。

出于对表达简洁、语用丰富的需求，词语的离用法可以灵活变化，越是表现日常行为，离合词离析频率越高、离用法越丰富^[2-3]。原本不可离合使用的词，在类化使用中产生离用法。随着使用者增多和时间推移，最终会进入到离合词的范畴中^[4]。随着互联网的发展以及社交媒体的繁荣，人们的语言习惯日趋随意，催生了许多新的偶发性离合词。如新词“破防”是指“因遇到一些事或看到一些信息后情感上受到很大冲击，内心深处被触动，心理防线被突破”，人们常使用其离用法“破大防”来加重其程度。本文将这种已产生离用法但未被收录为离合词的词语称为未登录离合词。

在以往的离合词识别研究中，很多学者注意到了离用法较强的规则性，多采用规则匹配的方法。目前规则匹配方法将中间插入成分限制在四个词以内，而离合词使用灵活，可能会超出模板匹配范围；插入成分也可以是复杂的短语，如“帮了他一点力所能及的忙”，因此造成规则失效，识别效果不佳。此外，目前的方法无法识别偶发的离合词。

针对上述问题，本文提出了字符级神经网络模型 BERT_MASK + 2BiLSTMs + CRF 模型，着重研究自然文本中二字动词离合现象的自动识别。该模型直接面向离用法边界的识别，对中间插入成分不做限制；为增强对离用法构成规律的学习，引入掩码(MASK)机制^[5]，用掩码符号遮蔽发生分离的词语前后语素，从而弱化词语本身的影响，对偶发及低频现象有更好的识别效果。

实验结果显示，在赵聿夕^[6]构建的离合词语料库上，本文模型相比该文提出的最优模型，提高了 2.85% 的 F_1 值。

1 相关工作

1.1 离合词语料标注

目前，公开的离合词语料少，且标注规范不统一。周卫华等人^[7]总结了 8 种离用法，在构建动宾式和并列式离合词的词库时对每个离合词离用法做了详细的描写，在标注时利用词库将语料文本中的离用法标注出来。CCL 现代汉语语料库^[8]对离合

词的标注方法为：只有当后字为语素时，将其标注为语素，但并不标注与前字的关系，可见这种标注方法并不全面。

中文抽象语义表示(CAMR)标注体系^[9]对离合词进行了较为全面的标注。戴茹冰等人^[10]将动宾式离合词的两个分离节点做合并处理。如图 1 所示，词语“敬礼”的两个语素在句子中的位置是分离的，在 CAMR 图中被合并为一个节点“x5_x8”，其概念为原词“敬礼”，中间插入成分根据其语义在下一层进行标注。

x1_警卫员 x2_向 x3_毛 x4_主席 x5_敬 x6_了 x7_个 x8_礼 x9_。
(x5_x8 / 敬礼-01 :arg0() (x1 / 警卫员) :arg1(x3/毛) (x27 / person :arg0-of() (x28 / have-org-role-91 :arg2() (x4 / 主席)) :name() (x3 / name :op1 x3/毛)) :aspect() (x6 / 了) :cunit() (x7 / 个))

图 1 CAMR 动宾式离合词表示示例

1.2 离合词的自动识别

1.2.1 基于规则匹配的方法

以往的离合词识别研究工作多基于大规模语料对离合词规则进行人工归纳，设计规则匹配算法。刘博^[11]基于 CCL 语料库，设计了基于词表和规则匹配的离合词自动识别系统。藏娇娇等人^[12]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 个常用的动宾式离合词进行研究，以“0/1”标注句子中是否存在离用法，无法定位离合词的位置，也不能处理一个句子中有多个离合词的现象。

前文所述的规则匹配算法主要依赖人为制定特征模板，需要在使用中不断更新。

1.2.2 基于机器学习和神经网络的方法

针对离合词分开后保持词义不变的性质，张振景^[13]将词义消歧的方法应用于离合词识别，根据句子的词特征、词性特征、词及词性特征设计特征模板，采用 SVM 模型建立句子分类器。赵聿夕^[6]以新华社十四年新闻语料作为原始语料，根据人工制定的离合词规则动态生成离合词词表。在识别过程中，首先根据词表进行规则匹配，对于无法利用规则匹配识别的句子则通过特征模板获取句子表示，分别用 KNN 模型、SVM 模型和 CNN + LSTM +

Attention 模型进行句子二分类任务,将它们的预测结果进行投票作为最后的输出,最终构建了一个离合词自动识别级联模型。

此类识别模型利用大规模标注数据进行监督学习,但仍需要人为制定模板以获取句子特征向量。目前利用深度神经网络进行离合词识别的工作较少,但鉴于其在多个领域所展现的强大性能,本文采用深度学习模型,以期提高离合词识别性能。

1.2.3 存在的问题

以往的离合词自动识别模型主要采用两类方法,一类主要依据词性制定规则进行匹配,另一类则是提取词语级别的句子特征来进行分类,两类模型都需要先利用分词标注器对文本进行自动分词和词性标注,而离合词的存在本身就影响了分词性能,其误差将直接导致下游离合词识别错误。如例 2 中存在词语“获奖”的离用法,本文用分词工具 jieba^① 对该句进行分词和词性标注,结果为例 2(a)。

例 2 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

其中,量词“篇”的词性标注有误,可能会影响根据词性制定的规则匹配结果。同时“佳作奖”被分为一个词,现有的模型无法分别获得后语素“奖”和中间插入成分“佳作”与模板进行匹配。当句子被切分标注为例 2(b)的情况时,识别模型才能匹配单字“获/v”和“奖/n”,进而识别出句子中的离用法。

此外,由于句子具有可扩展性,离合词离用法结构多样^[14],依靠人工总结离用法规则显然无法将其穷尽,无法识别未登录离合词;识别效果直接受到分词及词性标注结果的影响,且对插入成分有所限制,无法识别插入成分复杂的离用法。利用特征模板获取句子表示的有监督学习模型对高频离合词的识别效果远好于低频离合词,但也容易导致过拟合问题。

综上所述,如何实现端到端的离合词识别,避免人工特征工程和模型的系统错误传播是本文的重点研究内容。

2 离合词识别任务研究

2.1 任务定义

本文采用字符级序列化标注方法进行离用法识别任务。模型输入为一个句子 $S=c_1,c_2,\cdots,c_n$,其中 c_i 表示句中第 i 个字符, $i\in\{1,2,\cdots,n\},n=|S|$,

预测输出为对应句子中每个字符的标签序列 $Y=y_1,y_2,\cdots,y_n$,其中 $y_i\in\{B,I,E,O\}$ 。

本文将存在离用法的句子称为离合句,表 1 中给出“帮忙”所在的一句离合句的标注示例。

表 1 离合句标注示例

文本	他	帮	了	我	一	个	忙
标签	O	B	I	I	I	I	E

标签 B 表示该字符是离用法的起始字符,I 表示该字符属于中间插入成分,E 表示该字符是离用法的结尾字符,O 标注了句子中离用法外的其他成分。

2.2 数据集构建

本文采用的数据集源自赵聿夕工作中构建的离合词语料^[6],称为原语料,该语料中每一条数据的格式为“离合词 离合句”,词句一一对应,如例 3~例 6 所示。

- 例 3 帮忙 他帮了我一个忙
- 例 4 鼓劲 鼓实劲而不鼓虚劲
- 例 5 洗澡 可等洗完澡下了班
- 例 6 下班 可等洗完澡下了班

例 3 中离合词“帮忙”对应其后离合句。当同一离合词在一个句子中出现多次离用法,如例 4 句中存在“鼓劲”的两次离用法,原语料不做表示,无法定位到其在句中出现的位置。当一个句子中有多个离合词的离用法时,原语料只能将不同的离合词与该句分别组合,导致该句重复出现,如例 5、例 6 所示,两个离合词分别是“洗澡(洗完澡)”和“下班(下了班)”,句前离合词不同而句子相同。由此可见,一句多例的情况在原语料中无法直接标注出来。

为解决以上标注问题,本文根据句前离合词,在句中找到相应的离用法进行标注,一组 BIE 标签即为一个离用法,表 2 为例 5、例 6 所对应的同一离合句的标注示例。

表 2 一句多例标注示例

文本	可	等	洗	完	澡	下	了	班
标签	O	O	B	I	E	B	I	E

本文构建的语料以一个离用法作为一个样本,因此一个句子中可以出现多个样本,统计单句中样本数分别为 0、1、2 和 3 的情况,其中单句样本数为 0,即不包含离合词的负例句。为进行有监督学习,

① <https://github.com/fxsjy/jieba>

随机划分数据集,相关统计数据如表 3 所示。

表 3 数据集情况

数据集		训练集	开发集	测试集
单句 样本数	0	4 427	1 501	687
	1	23 100	7 642	3 902
	2	92	34	19
	3	2	0	0
总样本数		23 290	7 710	3 940
总句数		27 621	9 177	4 608

3 模型

本文提出了一个汉语词语离合现象自动识别模型,即 BERT_MASK + 2BiLSTM_s + CRF 模型,整体结构如图 2 所示。

首先,为避免自动分词及词性标注错误,本文采用字符级序列化标注方法,可以直接描述离用法边界,对中间插入成分的数量和长度不做限制,赋予模型识别长而复杂的离用法的能力。

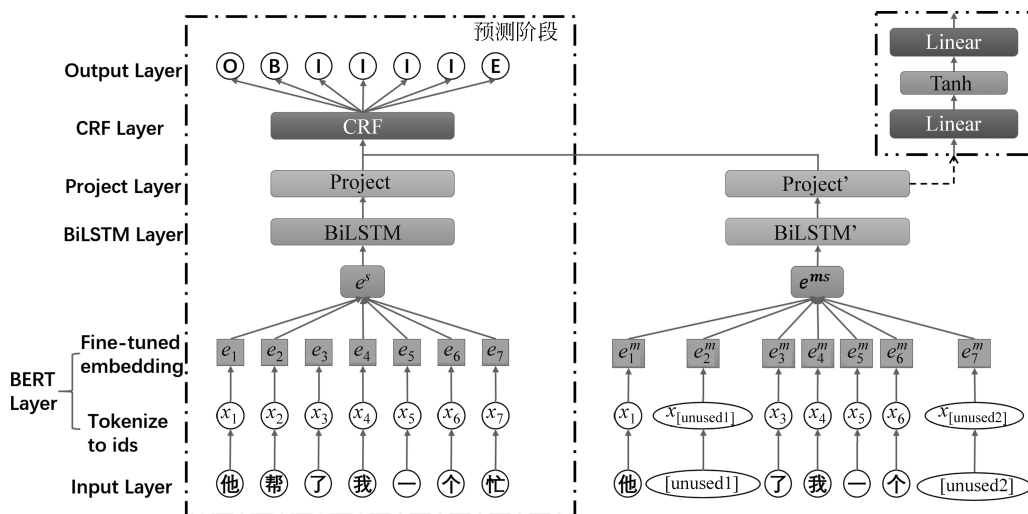


图 2 模型结构图

其次,为减轻词频对模型识别的影响,引入了 MASK 机制。对词语分离的两个语素用掩码字符进行遮蔽,强化模型对中间插入成分结构的学习。为强调顺序关系,分别使用不同的掩码字符来表示前后语素的位置区别。

识别自然文本中的离合词是模型的预测任务,因此在预测阶段,模型并不知道句中是否存在离用法,不能对分离的前后语素采用掩码机制,造成了模型在训练与预测两个阶段的结构不对等。因此本文在训练阶段针对句子的原始信息和带掩码的信息设计了双编码模型,在预测阶段仅使用对原始文本进行编码的模块。

下文将对模型的三个主要模块依次展开介绍,分别是字符嵌入层、句子编码层和预测输出层。

3.1 字符嵌入层

为应对离用法复杂多样的构成形式,本文提出字符级神经网络模型。模型的输入为一个自然句子

S , 将其转化为字符序列 $c_1, c_2, \dots, c_n$ 。首先根据 BERT^[15] 中文预训练模型的字典得到每个字符 c_i 的对应序号 x_i , 然后调用模型得到每个字符的向量表示 e_i 。在训练过程中对 BERT 模型进行微调,使其更适应当前语料。句子的表示 e^s 由 e_i 拼接得到,如式(1)所示。

$$e^s = e_1 \oplus e_2 \oplus \dots \oplus e_n \quad (1)$$

在训练阶段,为缓解模型由于离合词频数不同而可能造成的影响,强化对中间插入成分的学习,在 BERT 层引入 MASK 机制。用词表中无意义的字符作为掩码字符,替换离用法中的前后语素;为了强调语素顺序,用不同的掩码分别替换前后语素字符,得到新的字符序列 $S' = c_1, \dots, '[\text{unused1}]', \dots, '[\text{unused2}]', \dots, c_n$ 。如式(2)所示,获得没有离合词语本身信息的句子表示,记为 e^{ms} 。

$$e^{ms} = e_1 \oplus \dots \oplus e_{[\text{unused1}]} \oplus \dots \oplus e_{[\text{unused1}]} \oplus \dots \oplus e_n \quad (2)$$

3.2 句子编码层

由字嵌入直接拼接得到的句向量缺少了对字符在句中位置的考虑,因此本文采用 BiLSTM 模型捕捉双向的字符间语义依赖。将 e^s 和 e^{ms} 分别经过相同结构、不同参数的网络处理: 经过 BiLSTM、线性网络层(记为 Project),降低向量维度,以整合深层网络学习到的抽象句子特征。最后将两个输出拼接,得到 h_i ,如式(3)~式(5)所示。此时,离合句的表示既包含原句信息,也包含隐藏离合词的信息,即模型对离合词和离用法的构成规律都给予了关注。

$$h_i^s = \text{bilstm}(e^s) \quad (3)$$

$$h_i^{ms} = \text{bilstm}'(e^{ms}) \quad (4)$$

$$h_i = \text{project}(h_i^s) \oplus \text{project}'(h_i^{ms}) \quad (5)$$

3.3 预测输出层

离用法的标签之间存在一定的依赖性,一组完整的 BIE 标签才能算作一个离用法,即顺序出现的“B”“E”之间必须存在“I”标签。 h_i 的每个元素只表示每个字符标签的预测概率,而无法考虑这些约束条件。

本文利用 CRF 算法的转移矩阵来强调标签顺序,用维特比算法^[16]解码得到标签预测序列。在训练过程中,使用 CRF 损失函数,如式(6)~式(8)所示。

$$P_i = e^{\text{EmissionScore} + \text{TransitionScore}} \quad (6)$$

$$P_{\text{total}} = \sum_{i=1}^N P_i \quad (7)$$

$$\text{LossFunction} = -\log \frac{P_{\text{RealPath}}}{P_{\text{total}}} \quad (8)$$

设标签序列的每种可能路径的分数为 $P_i, i \in \{1, 2, \dots, N\}$,共有 N 条路径,总分为 P_{total} ,其中真实路径分数为 P_{RealPath} 。发射分数 EmissionScore 表示当前标签被预测为某一分类的分数,转移分数 TransitionScore 表示从一个标签类别到另一个标签类别的概率分数, e 是自然常数。

在预测过程中,模型无法事先定位离用法,仅使用输入 e^s 的模块输出作为 CRF 层的输入,如图 2 左侧“预测阶段”所示。

4 实验

4.1 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-12^①,并在训练过程中进行微调。初始学习率的设置为 BERT 层 $2e-5$,其余各层 $1e-4$ 。使用 Adam 优化器^[17],每次循环结束进行学习率衰减,衰减率为 0.7,共进行 40 次循环训练。其他超参数如表 4 所示。

表 4 超参数设置

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	768
Batch Size	30
Dropout Rate	0.5

4.2 评价方法

以前的工作将离合词识别任务视为句子二分类任务,以句子为单位计算评价指标。而在本文的实验中,利用多组 BIE 标签,单句中可以出现多个离合词的离用法,因此本文对应地将一个离用法视作一个样本数据。

为方便表示,单个离合句的标注形式可转化为“S| p_1 | p_2 |...| p_m |”,用“|”符号隔开多个离用法标注。若句中不存在离用法,则样本形式为“S”。 p_j 表示句中第 j 个离用法的标注, $j \in \{1, \dots, m\}$, m 为句中离用法的数量。 $p_j = \text{“}B_jE_j\text{”}$, B_j 表示当前离用法起始位置索引, E_j 表示当前离用法结尾位置索引,用空格隔开两个索引标注。

例 7 鼓实劲而不鼓虚劲| 0 2 | 5 7 |

例 8 可等洗完澡下了班| 2 4 | 5 7 |

在例 7 中,存在“鼓劲”的两次离用法,分别为下标 0 到 2 和下标 5 到 7;在例 8 中存在两个离合词的离用法,分别是下标 2 到 4 “洗完澡”,对应词语“洗澡”,下标 5 到 7 “下了班”,对应词语“下班”。

假设模型预测结果 R 中有 m 个离用法,测试集黄金标注 G 中有 r 个离用法。记模型识别为负例、黄金标注为正例的样本数为 FN,识别为正的负例数为 FP,识别为负的负例数为 TN,识别为正的负例数为 TP。使用准确率 P 、召回率 R 和 F_1 值对模型的性能进行评价,计算如式(9)~式(11)所示。

① https://storage.googleapis.com/bert_models/2018_11_03/chinese_L-12_H-768_A-12.zip

$$P = TP / (TP + FP) \quad (9)$$

$$R = TP / (TP + FN) \quad (10)$$

$$F_1 = 2 \times P \times R / (P + R) \quad (11)$$

4.3 实验结果

实验中保存在开发集上性能最优的模型并用该模型进行测试。本文在相同的数据集上复现了目前性能最优的离合词自动识别级联模型(赵聿夕^[6])与本文实验进行对比,采用 4.2 节的评价方法进行模型评价。如表 5 所示,本文提出的模型评价有显著性能提高, F_1 值提高了 2.85%。

表 5 实验结果评价 (单位: %)

模型	P	R	F_1
赵聿夕 ^[6]	94.59	97.26	95.91
本文模型	98.48	99.04	98.76

4.4 实验结果分析

4.4.1 低频离合词

为了评价模型对低频离合词的识别效果,本文观察了 416 个训练集中出现 1 次的离合词。其中 68 个离合词出现在测试集中,模型识别出 65 个,识别率达到 95.59%,由此可见本文模型对低频离合词有良好的识别效果。例如,离合词“加班”在训练集中和模型在测试集上的识别结果中出现的离用法如表 6 所示。本文模型识别出的离用法与训练集中出现的形式并不相同,可见模型学习到了离用法的构成规律。

表 6 低频离合词识别情况

数据集	离合句
训练集	他曾加过班
测试集 识别结果	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

4.4.2 未登录离合词

为测试模型的泛化能力,在测试集中设置了 45 个训练集中未出现的离合词。在模型的识别结果中,这些离合词均被识别出。此外,模型还识别出 2 个未被标注的未登录离合词:“拄拐”和“失球”,经过人工判断,其对应的样本均为正确的离用法,如表 7 所示。

表 7 未登录离合词识别情况

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

由此可知,模型学习到了离用法的构成规律,从而可以识别出以离用法形式出现的未登录离合词。这些未登录离合词可能是训练集中未出现的,也可能是尚未被归入离合词范畴的词语。

本文认为模型强化离用法构成规律、弱化词语本身的能力得益于 MASK 机制的引入,在 4.6 节将通过消融实验深入探究其为模型带来的学习效果。

4.4.3 一句多例

在测试集上,得到 10 句模型识别为含有多个离用法的句子,共计 20 个离用法,其中 16 个识别正确,占比 80%。

4.5 对比实验

依存句法分析是分析句子内部语言成分之间的依存关系、揭示句法结构的重要自然语言处理任务。考虑到离合词在分离后仍作为一个整体来表义的特性,本文认为,在离合句中分隔开的离合词前后语素语义上也是相互依存的。

为减轻依存句法分析器对语料进行分词和词性标注处理的干扰,本文选用基于深度学习平台飞桨和大规模标注数据研发的 DDParse (Baidu Dependency Parser)^[18]。

本实验提取测试集中的正例作为实验语料,共计 3 824 句。为了对比两种模型的输出结果,将离用法视为前后语素间存在依存关系,实验结果如表 8 所示。

表 8 对比实验结果

	正确识别	未识别出前后 语素间关系	识别正确率/%
DDParser	3 659	165	95.69
本文模型	3 759	65	98.30

本文识别出离用法后,将前后语素合并成原词处理。为方便比较展示,本文在句法分析树上加入模型的离用法识别结果,表示为“LH”类型的弧,由前语素指向后语素。

例 9 曾为法国国家队踢了 26 场球的洛科
输出例 9 的依存句法分析树,如图 3 所示。

DDParser 未识别出词语“踢球”分离的前后语素之间的依存关系。根据依存分析结果,“踢”与“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,“踢”和“球”具有直接的依存关系,增加由“踢”指向“球”的 LH 类型弧,可以将两个节点概念合并分析,更易理解句子的句法结构及句义。因此,正确识别离用法能够为句法分析提供帮助,从而提高句法分析器性能。

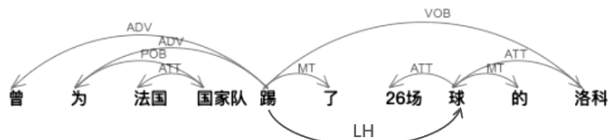


图3 例9的依存句法分析树

4.6 消融实验

为探究引入 MASK 机制对离合词自动识别的有效性,以本文模型为基线模型(Baseline)做了消融实验,-MASK 表示不引入该机制,实验结果对比如表 9 所示。

相较-MASK,Baseline 在 R 和 F_1 值上取得了提高,但在 P 上有所下降,通过对模型错误的识别结果进行分析,其原因主要有:语料本身标注不正确;有遗漏标注的离用法,即人工校对认为模型识别结果是正确的,但在测试集中未被标注出来;大部分离用法中间插入成分单一,模型学习遇到过拟合问题,如简单结构“A了B”,模型将“下了班车”错误识别为离合词“下班”的离用法。

表9 消融实验结果评价

模型	-MASK	Baseline	差异
TP	3 809	3 829	+20
FP	56	59	+3
FN	55	37	-18
TN	690	687	-3
P	98.55%	98.48%	-0.07%
R	98.58%	99.04%	+0.46%
F_1	98.56%	98.76%	+0.20%

本文通过模型能否识别未在训练集中出现过的离合词来衡量识别未登录离合词的效果。如表 10 所示,相较-MASK,Baseline 识别出未在训练集中出现的词语数量有所增多,甚至识别出了在测试集

中未被标注的离合词,例如 4.4.1 节中提及的“拄拐”和“失球”,因此本文认为引入 MASK 机制可获得未登录离合词。

表10 模型在识别未登录离合词上的表现

词语数\模型	-MASK	Baseline
未在训练集中出现	60	67
未在测试集中标注	16	22

5 总结与展望

本文将二字动词离合现象的自动识别视作序列化标注任务,采用字符级标注避免了中文自动分词和词性标注带来的误差传播,节省了大量人工总结规则、制定特征模板的工作,并且直接面向离用法识别,跳出了规则模板有限性的制约。引入 MASK 机制,强化了模型对离用法规则的自动学习,可以有效识别低频离合词和未登录离合词,缓解模型学习高频离合词特征而忽略离用法构成的规则性的问题。实验结果表明,该方法提高了离合现象自动识别性能, F_1 值达到 98.76%。

目前关于词语离合现象形式的标注不够全面,在未来,希望能针对多字离合词、倒装形式等特殊离用法设计更科学的标注方法,提升有监督离合现象识别模型的学习效果。

今后的工作将在以下方面开展:尝试将离合现象自动识别模型应用于 CAMR 标注中,提高 CAMR 对离合现象标注的性能;利用离合现象语素间依赖关系的识别,为分词系统、依存句法分析等自然语言处理技术提供语言内部成分的分析支持。

参考文献

- [1] 王燕明.对外汉语离合词扩展形式及教学[D].北京:北京语言大学硕士学位论文,2009.
- [2] 王海峰.基于语料库的现代汉语离合词语义特征考察[J].河北师范大学学报(哲学社会科学版),2010,33(01): 96-100.
- [3] 王海峰.离合词离析结构句的元语用功能考察[J].汉字文化,2012(06): 10-15.
- [4] 王慧森.试论汉语词汇中的离合词现象[D].哈尔滨:黑龙江大学硕士学位论文,2013.
- [5] LIU J, CHEN Y, ZHAO J. Knowledge enhanced event causality identification with mention masking generalizations[C]//Proceedings of the 29th International

- Joint Conference on Artificial Intelligence, Yokohama: International Joint Conferences on Artificial Intelligence, 2020: 3608-3614.
- [6] 赵聿夕. 面向应用的汉语离合词识别[D]. 南京: 南京师范大学硕士学位论文, 2019.
- [7] 周卫华, 胡家全. 中文信息处理中离合词的处理策略[J]. 三峡大学学报(人文社会科学版), 2010, 32(06): 39-41.
- [8] 俞士汶, 段慧明, 朱学锋, 等. 北京大学现代汉语语料库基本加工规范[J]. 中文信息学报, 2002, 16(5): 49-64.
- [9] 曲维光, 周俊生, 吴晓东, 等. 自然语言句子抽象语义表示 AMR 研究综述[J]. 数据采集与处理, 2017, 32(01): 26-36.
- [10] 戴茹冰, 侍冰清, 李斌, 等. 基于 AMR 语料库的汉语省略与论元共享现象考察[J]. 外语研究, 2020, 37(02): 16-23.
- [11] 刘博. 基于语料库的离合词扩展形式自动识别研究[D]. 保定: 河北大学硕士学位论文, 2015.
- [12] 臧娇娇, 苟恩东. 基于 BCC 的离合词离析形式自动识别研究[J]. 中文信息学报, 2017, 31(01): 75-83.
- [13] 张振景, 李新福, 田学东, 等. 基于 SVM 的离合词词义消歧[J]. 计算机科学, 2016, 43(02): 239-244.
- [14] 周卫华. 现代汉语离合词的扩展形式及特点[J]. 三峡论坛(三峡文学·理论版), 2010(06): 123-127.
- [15] DEVLIN J, CHANG M, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis: Association for Computational Linguistics, 2019: 4171-4186.
- [16] GRAVES A, SCHMIDHUBER J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures[J]. Neural Networks, 2005, 18(5-6): 602-610.
- [17] KINGMA D, BA J. Adam: A method for stochastic optimization[C]//Proceedings of ICLR, 2015: 1-15.
- [18] ZHANG S, WANG L, SUN K, et al. A practical Chinese dependency parser based on a large-scale dataset [J/OL]. arXiv preprint arXiv: 1611.01734, 2020.



周露(1996—), 硕士研究生, 主要研究领域为自然语言处理。

E-mail: lu_zhou2@163.com



魏庭新(1980—), 博士, 讲师, 主要研究领域为计算语言学。

E-mail: tingxin_wei@163.com



曲维光(1964—), 通信作者, 博士, 教授, 博士生导师, 主要研究领域为自然语言处理。

E-mail: wgqu_nj@163.com