文章编号：1003-0077（2011）00-0000-00

基于多模型融合的汉语介词短语识别*

刘彤，黄德根，张聪

（大连理工大学 计算机学院，辽宁省 大连市 116024）

摘要：提出了一种多模型融合的介词短语识别方法，不仅能识别并列型介词短语，而且提高了嵌套型介词短语的识别精度。首先，利用简单名词短语识别模型识别出语料中的短语信息并进行融合，简化语料，降低介词短语内部复杂性；其次，用 CRF 模型识别嵌套的内层介词短语，即若存在嵌套则识别嵌套的内层，若无嵌套则识别该介词短语；最后，将初始语料中识别出来的内层介词短语进行分词融合并修改其特征信息，重新训练外层介词短语识别模型进行识别。在内、外层介词短语自动识别后，利用双重错误校正系统对识别的介词短语进行校正。在 2000 年《人民日报》语料中进行五倍交叉实验，结果表明，该方法识别的介词短语的正确率、召回率、F 值分别为 94.11%，94.02%，94.06%，比基于简单名词短语的介词短语识别方法（baseline）分别提高了 1.09、1.07、1.08 个百分点，有效提高了介词短语识别的性能。

关键词：简单名词短语；分词融合；分层嵌套结构；双重错误校正系统

中图分类号：TP391　　**文献标识码：**A

Chinese Prepositional Phrase Recognition Based on Combined Multiple Models

LIU Tong, HUANG Degen, ZHANG Cong

(School of Computer Science and Technology, Dalian University of Technology, Dalian, Liaoning 116024, China)

Abstract: A method of prepositional phrase recognition based on merged multiple model is proposed to recognize coordinate prepositional phrases and improve the performance of nested prepositional phrase recognition. First, a simple noun phrase recognition model is used to identify and merge the phrases in the corpus in order to simplify corpus and reduce internal complexity of prepositional phrases; Then, the CRF model is used to identify the inner layer of the nested prepositions phrases, i.e. if the preposition phrases is nested, recognize the inner layer, otherwise, recognize the whole preposition phrase; Finally, merge the recognized inner prepositional phrases in the corpus and modify the feature information in order to train a new model for outer prepositional phrase recognition. In addition, after the recognition of both inner and outer prepositional phrases, a double error correction system is used to correct the recognized phrases. Five-fold experiments are conducted on the corpus of People's Daily of 2000 including 7028 prepositional phrases, and the results achieve 94.11% in precision, 94.02% in recall, and 94.06% in F-measure, which are improved by 1.09%, 1.07%, 1.08% respectively than the simple noun phrases based prepositional phrase identification method (baseline).

Key words: simple noun phrase; word segmentation fusion; hierarchical nested structure; double error correction system

1 引言

介词短语（Preposition Phrase，PP）作为一个重要的短语类型，在汉语中占有很大的比例。文献[1]曾对包含十万字、六万词的语料所包含的介词短语的句子书进行过统计分析，结果表明，科技类文章含有介词短语的句子占 57%，而政论类文章包含介词短语的句子占 63%。介词短语大多作为句子状语和补语，正确识别介词短语能够提高句子结构的清晰度，

* **收稿日期：**　　**定稿日期：**

基金项目：国家自然科学基金（61672127，61672126）

降低句子复杂度，为下一步句法分析提供有效信息。提升介词短语的识别精度对于信息检索及文本分类效果都有较大的提升，对于浅层句法分析、机器翻译等研究具有极其重要的意义。

现有的介词短语识别研究主要集中在介词短语学习模型的选择和介词短语的层次关系分析两方面。在学习模型的选择方面，文献[2]提出了一种结合可信搭配关系和三元边界统计模型的识别方法，根据固定搭配制定两个搭配模板，利用模板获取可信搭配关系，根据其识别介词短语，结合三元模型和规则识别剩余的介词短语；文献[3]提出了基于最大熵模型的识别方法，首先对介词短语抽取标记介词短语的特征，然后利用最大熵模型识别语料中的介词短语，最后利用依存树库中的介词短语边界词的语法知识对识别结果进行校正；文献[4]提出了基于 HMM 模型的识别方法，先利用 HMM 模型识别语料中的介词短语，然后利用依存语法对识别结果进行校正。

在层次关系方面，文献[5]提出了基于双层 CRF 模型的识别方法，针对介词短语的特点选择双层 CRF 模型进行识别，并制定规则对结果进行校正；文献[6]在文献[5]基础上提出了基于多层 CRF 模型的介词短语识别方法，通过 CRF 模型利用多个有效特征及符合特征模板从后向前逐个识别语料中的介词短语，然后利用基于转换的驱动学习方法制定了规则转换集，并用其对识别结果进行校正。

另外，文献[7]提出了基于简单名词短语的介词短语识别方法，简单名词短语（Simple Noun Phrase，SNP）是文献[8]提出的内部不包含复杂修饰成分的名词短语，先识别出介词短语中的 SNP 并进行融合，简化介词短语的内部结构，降低介词短语识别的复杂性，再进行介词短语识别，是目前发表的识别效果相对较好的方法。

通过对以往的研究进行分析发现，当前介词短语识别中认可度较高的模型是 CRF 模型，但在介词嵌套方面的研究还不够细致。文献[6、7]已经考虑过介词短语嵌套情况，但并未对介词短语的结构层次进行深入的分析。他们采取的是将句子介词短语从后向前依次识别的方法，不能很好的解决介词短语的嵌套、并列结构并存的情况。

本文提出多模型融合的介词短语识别方法，通过分词融合将语料中的简单名词短语信息融合以简化语料，并对其训练得到内层训练模型，使用该模型识别测试语料中的内层介词短语，规则校正后将初始语料中的内层介词短语进行融合并修改其标注信息，重新进行训练得到嵌套介词识别模型，再将测试语料识别出的内层介词短语融合修改标注信息后用嵌套模型进行识别，规则校正后得到最终结果。本文在介词短语识别时着重考虑介词短语层次特点，将同等层级的介词短语同时识别，降低某层识别错误给其他层次所带来的影响。

2 理论基础

2.1 序列标注

条件随机场(Conditional Random Fields，简称 CRFs)模型[9]能够充分利用词语的上下文信息特征，适用于序列标注工作。CRF 通过学习训练数据获得使训练样本标注序列在标注序列集合中条件概率最大的特征集合和特征权重。

序列标注需要将语料进行分词及词性标注，经过分词及词性标注后的汉语句子 $S= W_1/P_1 W_2/P_2 W_3/P_3 \ldots W_i/P_i \ldots W_N/P_N$（$W_i$ 为第 i 个词，P_i 为第 i 个词的词性，N 为词的个数）。

简单名词短语识别使用 BIO 标记边界状态，其中，B 表示简单名词短语的左边界，I 表示内部词语或右边界，O 表示不在短语内部的词语。即，对于输入的词语序列 $S= W_1/P_1 W_2/P_2 W_3/P_3 \ldots W_i/P_i \ldots W_N/P_N$，任务的目标为获得一个对应的标注序列 $T^*=T_1 T_2 T_3 \cdots T_N$，使得该序列在所有可能的标注序列中概率最大，其中 $T_i \in \{B, I, O\}$。

介词短语自动识别的任务是标注出句子中所有介词短语，而不对介词短语的内部成分进行分析。首先，把句子 S 经过分词及词性标注处理为“word(1)/pos(1) word(2)/pos(2) …word(i)/pos(i)…word(n)/pos(n)”的格式（word(i)为第 i 个词，pos(i)为第 i 个词的词性）。然后，获

得对应的标注序列 $T^*= T_1 T_2 T_3 \cdots T_N$，使该序列在所有可能的标注序列中概率最大，其中，$T_i$ 可能取值有 BIEO，“B”表示介词短语的首词，“I”表示介词短语的内部词，“E”表示介词短语的尾词，“O”表示介词短语的外部词语。最后，输出标注序列不为“O”的所有词。

2.2 分词融合

本文中分词融合是指根据已经识别出来的序列标注结果进行词语合并，并制定规则修改合并后的词的词性等特征。主要包括两个方面：SNP 融合、内层介词短语融合。

SNP 融合：首先识别出语料中的简单名词短语，然后将相应的词语进行合并，并将融合后的短语词性标注为“COM-NOUN”。例如短语“在嫌疑人家中”的处理过程如表 1 所示：

表 1 SNP 分词融合示意表

类目	内容			
初始短语	在嫌疑人家中			
词性标注	在/PREP	嫌疑/COM-NOUN	人/SUF-AF	家中/COM-NOUN
序列标注	O	B	I	O
分词融合	在/PREP	嫌疑人/COM-NOUN	家中/COM-NOUN	

介词短语融合应用在介词短语模板训练部分、介词短语识别模块。在介词短语模板训练部分，将训练语料内层介词短语融合，并将词性标注为 PP，训练外层介词短语识别模板；在介词短语识别模块，将测试语料内层介词短语识别后，若介词短语无嵌套情况识别后可进行去除，若有嵌套需将介词短语原语料中识别结果所对应的词语进行合并，并将合并后的介词短语词性标注为 PP，简化语料以适应外层介词短语识别。例如嵌套短语“本着对亲人、对家庭负责的态度”处理过程如表 2 所示：

表 2 内层介词短语分词融合示意表

类目	内容					
初始短语	本着对亲人负责的态度					
分词	本着	对	亲人	负责	的	态度
词性	<PREP>	<PREP>	<COM-NOUN>	<NVERB>	<DE-1>	<COM-NOUN>
序列标注	0	B	E	0	0	0
PP 融合	本着	**对亲人**	负责	的	态度	
融合词性	<PREP>	**PP**	<NVERB>	<DE-1>	<COM-NOUN>	

3 介词短语分层识别方法

具有嵌套并列结构的介词短语识别采用 CRF 模型，具体步骤如图 1 所示。

3.1 语料预处理

本文首先使用 CRF 模型对语料中的简单名词短语进行识别，由于简单名词短语选取特征与介词短语不同，因此要将语料形式进行更改，只需要留下词和词性;然后针对 PP 内部短语的特性制定规则库,并将结果进行校正；最后依据识别校正后的简单名词短语将初始语料中相应的词语进行分词融合，使语料更加简洁，适合介词短语识别。

3.1.1 特征抽取及特征模板

本文识别简单名次短语使用的特征为词特征（word）、词性特征（pos），选取特征窗口大小为 5，特征模板如表 3 所示，括号中的数字表示词的相对位置。

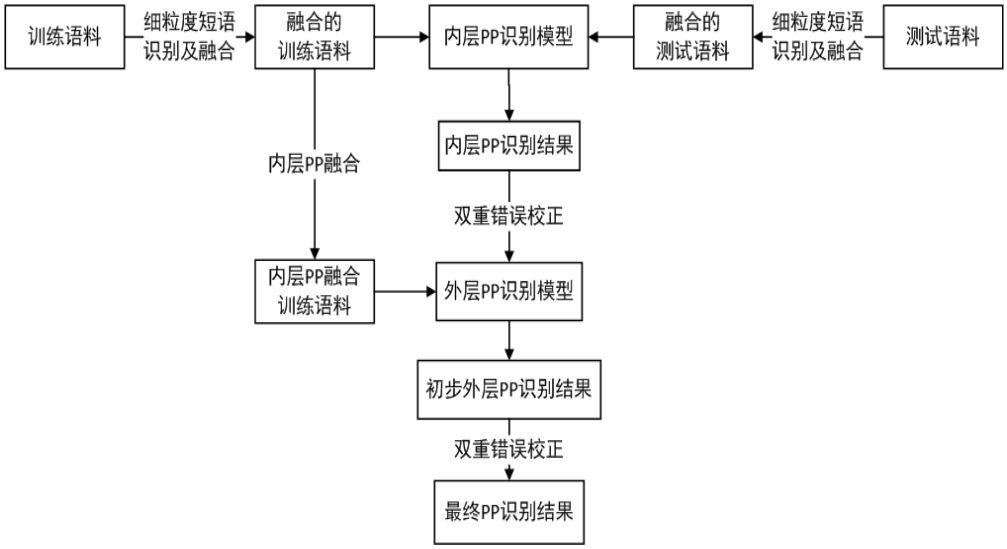


图 1 分层识别流程图

表 3 SNP 特征模板特征描述及特征表示

序号	特征描述	特征表示
1	当前词分别与窗口范围内任意两个相邻词性组合	word(0)pos(i)pos(i+1)
2	当前词、当前词性分别与窗口内词组合	word(0)pos(0)pos(i)
3	窗口范围内三个相邻的词性组合	pos(i)pos(i+1)pos(i+2)

3.1.2 规则库

依据介词短语内短语的特性制定规则库，修正简单名词短语识别结果，使其更适宜介词短语识别，部分规则如下：

1) 若前词为程度副词，该程度副词修饰名词短语的第一个词，且第一个词为形容词时，则将程度副词合并到名词短语中。

如:“高层次”，“高”的前词为副词“更”，合并“更”到短语内得到“更高层次”。

2) 若名词短语后界为“全部”等副词，则名词短语的后界为副词的前词。

3) 当名词短语前词为“沿”、“依”时，若组成名词短语的前两个词为名词，且名词短语由三个或三个以上的词构成时，则其前界为名词的后词，否则，标记不是名词短语。

4) 若后界为“你”等人称代词，将人称代词的前词标记为简单名词短语的后界。

3.2 介词短语分层识别

3.2.1 多模型训练

由于介词短语内部结构复杂，上下文联系密切，特征的选择对介词短语的识别效果有着重要影响。本文结合其他文献特征的选择，最终决定采用 6 个基本特征，具体如下：

（1）词特征(word)

（2）词性特征即词性标注(pos)

（3）候选介词前界特征(CFB)：当前分句中该词之前是否存在候选介词

（4）候选介词后界特征(CLB)：当前词是否可以作为介词短语后界。使用公式计算当前词作可以为后界的概率，阈值设置为 0.05

后界概率=当前词作为后界出现的次数/对应介词出现的总次数

（5）候选介词后词特征(CLW)：当前词是否可以作为介词短语后面的词。利用公式计算当前词可以作为后词的概率，阈值设置为 0.05.

（6）词长特征（CL）

本文使用原子特征模板和复合特征模板，选择特征窗口大小为 5 进行实验。通过基本特

征构成的集合作为 CRF 模型的原子特征模板，如表 4 所示；复合特征模板侧重特征间的搭配关系，提高了介词短语识别的精度，复合模板如表 5 所示，其中括号中的数字表示词的位置。

表 4　原子特征模板特征描述及特征表示

序号	特征描述	特征表示
1	词特征	word(-2),word(-1),word(0),word(1),word(1)
2	词性特征	pos(-2),pos(-1),os(0),pos(1),pos(2)
3	候选前界特征	CFB(-2),CFB(-1),CFB(0),CFB(1),CFB(2)
4	候选后界特征	CLB(-2),CLB(-1),CLB(0),CLB(1),CLB(2)
5	候选后词特征	CLW(-2),CLW(-1),CLW(0),CLW(1),CLW(2)
6	词长	CL(-2),CL(-1),CL(0),CL(1),CL(2)

表 5　复合特征模板特征描述及特征表示

序号	特征描述	特征表示
1	当前词与其词性组合	word(0)pos(0)
2	词、词性分别与候选前届的组合	word(i)pos(i)CFB(i)
3	当前词的候选后词与其前词的候选前后界组合	CLB(-1)CLW(0)CFB(-1)
4	当前词词性、候选前届与其前后词的候选组合	pos(i)pos(0)CFB(0)
5	当前词词性、候选后词与其前词的候选前后界组合	CLB(-1)pos(0)CFB(-1)CLW(0)

介词短语嵌套、并列现象的存在，使得介词短语识别难度加大，如句子“他们把对战士的爱、对边防的情一一送上哨卡”，包含并列结构“对战士的爱”“对边防的情”以及嵌套结构“把（对战士的爱）、（对边防的情）”。在文献[6]、文献[7]所采用的从右向左逐个介词短语识别的方法中，某个介词短语识别的错误会对其他介词短语的识别产生影响，如表 6 所示。逐个识别介词短语不能很好适用这种结构，本文将介词短语分层识别，从内层至外层逐层用 CRF 进行介词短语识别。另外，由于嵌套的内外层的上下文信息不同，本文提出，需要训练不同的模型对不同层的介词短语进行识别。

内层介词短语训练模板需将经过简单名词短语融合后的语料进行训练，外层介词短语训练模板需要将语料内层介词短语融合，并修改词性等相应的特征，重新训练生成。

表 6　从右向左逐个介词短语识别错误示例

次数	分类	内容									
第一次	分词	他们	本着	对	亲人	、	对	家庭	负责	的	态度
	识别	O	O	O	O	O	B	E	O	O	O
第二次	分词	他们	本着	对	亲人	、	负责	的	态度		
	识别	O	O	B	E	O	O	O	O		
第三次	分词	他们	本着	、	负责	的	态度				
	识别	O	B	I	I	I	E				

3.2.2 分层识别

分层识别过程如下：

首先，将测试语料处理成适合内层介词短语识别的形式，修改前界、后界、后词等特征，同时修改人工标注结果以方便比对，修改方式为：

1.若有多层嵌套的介词短语，则只标注最内层介词短语；

2.若只有一层介词短语，则标注该层介词短语；

3.去掉不含介词短语的句子，并用 CRF 识别内层介词短语。

如测试语料中的句子“他们本着对亲人、对家庭负责的态度”，经过分词、词性标注以及 SNP 识别融合后的结果是“他们/<PERSON_PRON> 本着/<PREP> 对/<PREP> 亲人/<COM-NOUN> 、/W 对/<PREP> 家庭/<COM-NOUN> 负责/<NVERB> 的/<DE-1> 态度/<COM-NOUN>”,测试语料的人工标注为“O B B E I B E I I E”，将语料处理成内层识别模式，人工标注序列更改成“O O B E O B E O O O”。

随后，使用 CRF 工具利用训练好的介词短语内层识别模型识别出介词短语，并根据双重错误校正系统进行校正。

然后，将识别校正后的内层介词短语进行融合并修改相关特征。如上例分词融合后结果为“他们/<PERSON_PRON> 本着/<PREP> **对亲人/PP** 、/W **对家庭/PP** 负责/<NVERB> 的/<DE-1> 态度/<COM-NOUN>”（加粗为融合修改部分），同时人工标注序列也修改为“O B I I I I I E”。

最后，利用训练好的适合外层的模型对外层介词短语进行识别并进行双重错误校正。

3.2.3 转换规则集

本文在序列标注后规则处理时使用的转换规则集由两部分构成：错误驱动学习（Transformation-based error-driven learning, TBL）和语义分析得到的固定搭配。

TBL 基本思想是通过错误驱动来修改识别结果，根据预先设计好的转换模板和目标函数寻找修正错误最多的转换规则，用生成的规则对标注结果进行修正，这部分规则由触发条件和转换规则组成。在进行结果校正时，若满足触发条件则进行修正。

例：句子“统统记在参加保险者的名下。”满足触发条件的介词为“在”且其前词词性是动词，若分句中存在“的”，则标记“的”后面的词为“E”，介词后的词到“的”标记为“I”，结果如表 7 所示。

表 7 转换规则集示例

类目	内容						
分词	统统	记	在	参加保险者	的	名下	。
词性标注	ADV	COM-VERB	PREP	COM-NOUN	DE-1	COM-NOUN	W
PP 识别	O	O	B	E	O	O	O
标注修正	O	O	B	I	I	E	O

固定搭配是通过对介词短语进行语义分析得到的，本文参考分析国内的语言学家们对介词及介词短语的研究成果，包括范晓[10]的《介宾短语·复指短语·固定短语》，张斌[11]的《现代汉语虚词》，陈昌来[12]的《汉语“介词框架”研究》等，总结出一系列适用于本文语料的固定搭配，如“对……来说”“当……时”。当进行结果校正时，若当前分句满足固定搭配，则修改其标注结果。

4 实验设置及结果分析

4.1 实验设置

本实验语料选用《人民日报》2000 年 1 月语料，包含 7037 个介词短语信息。该语料经过分词工具[13]进行分词及词性标注，并进行了人工校正。此外，需将训练语料格式化使其适合 CRF 训练，删除测试语料中不包含介词短语的句子，并对其同样进行格式化处理，再使用 CRF 工具进行序列标注。

实验方法方面，本文采取五倍交叉实验：将语料平均分成五份，每份介词短语数目如表 8 所示。使用其中一份语料作为测试语料，其他四份作为训练语料，重复进行五次实验，取平均值作为最终结果。

表 8 语料中介词短语数目统计

实验	语料 1	语料 2	语料 3	语料 4	语料 5
PP 总数	1414	1424	1405	1401	1393
嵌套 PP 数目	149	147	133	141	142

实验结果采用 CoNLL2000 评价标注，使用精确率（P）、召回率（R）和 F 值进行评价。精确率表示正确识别的介词短语所占识别出的介词短语百分比，反映了模型的识别能力。召回率表示正确识别的介词短语占语料中所有介词短语的百分比，反映了模型的查全能力。F 值综合表征了精确率和召回率，体现了算法综合性能。P、R、F 值的公式如下：

$$精确率\ P=\frac{Nc}{Ni}\times 100\% \qquad (1)$$

$$召回率\ R=\frac{Nc}{Ny}\times 100\% \qquad (2)$$

$$F=\frac{2\times P\times R}{P+R}\times 100\% \qquad (3)$$

其中，N_c 代表正确识别的介词短语数，N_i 代表识别出的介词短语数，N_y 代表语料中的介词短语总数。

4.2 实验结果及分析

本文进行了 5 个对比实验，实验 1 是融入简单名词短语的介词短语识别结果；实验 2 是在实验 1 的基础上对多层嵌套分层识别改进的结果，将从右向左逐个识别改为逐层识别；实验 3 对实验 2 进行了改进，对分层嵌套结构识别结果进行融合并对特征进行更新；实验 4 在实验 3 的基础上，外层介词短语识别时重新训练新的模型，识别后再加规则处理的结果。实验 5 是在实验 4 基础上，将外层介词短语识别完后更换规则得到的结果。

表 9 实验结果统计

序号	模型	P (%)	R (%)	F (%)
1	融合 SNP（baseline）	93.02	92.95	92.98
2	分层识别	93.34	93.28	93.31
3	分层识别+识别融合特征修改	93.81	93.74	93.77
4	分层识别+识别融合特征修改+多模型	94.03	93.95	93.99
5	分层识别+识别融合特征修改+多模型+不同规则	94.11	94.02	94.06

实验结果如表 9 所示：实验 2 的 P、R、F 值相比实验 1 分别提高了 0.32%，0.33%，0.33%，说明在分层嵌套的情况下，以层为单位能够降低某个介词短语识别错误对其他介词短语造成的影响，这种方式更适合介词短语的识别；实验 3 每层识别后，不再去掉已经识别出来的介词短语而是将其合并并修改标注信息，P、R、F 值比实验 2 分别提高了 0.47%，0.46%，0.46%，说明单纯去掉识别出来的介词短语会影响介词短语的上下文信息，可能会导致接下来的介词短语识别错误；实验 4 在不同层介词短语识别时采用不同的训练模型进行识别，P、R、F 值比实验 3 分别提高了 0.22%，0.21%，0.22%，说明不同层次的介词短语上下文信息也会不同，同一个训练模型不能很好地处理嵌套结构。实验 5 相比实验 4 的 P、R、F 值分别提高了 0.08%，0.07%,0.07%，说明不同层的介词短语由于结构不同所使用的校正规则信息也应不同。

表 10 给出了本文方法和融合 SNP 方法（baseline）对嵌套并列结构介词短语识别的改进效果对比：

表 10　嵌套并列结构介词短语总数及识别错误数目

	语料 1	语料 2	语料 3	语料 4	语料 5
嵌套并列结构 PP 总数	149	147	133	141	142
文献[7]识别错误数	29	47	33	45	32
本文错误数	17	17	20	24	22

为了进一步说明本文方法的有效性，本文在同一语料上重现了相关的研究方法，表 11 为本文方法与其它方法的实验对比：

表 11　与其它方法的结果比较

模型	P (%)	R (%)	F (%)
三元模型（文献[2]）	87.48	87.27	87.37
HMM（文献[4]）	86.50	85.40	85.64
最大熵模型（文献[3]）	89.52	88.93	88.22
多层 CRF（文献[6]）	91.98	91.92	91.95
融合 SNP（文献[7]）	93.02	92.95	92.98
多模型融合（本文）	94.11	94.02	94.06

由表 11 的实验结果可见，与其他模型相比，CRF 模型能够较好地利用上下文信息，并通过特征的重要性对其加权，识别结果精度较高；文献[7]的结果说明融入简单名词短语能够降低句子的复杂程度，提高识别精度；文献[6、7]采用的从右向左逐个介词短语的识别方法，某个介词短语识别错误会对接下来要识别的介词短语产生影响，本文对识别方法的改进，降低了复杂嵌套、并列结构介词短语的识别难度，不同层次采用不同的训练模型进行识别，能够更好地获得嵌套介词短语的特征信息，提高识别效果。

5 总结及展望

中文介词短语中，介词短语嵌套和并列现象是影响介词短语识别性能的重要问题之一。为此，本文提出了多模型融合的中文介词短语识别方法。实验结果表明：

（1）介词短语具有嵌套、并列的复杂结构，从右向左识别介词短语的方法某个介词短语的识别错误会影响到后续的介词短语识别。分层识别方法不是每次只识别一个，而是将同一层次的介词短语同时进行识别，更适合存在嵌套、并列的介词短语。

（2）内外层介词短语结构不同，上下文信息也不同，需要不同的训练模型进行识别，训练语料需要将标记的内层介词短语融合并进行特征修改后训练嵌套模型以适应外层介词短语识别。

（3）识别出的内层介词短语不再进行去除，而是根据识别结果将测试语料中相应的词语进行分词融合，并将相应特征进行修改，以适应外层介词短语识别。在外层识别时，重新训练新的模型使之适合当前层的识别，提高识别效果。

在语料处理过程中，简单名词短语的识别错误可能会将介词短语的后界与后词合并在一起，导致识别介词短语错误。如：句子“加大 对 大要案件 侦办力度”经过简单名词短语识别融合后结果为“加大　对 大要案件侦办力度”介词短语识别结果为“对大要案件侦办力度”，而正确结果为“对大要案件”。因此后续的研究要改善简单名词短语的识别方法，使简单名词短语的粒度细化，提高精确率和召回率。

参考文献

[1] 吴云芳. 现代汉语介词结构的自动标注[D]. 北京：北京语言文化大学, 1998.

[2] 干俊伟, 黄德根. 汉语介词短语的自动识别[J]. 中文信息学报, 2005, 19(4):17-23.

[3] 卢朝华, 黄广君, 郭志兵. 基于最大熵的汉语介词短语识别研究[J]. 通信技术, 2010, 43(5):181-183.

[4] 奚建清, 罗强. 基于HMM的汉语介词短语自动识别研究[J]. 计算机工程, 2007, 33(3):172-173.

[5] 胡思磊. 基于CRF模型的汉语介词短语识别[D]. 大连：大连理工大学, 2008.

[6] 张杰. 基于多层CRFs的汉语介词短语识别研究[D]. 大连：大连理工大学, 2013.

[7] 桑乐园, 黄德根. 基于简单名词短语的汉语介词短语识别研究[J]. 中文信息学报, 2015, 29(6):8-12.

[8] 孙玉祥. 汉语简单名词短语自动识别的研究[D]. 大连：大连理工大学, 2014.

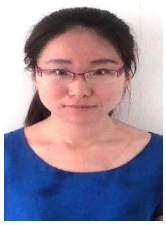
[9] Lafferty, John D, McCallum, et al. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data[M]// Departmental Papers (CIS). 2001.

[10] 范晓. 介宾短语·复指短语·固定短语[M]. 北京：人民教育出版社, 1990.

[11] 张斌. 现代汉语虚词[M]. 上海：华东师范大学出版社, 2000.

[12] 陈昌来. 汉语“介词框架”研究[M]. 北京：商务印书馆, 2014.

[13] Degen H, Deqin T. Context information and fragments based cross-domain word segmentation[J]. China Communications, 2012, 9(3):49-57.



刘彤（1993—），硕士研究生，主要研究领域为自然语言处理。

E-mail：liutongdl@mail.dlut.edu.cn

黄德根（1965—），通讯作者，教授，博士生导师，主要研究领域为自然语言处理、机器翻译、人工智能等。

E-mail：huangdg@dlut.edu.cn

张聪（1993—），硕士研究生，主要研究领域为自然语言处理。

E-mail：cccaaag@mail.dlut.edu.cn