基于简单名词短语的汉语介词短语识别研究 *

桑乐园，黄德根

（大连理工大学电信学部计算机学院，辽宁省 大连市 116024）

摘要：提出一种融入简单名词短语信息的介词短语识别方法。该方法首先使用 CRF 模型识别语料中的简单名词短语，并使用转换规则对识别结果进行校正，使其更符合介词短语的内部短语形式；然后依据简单名词短语识别结果对语料进行分词融合；最后，通过多层 CRFs 模型对测试语料进行介词短语识别，并使用规则进行校正。介词短语识别的精确率、召回率及 F-值分别为：93.02%、92.95%、92.99%，比目前发表的最好结果高 1.03 个百分点。该实验结果表明，基于简单名词短语的介词短语识别算法的有效性。

关键词：简单名词短语识别；CRF；分词融合

中图分类号：TP391 **文献标志码**：A

The Chinese preposition phrase recognition based on simple noun phrase

SANG Leyuan, HUANG Degen

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

Abstract: This paper proposes a new approach integrating simple noun phrase information into preposition phrase recognition. We recognize simple noun phrases through basic CRF model, and filter the phrases with conversion rules in order to adapt to the inner phrase patterns in the preposition phrases. Then we utilize the simple noun phrases to merge fragmental participles into a complete phrase in our corpus. Finally, we recognize the preposition phrases through multilayer CRFs, and use rules to correct the result. The optimized model performs 1.03 point higher than the current best model with precision, recall and F-value being 93.02%, 92.95%, 92.99%, indicating the affectivity of the recognition algorithm of preposition phrase based simple noun information.

Keyword: Simple noun phrase recognition；CRF；Participle fusion

1 引言

介词短语(Preposition Phrase，PP)是汉语中一种重要的短语类型，对句法分析、机器翻译、信息检索有着重要作用。介词[1]起标记作用，与名词、名词短语或其他词语构成 PP，充当状语、宾语、补语等成分，用于补充谓语或说明宾语。PP 的正确识别能够大大降低句法分析的难度，提高机器翻译的性能，对信息检索及文本分类效果都有较大的提升。因此，PP 识别作为自然语言处理的一部分，具有重要的意义。

国内外学者针对 PP 的自动定界问题展开了各种探索和研究。在英语方面的代表性方法包括：基于规则的转换算法[2]，启发式无监督的统计算法[3]，基于句法分析及语义分析的消歧算法[4]等，这些方法针对英语 PP 的构词规则，应用到汉语 PP 识别上效果较差。由于汉语 PP 内部结构复杂且定界不明，目前识别结果的 F-值大都在 90%左右。汉语 PP 识别的方法[5-8]集中在浅层句法分析上，即在分词及词性标注后，用一个模型将 PP 作为一个整体识别出来。干俊伟等人[5]提出了基于三元统计模型的方法，首先利用搭配模板获取可信搭

* **收稿日期：** **定稿日期：**

基金项目：国家自然科学基金(NO.61173100, NO.61173101, NO.61272375)，2013 教育部人文社会科学研究规划基金项目（NO.13YJAZH062）

配关系，依据可信搭配关系识别 PP，然后利用三元统计模型与规则相结合的方法识别可信搭配关系未识别出来的 PP，文中的三元统计模型中只考虑了介词、后界的词性及后词的词性三个特征，考虑的特征少，其 F-值仅为 87.37%；奚建清等人[6]提出了基于 HMM 模型的 PP 识别方法，并应用依存语法进行错误校正，由于 PP 内部结构比较复杂，利用简单特征函数无法涵盖其所有特性，而 HMM 模型无法使用复杂特征，其 F-值仅为 85.67%；卢朝华等人[7]提出了基于最大熵模型的 PP 识别方法，并采用基于依存语法的错误界定方法对识别结果进行校正，由于最大熵模型不能统计特征的强度，并且数据稀疏问题严重，其 F-值为 88.22%；张杰[8]提出了基于多层 CRFs 的 PP 识别方法，并采用基于转换的错误驱动学习方法对识别结果进行校正，识别 F-值达到 91.95%，是目前发表的识别结果最好的方法，但文中对 PP 的分析局限在词上，没有考虑 PP 的内部成分特点，仍有提升空间。PP 是由介词与其他实体短语一起构成，若先对语料进行实体短语识别，可以简化 PP 的内部结构，从而降低 PP 识别的复杂性。考虑到 PP 中介词后面的短语大多是由名词短语构成，本文提出基于名词短语识别的 PP 识别方法。

汉语名词短语识别[9-12]分为基于规则的方法和基于统计的方法两种方法。Cardie 等人[9]提出了一种基于基本名词短语(Base Noun Phrase，BNP)的词性串的规则剪枝方法；钱小飞等人[11]则提出了一种基于 CRF 模型的最长名词短语(Maximal Noun Phrase，MNP)识别方法，并制定了基于边界信息和内部结构信息的规则库对识别结果进行校正；孙玉祥[12]提出了基于 CRF 模型的简单名词短语（Simple Noun Phrase，SNP）的识别方法，并利用基于语义分析的规则库对识别结果进行校正。BNP 简单易识别，但易将作为整体结构的短语割裂细化，形成粒度过小的短语结构，失去了在 PP 识别中加入名词短语识别的意义；而 MNP 识别粒度大，有利于句子整体结构分析，但却合并了一些 PP 到 MNP 内，反而使识别 PP 的效果降低；SNP 是指内部不包含复杂修饰成分的名词短语，其复杂程度介于 BNP 和 MNP 之间，既能保留充分的语法信息，又能够减少歧义问题，进而提高 PP 识别的精度和效率，因此本文采用融入 SNP 信息到 PP 识别方法中对其进行优化。

综上，本文提出一种基于 SNP 的 PP 识别方法，即通过分词融合将 SNP 信息融入到语料中，并对其训练得到多层 PP 识别模型，再使用该模型识别测试语料中的 PP，最后使用规则校正其识别结果，得到最终识别结果。

2 CRF 模型及分词融合

本文把 SNP 识别问题及 PP 识别问题视为序列标注问题，即通过 CRF 模型对测试语料进行序列标注，识别出 SNP 及 PP。首先，把语料进行分词及词性标注，即把句子处理为 S=word(1)/pos(1) word(2)/pos(2)…word(*i*)/pos(*i*)…word(n)/pos(n)格式(其中 word(*i*)为句子中的第 *i* 个词，pos(*i*)为第 *i* 个词的词性，*n* 为句子 *S* 中含有词的个数)。目标为获得一个对应的标注序列 T = T(1) T(2)…T(n)，使得该序列在所有可能的标注序列中概率最大，其中 SNP 识别过程中 T(*i*)∈{B，I，O}，B 表示 SNP 的起始词，I 表示 SNP 的内部词语，O 表示 SNP 的外部词语，而 PP 识别过程中 T(*i*)∈{B，I，E，O}，B 表示 PP 的起始边界，I 表示 PP 的内部词语，E 表示 PP 的后边界，O 表示 PP 的外部词语。

2.1 条件随机场

通过CRF机器学习模型[13]能够充分地利用词语的上下文信息特征，使用无向图理论使序列标注的结果达到整个序列上的全局最优，适用于词性标注及浅层句法分析任务。

本文使用线性链CRF，即给定参数 $\Lambda=(\lambda_1,\lambda_2,...,\lambda_n)$，定义在观测序列 $X=x_1,x_2,...,x_T$ 上对应的状态序列 $Y=y_1,y_2,...,y_T$ 的条件概率为：

$$P_\Lambda(Y\mid X)=\frac{1}{Z_X}\exp\left(\sum_{t=1}^{T}\sum_{k}\lambda_k f_k(y_{t-1},y_t,x,t)\right) \tag{1}$$

其中 Z_X 是所有状态序列的归一化因子，$f_k(y_{t-1},y_t,x,t)$ 为关于整个观测序列X、位置t以及位置$t-1$标记的二值特征向量函数，λ_k 是在训练中得到的 f_k 的权重，k的取值范围取决于模版中特征的数量。训练的目标是为CRF模型找到最优的λ值，找到后即可用Viterbi算法对未标记序列（测试语料）进行序列标注。序列标注的任务就是求出使条件概率 $P_\Lambda(Y\mid X)$ 最大的Y，即最大可能的标记序列为：

$$Y^*=\arg\max_{Y} P_\Lambda(Y\mid X) \tag{2}$$

2.2 分词融合

对语料进行 SNP 识别后，依据识别出的 SNP 对词语进行分词及词性标注合并，将融合后的 SNP 的词性标记为“<COM-NOUN>”。举例如下：

初始分词及词性标注为：给/PREP 自身/PERSON-PRON 和/CNJ 他人/PERSON-PRON 的/DE-1 生命/COM-NOUN 财产/COM-NOUN 安全/COM-NOUN 造成/NVERB 严重/ADJ 威胁/NVERB-N。/WJ

识别出来的 SNP 为：生命财产安全、严重威胁

分词融合后的分词及词性标注为：给/PREP 自身/PERSON-PRON 和/CNJ 他人/PERSON-PRON 的/DE-1 生命财产安全/COM-NOUN 造成/NVERB 严重威胁/COM-NOUN 。/WJ

3 介词短语识别方法

首先，使用 CRF 构建 SNP 识别模型，并使用该模型识别语料中的 SNP，使用规则库校正其识别结果得到 SNP 识别结果；之后，依据 SNP 识别结果对语料进行分词融合，采用 CRF 构建多层 PP 识别模型；最后，利用建立的 PP 识别模型识别 PP，并通过错误驱动方法及语义分析确定转换规则集，校正识别出的 PP，得到最终结果。

3.1 简单名词短语识别

本文使用 CRF 模型对语料进行 SNP 识别，并且针对 PP 内名词短语的特性制定了规则库进行结果校正。

3.1.1 特征抽取及特征模板

本文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1所示，其中括号中的数字表示词的位置，如word(-1)表示当前词的前词，word(0)表示当前词，word(1)表示当前词的后词。

表1 SNP识别特征模板

Tab.1 The feature template of SNP recognition

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), i∈[-2,1]
2	当前词、当前词性分别与窗口内的词组合	word(0)pos(0)word(i), i∈{-2,-1,1,2}
3	窗口范围内三个相邻的词性组合	pos(i)pos(i+1)pos(i+2), i∈[-2,0]

3.1.2 规则库

依据PP内名词短语的特性制定以下规则，该规则能较好地校正SNP识别结果，并在最大程度上不合并PP的后界和后词，明显提升PP的识别效果。

1) 若识别出的SNP的前词为程度副词时，该程度副词修饰SNP的第一个词，且第一个词为形容词，则将程度副词合并到SNP中。如识别出的SNP为“好结果”，好的前词为副词“更”，合并“更”到短语内，则SNP为“更好结果”。
2) 短语内部包含并列成分，采用语义相似度和词语组合数据库方法进行并列消歧，分为三种情况，如表2所示：

表2 并列歧义的三种情形

Tab.2 Three kinds of parallel ambiguity

并列类型	举例	正确切分	可能的错误切分
修饰词并列	物质和精神财富	[物质和精神财富]	物质和[精神财富]
中心词并列	各项工作和政策	[各项工作和政策]	[各项工作]和政策
整体并列	各族人民和海外侨胞	[各族人民]和[海外侨胞]	[各族人民和海外侨胞]

3) 若SNP后界为机构名时，则SNP的后界为其前词；
4) 若SNP的后界为“全部”、“全程”等副词，则SNP的后界为副词的前词；
5) 当SNP的前词为介词“沿”、“依”时，若组成SNP的前两个词为名词，且SNP由3个或3个以上词构成时，则其前界为名词的后词，否则，标记不是SNP；
6) 若SNP的后界为“每个”等指示性代词，则SNP的后界为其前词。

3.2 介词短语的识别

3.2.1 特征抽取及特征模板

本文使用张杰[8]的特征抽取方式，使用原子特征模板和复合特征模板，选择特征窗口的大小为5。原子特征模板即基本特征，选择以下基本特征：

1) 词特征(word)；
2) 词性特征(pos)：即词性标注；
3) 候选前界特征(CFB)：即当前分句中该词前是否存在候选介词。若存在候选介词，则标记为该介词，若不存在，则标记为“N”；
4) 候选后界特征(CLB)：即当前词是否可以作为介词短语的后界。使用公式（3）计算当前词作为后界的概率，本文选择阈值为0.05，即若概率大于0.05，则标记该特征为“Y”，否则标为“N”；

$$后界的概率=\frac{当前词作为后界的次数}{对应介词出现的总次数} \tag{3}$$

5) 候选后词特征(CLW)：即当前词是否可以作为介词短语后面的词。利用公式（4）计算当前词作为后词的概率，本文选择的阈值为0.05，即若概率大于0.05，则标记该特征为“Y”，否则标为“N”；

$$后词的概率=\frac{当前词作为后词的次数}{对应介词出现的总次数} \tag{4}$$

6) 词长特征：即当前词的长度。

复合模板侧重特征间的搭配关系，提高了介词短语识别的精度。复合特征模板如表 3 所示，其中括号中的数字表示词的位置，如 word(0)表示当前词。

表3 PP识别特征模板

Tab.3 The feature template of PP recognition

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、词性分别与候选前界的组合	word(*i*)pos(*i*)CFB(*i*)，$i\in\{-1,0,1\}$
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、候选前界与其前、后词的词性组合	pos(*i*)pos(0)CFB(0)，$i\in\{-1,1\}$
5	当前词词性、候选后词及前词的候选前、后界组合	CLB(-1)pos(0)CFB(-1)CLW(0)

3.2.2 转换规则集

该转换规则集由两部分构成，一部分是通过错误驱动学习（Transformation-based error-driven learning，TBL）自动获取；另一部分是通过语义分析得到的固定搭配[1]。TBL 的基本思想是通过错误驱动来修改标记结果，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。重复上述过程，直到无新规则产生。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，用相应的转换规则对当前结果进行修改。例如，若分句为“统统/<ADV> 记/<COM-VERB> 在/<PREP> 参加保险者/<COM-NOUN> 的/<DE-1> 名下/<COM-NOUN>”，其标注结果为“O O B E O O”，满足触发条件介词为“在”且其前面是动词，若分句中存在“的”，则标记“的”后的词为“E”，介词后的词到“的”标记为“I”（转换条件），因此修改标注结果为“O O B I I E”。固定搭配是通过对 PP 进行语义分析得到的，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。例如，若一个分句满足“对……来说”规则，则将 “对”和“来说”两词中间的词标注结果改为“I”， “来说”的标注结果改为“E”，“来说”后词的标注结果改为“O”。

4 实验设计及结果分析

本文的实验语料是人民日报2000年的语料，该语料经过NIHAO分词工具[14]进行分词及词性标注，为保证实验结果的准确性，进行了人工校正。训练语料需格式化使其适合CRF训练，而测试语料需删除不包含PP的句子后再进行格式化，然后使用CRF进行序列标注。所有语料共包含7049个PP。本文将语料平均分成五份，即语料1，语料2，语料3，语料4，语料5。实验采用五倍交叉验证，即用其中四份作为训练语料，另一份作为测试语料，进行五次实验。本文将五次实验结果的平均值作为最后的识别结果。

4.1 实验设计

本文针对PP识别进行了4个对比实验：实验1是直接使用PP识别模型对测试语料进行PP识别得到的实验结果；实验2是首先对测试语料进行SNP识别，分词融合后使用PP识别模型对测试语料进行PP识别得到的实验结果；实验3是对实验1的实验结果利用规则库处理后得到的实验结果；实验4是对实验2的实验结果进行规则处理后得到的实验结果。

4.2 实验结果及分析

实验结果如表2所示：实验2的精确率、召回率及F值比实验1分别提高了0.57%、0.56%、

0.56%，说明加入简单名词短语识别后的PP识别的效果有了明显的提高；加入规则后，实验3和实验4的精确率、召回率、F-值分别提高了0.53个百分点和1.28个百分点，说明规则库对识别效果是有明显的提升作用，实验4比实验3的F-值多提升0.75个百分点，说明规则库更适合SNP识别后的PP识别。

表4 PP识别结果

Tab.4 The result of the PP recognition

	精确率（%）	召回率（%）	F值（%）
实验 1	91.17	91.11	91.14
实验 2	91.74	91.67	91.70
实验 3	91.70	91.64	91.67
实验 4	93.02	92.95	92.99

参照表5中各个参考文献PP的识别结果可知，HMM模型识别PP的效果最差，这是由于介词短语内部结构比较复杂，使用简单特征函数不能涵盖其关键特性，致使识别效果最差；三元模型只考虑三个基本特征，忽略了其他比较重要的特征，如后词、后界，致使识别结果的F-值仅为87.37%；最大熵模型不忽略PP的任意特征，使其识别效果高于前两个文献的识别结果，但最大熵模型不能统计特征强度，降低了部分重要特征的权重，使其实验结果仍差强人意；CRF模型能够较好的利用上下文信息，并且通过特征的重要性对其加权，使识别结果较好；本文通过对PP内部结构进行分析，把SNP信息融入到PP识别方法中，降低了PP内部的复杂结构，提高了识别的精度和效率，精确率、召回率及F-值分别比文献[8]方法高1.04%、1.03%、1.04%，说明该方法的有效性。

表5 与其他文献的结果对比

Tab.5 Comparison with other works

实验方法	精确率（%）	召回率（%）	F值（%）
文献[5]方法(三元模型)	87.48	87.27	87.37
文献[6]方法(HMM)	86.50	85.40	85.64
文献[7]方法(最大熵模型)	89.52	88.93	88.22
文献[8]方法(多层 CRFs)	91.98	91.92	91.95
本文实验(融入 SNP 信息)	93.02	92.95	92.99

5 总结及展望

本文提出了融合简单名词短语信息的介词短语自动识别方法，首先抽取语料中的简单名词短语；之后将简单名词短语融合为单一的名词，并标注其词性为普通名词；最后通过多层CRFs模型识别介词短语。该方法通过降低介词短语内部结构的复杂性，提高了识别结果，其F-值为92.99%。实验结果表明，本文方法比目前发表的最好的实验结果高1.03个百分点，验证了简单名词短语信息在介词短语识别中的重要性。接下来我们将加入简单名词短语内部的词性等细粒度信息，并且寻找更优的规则对简单名词短语识别结果进行校正，以进一步提高介词短语识别的性能。

参考文献

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

[2] Brill E，Resnik P．A rule-based approach to prepositional phrase attachment disambiguation[C]// Proceedings of the 15th conference on Computational linguistics-Volume 2．Associationfor Computational Linguistics，1994：1198-1204．

[3] Ratnaparkhi A．Statistical models for unsupervised prepositional phrase attachment[C]//Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 2．Association for Computational Linguistics，1998：1079-1085．

[4] Branigan H P，Pickering M J，McLean J F．Priming prepositional-phrase attachment during comprehension[J]．Journal of Experimental Psychology：Learning, Memory, and Cognition，2005，31(3)：468-481．

[5] 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．

[6] 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，2008，33（3）：172-173+182．

[7] 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(05)：181-183+186．

[8] 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．

[9] Cardie C，Pierce D．Error-driven pruning of treebank grammars for base noun phrase identification [C] //Proceedings of the 17th international conference on Computational linguistics-Volume 1．Association for Computational Linguistics，1998：218-224．

[10] 胡乃全，朱巧明，周国栋．混合的汉语基本名词短语识别方法[J]．计算机工程，2009，35(20)：199-201．

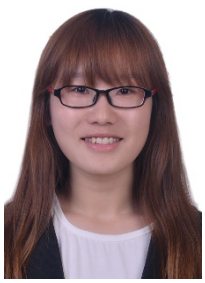
[11] 钱小飞，侯敏．基于混合策略的汉语最长名词短语识别[J]．中文信息学报，2013，27(6)：16-22．

[12] 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．

[13] Lafferty J，McCallum A，Pereira F C N．Conditional random fields：Probabilistic models for segmenting and labeling sequence data[C] //Proceedings of the 18th International Conference on Machine Learning 2001 (ICML 2001)：282-289．

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

作者简介：桑乐园（1991——），女，硕士研究生，主要研究领域为自然语言处理。
Email:Sangleyuan@mail.dlut.edu.cn；**通讯作者：黄德根**（1965——），男，教授，主要研究领域为自然语言处理、机器翻译。Email:huangdg@dlut.edu.cn。



桑乐园

黄德根