

文章编号: 1003-0077(2017)02-0001-10

汉语介词短语自动识别研究综述

李洪政, 晋耀红

(北京师范大学 中文信息处理研究所, 北京 100875)

摘要: 作为一种重要的短语类型, 介词短语在汉语中分布广泛, 正确识别汉语介词短语对自然语言处理领域的很多任务和应用都有重要的作用和意义。该文对近些年与识别汉语介词短语有关的研究做了梳理, 从研究对象、实验评价标准和具体研究方法等几个方面比较详细地介绍了相关工作, 最后归纳了汉语介词短语识别研究中表现出来的一些特点, 并对未来研究的发展提出了几点建议。

关键词: 介词短语; 识别; 规则; 统计

中图分类号: TP391

文献标识码: A

A Survey on Automatic Identification of Chinese Prepositional Phrases

LI Hongzheng, JIN Yaohong

(Institute of Chinese Information Processing, Beijing Normal University, Beijing 100875, China)

Abstract: As an important type of phrase, prepositional phrases (PP) are widely distributed in Chinese, Therefore proper identification of PPs has positive and important impacts on the various tasks and applications in the field of Natural Language Processing. This paper surveys related studies in identifying Chinese PPs in recent years, and discusses the works in detail from several perspectives: research objects, experimental evaluation and research methods. It finally concludes several features of research on Chinese PP identification and proposes several suggestions on the future work.

Key words: prepositional phrases; identification; rule; statistic

1 引言

长期以来, 介词短语问题一直是自然语言处理的难点问题之一, 并引起国内外的广泛关注和深入研究, 众多相关研究主要集中在汉语和英语上。但由于两种语言结构的差异性, 该研究在这两种语言中也有着很大不同。在英语中, 介词短语一般出现在句子末端, 这就很容易造成一种句法结构上的歧义, 即该介词短语是修饰前面的名词还是修饰动词。所以, 英语中介词短语研究非常重视解决介词短语的附加问题 (Prepositional Phrase Attachment)。国外很早就有人关注这个问题, 从最初的基于规则的方法, 到统计方法, 再到无监督和有监督学习, 以及目前十分流行的词汇向量表示等众多策略都被相

继提出。直到现在, 仍然不断有研究尝试不同的方法。与此密切相关的另外一个重点问题是介词词义消歧问题 (Preposition Sense Disambiguation)。这两个问题一个是句法歧义, 一个是语义歧义, 二者自然联系紧密。在自然语言处理领域的很多会议和期刊上, 经常会出现相关的论文。由于本文重点在汉语介词短语, 英语的介词问题在此不再赘述。

不同于英语, 汉语介词短语的研究重点则是把经过分词和词性标注的汉语文本中的介词短语作为一个整体识别出来, 一般属于浅层句法分析, 同时也是组块处理 (chunking) 的研究范围。假设在汉语句子 $S = W_1, W_2 \cdots W_n$ 中, 字符串 $W_i, W_{i+1} \cdots W_j$ 为待识别的介词短语, 那么 W_i 即为介词, 介词短语识别的主要任务就是将 W_i 和 W_j 分别识别为介词短语的前后边界, 并将整个字符串识别出来。进一步

讲,由于介词短语的左边界是介词本身,容易识别,所以识别的关键问题在于确定后边界的位置。

由于汉语自身的特点,自动识别汉语介词短语通常具有以下难点:

(1) 介词短语的内部构成相当复杂。介词短语由介词与其他语言成分构成。这些成分可以是简单的词语(名词、代词等),也可以是各种短语(动宾短语、名词短语、方位短语、时间短语等),甚至还可以是一个子句形式。复杂的内部结构很容易形成远距离的搭配关系。

(2) 兼类介词的存在。汉语介词可以兼做名词、量词、形容词、连词和动词等,有时候必须通过上下文语境才能判断具体词性,这给介词短语的识别带来了不小的困难。

(3) 在同一个句子中经常会出现多个并列的介词短语或者复杂的嵌套介词短语,即一个大的介词短语内部又包含其他的介词短语结构。这更增加了确定短语边界的难度。

(4) 部分介词短语本身存在歧义。类似于“对他的意见”这样的歧义结构在很多文献中已有研究。在有些情况下,仅仅根据句子的内部信息无法识别出介词短语,必须利用上下文信息才能将包含结构相同的词组的介词短语正确识别出来。

尽管存在以上诸多困难,但应该注意到,正确识别汉语介词短语对自然语言处理的诸多任务和应用都有十分积极和重要的影响。多年来国内的学者们对这个问题进行了积极的探索。在目前已有的文献中,最早面向自然语言处理领域进行介词短语自动识别研究的应该是吴云芳的硕士论文《现代汉语介词结构的自动标注》^[1],而后出现了更多相关研究,探索了许多有意义的识别方法,并产生了积极的成果与影响。本文将从该研究涉及的对象,实验及评价标准以及具体的研究方法等几个方面对介词短语识别研究进行比较详细的回顾与梳理。在这些研究中,既有专门只针对介词短语的,也有包括介词短语在内的不同类型的汉语短语的识别研究。

2 介词组块与介词短语

关于组块(chunk),Abney^[2]最早提出了一个完整的英语组块描述体系,对组块有着权威性的定义。他把组块定义为句子中一组相邻的属于同一个S-投射的词语的集合,建立了组块与管辖约束理论的X-bar系统的内在联系^[3]。认为组块是从句内的

一个非递归的核心成分。这种成分包含中心成分的前置修饰成分,而不包含后置附属结构。组块是严格按语法定义的,而不是在语义、功能或词法上定义的。Abney 还认为组块不一定能覆盖整个句子。

CoNLL-2000 会议首先把组块分析作为共享任务^[4],在 Abney 的基础上认为英文组块由一些短语构成,而每一个短语内是由句法相关的词构成,这些短语彼此不重叠、无交集,不含嵌套关系。并将组块分为包括 NP、VP、PP 等 11 种基本类型,其中介词组块只包括介词本身,而不包括介词后面的成分(如 NP)。

然而与英文不同,汉语组块至今并没有一个统一的定义。正如有研究提到的,很多研究者们根据自己的研究目的各自提出了不同的汉语组块描述体系,由此产生了数量不等的组块类别^[5-8]。但可以发现,无论如何定义,都坚持了一个原则,即认为汉语组块是非递归、不重叠、不嵌套的短语。具体对于介词组块,有的文献认为介词组块只包括介词本身^[9-10],有的虽然也认同介词组块只包含介词本身,但是对于一些有固定搭配的介词短语如“在……中”,也被划分为一个介词组块,但这种介词组块的长度一般不能超过 3 个词^[11-12]。还有的研究在提出的汉语组块类型中并未包括介词组块^[13-14]。

相对于汉语介词组块,对于介词短语的界定则具有普遍共识,即由介词与后面其他语法成分构成的短语,这些语法成分可以小至一个词语,大至从句形式。有时候也把这种界定称为基本介词短语。介词短语通常具有更复杂的句法结构,能够包含其他结构,同时允许嵌套结构的存在,这是与介词组块的较大区别。在目前搜集到的资料中,似乎只有文献^[15]认为介词短语只包括介词本身。这种界定其实是等同于了英语介词组块的定义。清华汉语树库^[16]定义的介词短语还可以包括基本介词短语前面的修饰成分(如副词)。但在介词短语识别的研究中,专指基本介词短语的识别,不会包括前面的修饰成分。我们认为,汉语介词短语,尤其是远距离搭配的介词短语的识别要远远难于介词组块的识别。

尽管介词组块和介词短语都与介词密切相关,而且一直引起国内外的广泛关注,但本文不打算讨论与介词组块研究有关的文献,而是将把重点放在介词短语的识别上,下文中论述的识别方法都是关于介词短语的。

3 介词短语识别实验指标

在介词短语研究涉及的实验中,通常采用正确率(P)、召回率(R)和 F₁ 值作为介词短语识别结果的评测指标。计算公式分别如下:

$$P = \frac{\text{正确识别的介词短语数量}}{\text{识别为介词短语的数量}} \times 100\% \quad (1)$$

$$R = \frac{\text{正确识别的介词短语数量}}{\text{语料中介词短语的总数}} \times 100\% \quad (2)$$

$$F_1 = \frac{2PR}{(P+R)} \times 100\% \quad (3)$$

在计算时,只有某个介词短语的前后边界完全识别正确,以及正确识别出短语的类型时,才认为完全识别正确。比如,假设一个介词短语的边界虽然识别正确,但被识别为其他类型的短语,那么这个短语也不是正确识别的结果。

4 介词短语识别方法

介词短语识别的方法主要包括规则方法、统计方法以及将二者相结合的混合方法。规则方法主要依赖于人工总结的语言学知识和规则,统计方法主要依靠统计和机器学习的模型来识别,这些方法是目前的主流。下面将分别介绍这些方法。

4.1 规则方法

郑州大学自然语言处理实验室一直致力于现代汉语虚词用法的研究。他们提出了“三位一体”的构建现代汉语介词知识库的思路^[17-18],包括:介词用法词典、介词用法规则库和介词标注语料库。从介词的实际用法入手,对介词用法进行形式化的规则描述,人工书写规则,构建了介词用法规则库。最后以人民日报作为真实语料库,对其进行人工标注工作。并不断根据现实语料、《现代汉语语法信息词典》《现代汉语虚词词典》和《现代汉语八百词》等对介词用法词典和规则库进行修改和完善。

梁猛杰等^[19]根据已有的工作,通过考察介词规则库的处理特点,依据规则的覆盖程度从低到高进行分类,重新调整了规则的前后排序方案,同时对排序的规则进行优选。实验选用 2000 年 3 月份人民日报作为测试语料,实验结果表明,通过调整排序方案,在保证时间复杂度较低的情况下,排序后的介词识别准确率较之未调整前有了大幅提高。

与处理人民日报等新闻语料不同,北京师范大

学中文信息处理研究所在概念层次网络理论(Hierarchical Network of Concepts, HNC)^[20]的指导下,面向汉语专利领域的文本,专门构建了较大规模的汉语专利语料知识库,在开展汉英专利机器翻译研究的过程中探索了介词短语识别方法和思想。Zhu^[21]、朱筠^[22]、胡韧奋^[23]对大规模汉语专利语料中介词短语的分布情况和语言学特征等进行了细致的考察。专利文本中的介词短语结构通常具有更多的字数,结构也更为复杂,甚至经常出现嵌套介词短语的情况。文献首先将介词分为引导主语义角色(与句中核心谓词直接关联的对象或状态,如施事、受事等)和辅语义角色(结构上可有可无的辅助性信息,如时间、地点、方式等)两类。前者典型的介词包括“把、由、将、被”等,后者主要包括“通过、除了、根据”等。然后分析了介词与介词短语右边界的搭配信息。例如:“在……中”、“当……时/时候”等固定的介词短语的边界组合。再次,根据句中核心动词的配价特点,指出了不同配价的动词可以与哪类介词一起使用。最后,将介词短语按照句法层次分为两类。一类是构成句子的直接组成成分,另一类则是短语(如 NP)内部的介词短语成分。

根据以上四类语义信息,分别为不同特点的介词人工设计书写了简洁有效、易于阅读而又具有较高覆盖范围的形式化语义规则,帮助系统识别相应的介词短语。规则对于字数更多、包含远距离搭配结构的介词短语的识别具有明显效果。

对专利语料中介词短语识别的封闭测试的准确率在 90%左右,开放测试中,两类介词的准确率分别在 88%和 94%,平均比基线系统高了 12~15 个百分点。另外,加载了介词短语识别规则的翻译系统的 BLEU 得分为 22.33%,比基线系统提高了 2.3%,同样表明了规则方法的有效性,介词短语的正确识别与分析有助于改善系统的译文流畅度。

4.2 统计方法

作为目前最早针对自然语言处理领域研究介词结构识别的文献,文献[2]中并未涉及后来流行的统计模型,只是对与介词短语相关的信息进行了概率统计。从一定意义上,也可以认为是采用了统计的方法。

该文献主要采用不完全句法分析的思路。在介词结构自动标注过程中,只观察介词结构的关键词—介词、介词结构的右边界词(文中称为“内相关词”)和紧邻右边界词的词语(文中称为“外相关

词”),然后进行关键词匹配。文中提出了两个简单的算法。算法一是从训练语料中分别抽取两类信息表,一类包括介词、内相关词词形、外相关词词形和前三者搭配后的共现次数,另一类包括介词、内相关词词性、外相关词词性和前三者搭配后的共现次数。具体做法是,从句子中先找到介词,作为介词短语的左边界,然后依次提取介词后的相邻两词,将介词、提取的相邻的两个词作为一个整体,在信息表中查找匹配,如果能够查到,则将表中三者的共现频率赋值给 Freq1;同理,依次提取介词后的相邻两词的词性标记,在表中查找匹配,如果能够查到,则将共现频率赋值给 Freq2。根据公式计算语料中每个候选介词短语右边界词语的得分,得分最高者即为右边界。利用算法一对手工标注的约 10 万字语料进行封闭测试,准确率为 83%左右。

算法二从上述训练语料中提取出两类信息,一类是介词与内相关词词形在同一子句中出现的频率 Freq1,一类是介词与内相关词共现的频率 Freq2。用 Freq2 除以两个频率的和,作为该词是否为介词短语右边界的效度。在测试语料中,依次计算每一个词的效度,值最大的即为介词短语的右边界。算法二对约 26 万字的语料进行了分批测试,开放测试准确率平均只有 40%左右,封闭测试准确率约为 93%。

文献[24]假设介词短语后边界的确定只跟其前面一个词及其词性和后一个词及其词性相关,采用统计模型对常用介词“在”后面可能充当其后边界的词进行最大似然估计,实验时对于数据稀疏的数据利用删除插值的策略进行平滑后能够得到较好的概率估计。以 60 万词人民日报语料为训练集,20 万词为测试集,对介词“在”的介词短语进行测试。封闭测试及开放测试的准确率分别达到了 97%和 93%。

随后的统计方法主要使用机器学习模型进行识别。常见的有隐马尔科夫模型(Hidden Markov Model, HMM)^[25]、最大熵模型(Maximum Entropy, ME)^[26]、支持向量机模型(Support Vector Model, SVM)^[27]、条件随机场模型(Conditional Random Field, CRF)^[28]等。

4.2.1 HMM 方法

Li 等^[29]利用 HMM 模型对包括介词短语在内的 11 种汉语短语进行识别,同时利用基于转换的错误驱动学习方法^[30]进一步改进识别效果。在包含 28 000 左右的测试语料(源自北大汉语语料库^[31])

上的实验中,介词短语的识别准确率达到 93.67%左右。

奚建清和罗强^[32]提出了一种基于 HMM 的介词短语界定模型,首先通过 HMM 的 Viterbi 算法计算汉语介词短语边界划分的最佳路径,初步对一个句子的介词短语进行界定,随后利用依存树库中介词短语的句法特征信息对初步识别的结果从有限多个右边界词语中选择一个最合适的词语与左边界词语形成介词短语搭配,以降低错误界定发生的几率。对哈工大共享依存树库中近 5000 句包含介词短语的语料进行了识别测试,准确率分别达到了 86.5%(封闭测试)和 77.7%(开放测试)。

4.2.2 ME 方法

于浚涛^[33]在充分考虑汉语介词短语结构和语言学信息的基础上,利用最大熵模型,设计了一系列模型所需的特征,对 2000 年人民日报语料中的七千多个介词短语进行了识别研究,五折交叉实验得到整体平均准确率达到 89%左右。

卢朝华等^[34-35]也利用最大熵模型识别 2000 年人民日报语料中的介词短语,但同时添加了依存语法错误界定校正处理。选择最大熵模型中右边界错误识别及没有识别的句子,从依存关系树库中提取介词短语的句法特征信息,寻找一个词语,使这个词语和介词短语的左边界具有最大的语义关联度。重复了文献[32]中的实验,结合依存语法后的测试平均准确率达到 90.77%,高于最大熵的平均准确率(88.9%)。

霍亚格和黄广君^[36]提出了一种基于互信息的最大熵模型识别包括介词短语在内的 15 种汉语短语结构,将短语结构识别问题转化为标注问题。但该文只针对非嵌套短语结构以及由相邻词语构成的短语结构,并未识别嵌套的和远距离搭配短语。首先利用语料库建立词语结合频次库,包含相邻两词或词性在语料中单独出现和共同出现的次数,根据两个词语之间的互信息知识对短语结构边界进行预测,然后应用最大熵模型识别短语结构。在人民日报语料的实验中,介词短语识别的准确率和召回率分别达到了 89%和 88%。

4.2.3 SVM 方法

温苗苗和吴云芳^[37]同样基于 2000 年 1 月人民日报语料,利用 SVM 模型建立了介词结构的自动识别系统,尝试加入了动词特征和语义类信息等不同的特征集合,对汉语中比较常见的多个并列和嵌套的多重介词短语情况做了重点研究。基于不同的

特征集分别对包含四万五千多个介词短语的测试集进行了五折交叉验证实验,准确率平均达到 90% 左右。

鉴萍和宗成庆^[38]根据汉语的特殊表现形式,首次从正向(由左至右)和反向(由右至左)两个方向对最长名词短语和介词短语进行标注,尝试基于 SVM 分类器的确定性标注模型识别这两类短语类型。文献认为,在序列标注中,沿某一方向第一个与另一方向标注结果不同的那个位置,才能真正反映该方向整个标记序列(或一个短语片段)的信任度,这个位置称作“分歧点”。据此提出了一种基于“分歧点”的概率融合算法。随后对宾州中文树库(V5.0)中《新华日报》语料 8000 多个介词短语进行了十折交叉验证实验。分别对包括文中提出的算法模型在内的四个标注融合系统进行了对比分析,其中介词短语的正向识别 F_1 值平均在 84% 左右,比反向识别 F_1 高了近 10 个百分点,融合了两个方向的 F_1 值约为 86%,均高于每个单向的测试结果。从而验证了基于“分歧点”的算法可以达到较高的融合精度,能更有效地识别出介词短语。

4.2.4 CRF 方法

文献^[15]基于宾州中文树库 5.1 版,分别利用一阶 CRF 和二阶 CRF 模型进行短语识别对比实验,其中介词短语准确率分别为 99.42% 和 98.95%,显示一阶模型优于二阶模型。

朱丹浩等^[39]基于清华汉语树库(TCT)^[40],详细分析统计了语料中出现次数大于 100 的介宾结构内部的词性序列和短语序列特征,以及介宾结构的外部短语序列的语言学特征。利用条件随机场模型,结合介宾结构的语言学特征,使用复杂特征模板对无嵌套和有嵌套的两种介宾结构进行自动识别。在开放测试中, F_1 值最高分别达到 90.29% 和 89.99%。

大连理工大学的多篇硕士毕业论文^[41-43]对基于 CRF 模型的介词短语识别做了比较连续深入的研究。他们相继设计了单层和多层的模型,分层次识别单一层次的以及含有嵌套结构的介词短语。同时利用基于转换的错误驱动学习模型对介词短语的识别结果进行校正。基于 2000 年人民日报语料的实验显示取得了不错的效果。图 1 是文献中提到的多层 CRF 识别的流程图。

张灵^[44]基于宾州中文树库语料,提出了一种基于层叠 CRF 的介词短语识别方法(图 2)。该方法将介词短语识别问题分成三个步骤:一是采用基于

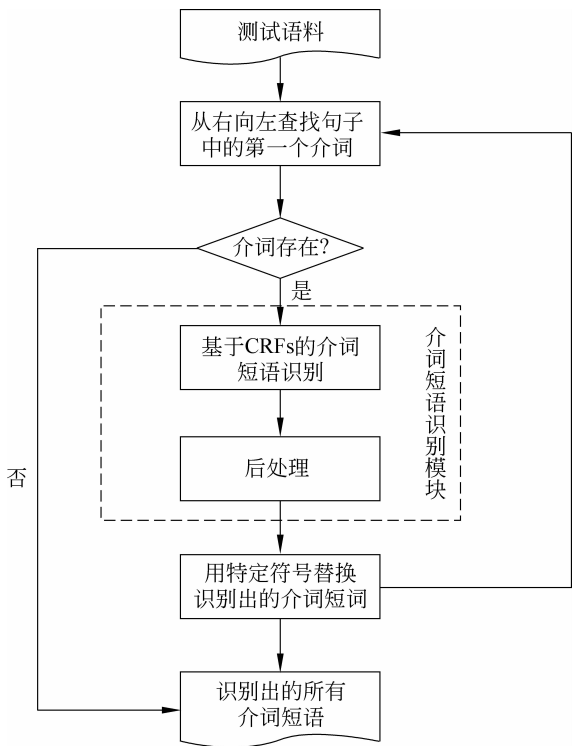


图 1 多层 CRF 识别方法流程图,摘自文献^[43]

搭配的方法对句子中的介词短语进行初步识别,并根据规则生成句子框架;二是基于 CRF 对句子框架进行短语结构分析,并采用深度优先的算法搜索最优句法分析结果;三是将第一步的介词短语识别结果和第二步的句法分析结果进行融合。对前两个步骤分别采用不同的条件随机场模型进行学习,第三步得到融合结果。对 CTB4.0 的语料进行了十折交叉验证实验,平均的 F_1 值达到了 90% 左右,两模型结果的融合使介词短语的识别准确性得到了提高,比采用单层的 CRF 模型的识别效果有所提升。该文献设计了基于 CRF 的介词短语识别模型和句法分析模型,利用句法分析的信息辅助识别介词短语,是一个比较大的创新。

Li 等^[45]同样利用 CRF 模型对汉语专利文献中的介词短语进行识别。针对汉语专利语料的特点和介词短语的语言学信息,设计了一组有效的识别特征,利用 CRF++ 工具包训练识别所需的模型,然后对 NTCIR9 汉英专利机器翻译评测使用的测试集进行了五折交叉验证测试,最终的准确率达到 93% 左右。与使用规则方法识别的 Baseline 的相比,准确率提高了两个百分点,召回率提高了六个百分点。

以上研究主要利用单独的某一个模型识别介词短语,还有一些文献同时运用了多种模型。例如, Sun 和 Huang 等^[46]把 HMM 与 ME 模型结合,提出

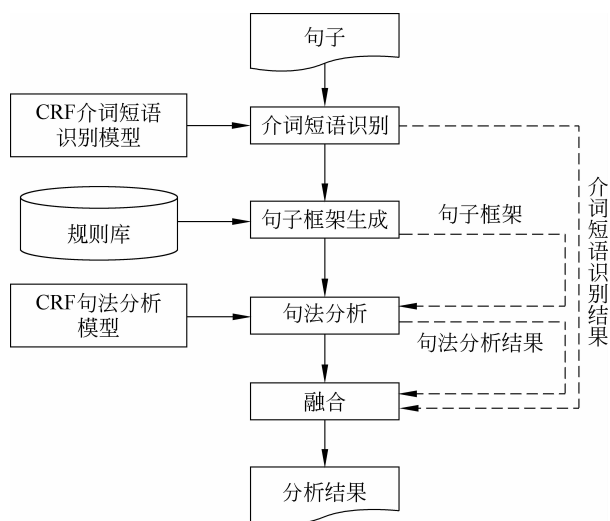


图2 基于层叠条件随机场的介词短语识别流程图^[44]

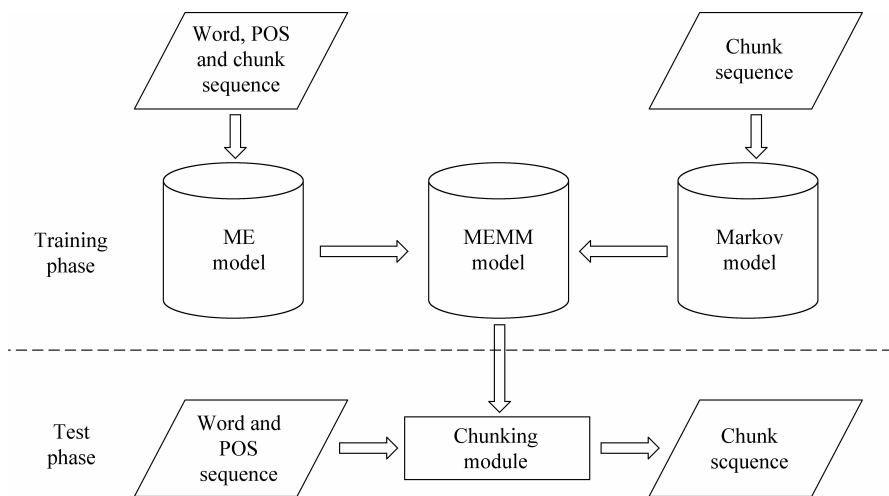


图3 基于 MEMM 的短语识别架构,摘自文献^[46]

以上就是介词短语识别中常用的几种统计方法。这几种模型都属于有监督的机器学习,其中 HMM 属于生成式模型,其他三种均属于判别式模型。这四种模型都可以使用成熟的工具包训练,而且每个训练模型都必须要选择合适的特征集,但具体的文本及特征格式,训练过程和训练时间等会有所不同。

4.3 混合方法

规则方法和统计模型有各自的优势与不足,近些年来,也出现了很多尝试充分利用两种方法的优势,将二者结合识别介词短语的研究,即混合方法的思路。

介词短语及其上下文中经常出现固定搭配现

了一种最大熵马尔科夫模型(Maximum Entropy Markov Models, MEMM)来识别汉语语块(识别过程如图3所示),同时利用平滑算法解决组块标注中数据稀疏问题。分别利用宾州中文树库和北京大学语料库进行实验,结果显示,MEMM 识别组块的效果均要优于单独的 HMM 和 ME 模型,在两种测试集中,介词短语识别的正确率分别达到了 99.11% 和 93.98%。

文献^[47]和^[48]则从训练模型所需的特征集和训练过程等方面对 ME、SVM 和 CRF 模型进行了横向的对比分析,分别利用这三种模型对 2000 年 1 月的人民日报语料进行了识别测试,平均的准确率达到了 80% 左右,同时得到的结论是,不同的模型对于介词短语边界的识别效果不同,其中 ME 模型最好,CRF 次之,SVM 最差,但三种模型在总体识别准确率上都明显优于基于规则的识别准确率。

象,比如“除……之外”、“就……而言”、“从……到”等,搭配特征是确定介词短语后界的重要依据。干俊伟和黄德根^[49]据此选择了两个搭配模板: $T1 = \langle \text{介词}, \text{后界}, \text{后界的词性} \rangle$ 、 $T2 = \langle \text{介词}, \text{后词}, \text{后词的词性} \rangle$,利用下面的公式计算搭配关系的可信度。其中 $\text{CorrectFrame}(\langle p, w, pos \rangle)$ 表示搭配关系, $\langle p, w, pos \rangle$ 在训练语料中出现的次数。 $\text{TotalFrame}(\langle p, w, pos \rangle)$ 表示词 w 标记为词性 pos 在介词 p 右方出现且 p, w 属于同一分句的次数。文献选取 TotalFrame 值大于等于 5 且搭配关系可信度大于 90% 的搭配关系作为可信搭配关系。

$$\text{Probability Frame}(\langle p, w, pos \rangle) = \frac{\text{CorrectFrame}(\langle p, w, pos \rangle)}{\text{TotalFrame}(\langle p, w, pos \rangle)} \quad (4)$$

对于可信度搭配无法识别的其他介词短语,接下来利用三元边界统计模型和规则相结合的方法识别。根据介词短语的语言学特征,制定了几条辅助性的识别规则。随后对 2000 年的人民日报语料进行了五折交叉验证测试,准确率和 F_1 值在 86%~87% 左右,比基线系统的实验结果提高了近个百分点,证明统计模型与规则结合以后比单独使用统计模型有效地提高了识别的精度。付禾芳和李朝霞^[50]后来同样利用最大熵模型和搭配关系可信度以及人工规则相结合的方法识别远距离的介词短语,虽然提到“基于词性的三元边界统计模型结合规则之后,识别效率明显地提高”,但并未给出具体的实验结果。

咎红英等^[51]在已有工作的基础上,分析对比了规则方法与统计方法的优劣,提出一种规则与 CRF 模型相结合的介词用法自动识别算法。文献首先分别利用人工书写的规则和 CRF 模型对 2000 年 5 月份人民日报语料中使用频率最高的 20 个介词进行自动识别测试,得到了每个介词的识别准确率和两种方法的总准确率: 67.38% (规则方法) 和 76.80% (CRF)。然后将这两种方法分别在宏观层面和微观层面进行结合,通过相同的实验得到宏观和微观的识别准确率以及 20 个介词各自的准确率。宏观总的准确率为 78.47%,比规则方法高 11.09%,比统计提高 1.67%。微观总的准确率为 82.02%,比规则方法高 14.64%,比统计方法高 5.22%。混合方法的识别准确率比单一方法的准确率有显著提高,同时微观结合效果要好于宏观结合。

郭丹丹和由丽萍^[52]基于框架语义学理论^[53-54],面向中文核心依存分析技术,运用规则和统计相结合的方法,尝试在指定支配性谓词的情况下识别一个分句内部中从属于谓词的介词短语。首先根据介词和介词短语右边界的规律抽取搭配模板,从训练语料中自动提取搭配关系,并用这些搭配关系在一定的搭配策略下对介词短语进行识别。然后,用基于词性的边界选择模型和规则方法相结合的技术对其它介词短语进行识别。以山西大学构建汉语的框架语义知识库语料^[55]为实验对象进行了五折交叉测试验证,最终的准确率达到 79% 左右。

Li 和 Jin^[56]针对中国专利局提供的汉语专利文本语料,分析了专利文本中介词短语的位置、分类等语言学特征,分别提出了一种基于规则的方法和 CRF 识别模型。以一千句包含介词短语专利语料为测试集,比较了两种方法的实验结果,同时统计了

测试集中出现频率最高的十个介词的实验数据。规则方法的总体准确率(96.86%)要高于 CRF 方法的准确率(92.65%),但 F_1 值要低一些。尽管两种方法的准确率达到 90% 以上,但该研究并未像文献[51]那样把规则方法和 CRF 方法结合在一起。

5 总结

本文对近些年来识别汉语介词短语的研究做了比较详细的梳理与说明,主要讨论了具体使用的方法。表 1 从每类方法中选择几项代表性的研究工作,横向对比这些方法的特点。

从前文的论述和表 1 可以大致归纳出目前汉语介词短语识别研究主要存在的几个特点:

(1) 从研究语料来看,大多数研究主要针对人民日报语料库以及新闻语料树库(如宾州中文树库)中的介词短语,文本领域比较单一。相比其他领域的语料,新闻语料在自然语言处理领域的各种任务中更为常见和普遍,而且在语料的标注处理等方面要更为成熟,对相关研究有较好的利用价值,这也是为什么多数研究多集中于该领域的原因。

(2) 从识别方法来看,识别方法呈现出多样性,无论是规则方法还是统计方法,国内的研究都做了比较全面的尝试,而且尽可能地涉及复杂的介词短语类型,研究方法也逐渐表现出规则与统计方法相结合的趋势。但从整体来看,统计方法的研究还是明显要多于规则方法。在统计方法的几种常用模型中,很多研究并不是单纯地使用某一种模型,而是尝试将几种模型相结合,或者将模型与其他策略(如互信息,基于转换的错误驱动方法等)有机结合,尽可能提高识别效果。受到语料类型和语料规模的影响,不同模型的识别效果有所不同,但总体上基本达到了比较满意的效果。

(3) 在统计方法中,相对于其他模型,大多数文献更倾向选择使用 CRF 模型。这是由 CRF 自身的优势决定的。隐马尔可夫模型容易引起数据稀疏等问题,而最大熵模型对于规则的描述又过于烦琐。CRF 作为一种用于序列标注的判别模型,以兼具生成式模型和序列分类器模型的优点著称,可以使用观测序列的任何特征并搜索全局最优标注结果,较好地克服了输出独立性假设和马尔可夫假设的局限性,并且能从上下文中任意地选择所需要的特征,可以有很出色的表现。

表 1 几种识别方法对比

方法分类		代表工作	所用语料	方法特点	实验设置	实验结果
规则方法		文献[23]	专利文本语料	利用包含语言学信息的词语知识库和人工书写的规则	封闭+开放测试,NTCIR9 汉英专利机器翻译评测测试集 2000 句。采用准确率(P)和召回率(R)评价。	封闭:(P)约 90%, (R)约 85% 开放:(P)平均 90%, (R)平均 82%
统计方法	HMM	文献[32]	哈工大依存树库	设计选取合适的特征,利用工具包训练模型	封闭+开放测试,训练集 3 466 句,封闭测试集 1 500 句,开放测试集 1 500 句。采用(P)和(R)评价。	封闭:均在 85%左右 开放:均在 76%左右
	ME	文献[36]	人民日报语料	首先利用互信息预测短语边界,然后利用 ME 识别	训练集 500 篇,测试集 100 篇。采用(P)和(R)评价。	均在 88%左右
	SVM	文献[38]	宾州中文树库(V5.0)	正向和反向标注字符串,找到分歧点,然后利用 SVM 训练模型	测试集 9 493 句,其中 8 282 个 PP。采用 F ₁ 值评价。	正向 83%,反向 74%, 融合 85%
	CRF	文献[44]	宾州中文树库(V4.0)	基于 CRF 的介词短语模型与基于 CRF 的句法分析相结合	测试集约 15 000 句。十折交叉验证,采用 F ₁ 值评价。	均值 90%
混合方法		文献[51]	人民日报语料	人工规则与统计模型相结合	测试集规模未知。采用(P)评价。	宏观:78%, 微观:82%

(4) 实验大多采用了五折或十折交叉验证的方法,以保证测试效果的均衡性。

尽管目前的研究已对介词短语识别问题做了很多积极的探索,并取得了令人可喜的成果,但尚存有一些不足,仍有继续深入研究的空间。对于未来的研究发展方向,笔者尝试提出几点建议,希望能对感兴趣的研究者具有一定的参考作用。

首先,目前大多数研究还是主要停留在结构比较单一、字数比较少的介词短语的层面,但由于复杂的自然语言具有递归性和嵌套性的特点,未来应该利用现有的成熟技术重点解决字数更多、结构更为复杂的介词短语(如多层嵌套)的识别问题。如果有效处理了难度更大的结构,介词短语的识别技术必然会达到新的水平,从而促进自然语言处理和自然语言理解的发展。

其次,由于自然语言处理已应用到越来越多的领域,希望未来可以从传统的新闻领域逐渐扩展到其他更多领域,如军事、农业、天气和科技文本等,结合不同文本的语体风格和表达特点等研究介词短语的识别和相关问题,以满足不同的需求。我们大胆推测,即使是结构相同的甚至是同一个介词短语,如果处在不同领域文本中,其识别效果很可能也是不同的。另一方面,还可以探索把识别方法从文本处理转移到语音识别等语音信息处理领域的可行性,或许也会有意想不到的结果。

再次,无论是规则方法还是统计方法,未来都可

以尝试加入更多颗粒度更细致的句法语义特征等,以降低词语的歧义问题和难以确定边界的困难。尝试“分而治之”的思想,加强对介词短语的内部结构的分析,在首先识别其他短语的基础上,再进行介词短语的识别。同时,需要扩大语料训练和测试规模,使训练的模型更加有效,以有效避免数据稀疏等问题。

第四,介词短语识别的最终目的是服务于自然语言处理的众多任务和应用。未来需要进一步与机器翻译、问答系统、信息抽取、文本分类等热门领域相结合,在具体应用中检验介词短语的识别效果,以及由此产生的实际影响。如果脱离了具体应用,只是孤立单纯地谈论实验结果,是不够的。

最后,近几年来随着机器学习的普遍流行,未来可以尝试将半监督的机器学习方法应用到语料标注和处理中,初步实现自动标注的目标,以减少人工标注语料费时费力的问题,提升标注效率。

参考文献

[1] 吴云芳. 现代汉语介词结构的自动标注[D]. 北京语言大学硕士学位论文,1998.

[2] Abney S. Parsing by Chunks[A]. In: Berwick R. , Abney S. and Carol T. (Eds.), Principle-Based Parsing. Dordrecht: Kluwer Academic Publisher. 1991: 257-278.

- [3] 李业刚, 黄河燕. 汉语组块分析研究综述[J]. 中文信息学报, 2013, 27(5): 1-9.
- [4] Erik F. Tjong Kim Sang and Sabine Buchholz. Introduction to the CoNLL-2000 Shared Task: Chunking [C]//Proceedings of CoNLL-2000 and LLL-2000, 127-132.
- [5] 李素建, 刘群, 白硕. 统计和规则相结合的汉语组块分析[J]. 计算机研究与发展, 2002, 39(4): 385-391.
- [6] 张昱琪, 周强. 汉语基本短语的自动识别[J]. 中文信息学报, 2002, 16(6): 1-8.
- [7] 周强. 汉语基本块描述体系[J]. 中文信息学报, 2007, 21(3): 21-27.
- [8] 孙广路. 基于统计学习的中文组块分析技术研究[D]. 哈尔滨工业大学博士学位论文, 2008.
- [9] 邹宏梅, 王挺. SVM 和基于转换的错误驱动学习相结合的汉语组块识别[J]. 计算机工程与科学, 2007, 29(4): 91-94, 123.
- [10] 秦颖, 王小捷, 钟义信. 级联中文组块识别[J]. 北京邮电大学学报, 2008, 31(1): 14-17.
- [11] 王莹莹. 汉语组块识别的研究[D]. 大连理工大学硕士学位论文, 2006.
- [12] 高红. 基于统计语言模型的汉语浅层分析研究[D]. 大连理工大学博士学位论文, 2007.
- [13] 谭咏梅, 姚天顺, 陈晴, 李布, 朱靖波. 基于 SVM+Sig-moid 的汉语组块识别[J]. 计算机科学, 2004, 31(8): 142-146.
- [14] 李珩, 朱靖波, 姚天顺. 基于 SVM 的中文组块分析[J]. 中文信息学报, 2004, 18(2): 1-7.
- [15] 郭永生. 基于条件随机场的汉语短语识别研究[D]. 东北大学硕士学位论文, 2008.
- [16] 周强, 张伟, 俞士汶. 汉语树库的构建[J]. 中文信息学报, 1997, 11(4): 1-11.
- [17] 咎红英, 张坤丽, 柴玉梅, 俞士汶. 现代汉语虚词知识库的研究[J]. 中文信息学报, 2007, 21(5): 107-111.
- [18] 俞士汶, 朱学锋, 王惠等. 现代汉语语法信息词典详解[M]. 北京: 清华大学出版社, 2003.
- [19] 梁猛杰, 宋玉, 韩英杰等. 基于规则排序的介词用法自动识别研究[J]. 河南师范大学学报(自然科学版), 2013, 41(3): 152-155.
- [20] 黄曾阳. HNC(概念层次网络)理论[M]. 北京: 清华大学出版社, 1998.
- [21] Zhu Yun, Jin Yaohong. A Chinese-English patent machine translation system based on the theory of hierarchical network of concepts [J]. The Journal of China Universities of Posts and Telecommunications, 2012, 19(Suppl. 2): 140-146.
- [22] 朱筠. 基本句群处理及其在汉英专利机器翻译中的应用[D]. 北京师范大学汉语文化学院硕士学位论文, 2013.
- [23] 胡韧奋. 面向汉英专利机器翻译的介词短语自动识别策略[J]. 语言文字应用, 2015, 1: 136-144.
- [24] 王立霞, 孙宏林. 现代汉语介词短语边界识别研究[J]. 中文信息学报, 2005, 19(3): 80-86.
- [25] Rabiner L R. A tutorial on hidden Markov models and selected applications in speech recognition [C]//Proceedings of the IEEE, 1989, 77(2): 257-286.
- [26] E T Jaynes. Information theory and statistical mechanics [J]. Physics Reviews, 1957, 106: 620-630.
- [27] Vapnik V N. Statistical Learning Theory [M]. Wiley-Interscience Publication: John Wiley & Sons, Inc. 1998.
- [28] J Lafferty, A McCallum, F Pereira. Conditional random fields: probabilistic models for segmenting and labeling sequence data [C]//Proceedings of International Conference on Machine. 2001: 282-289.
- [29] Li Hongqiao, Huang Changning, Gao Jianfeng and Fan Xiaozhong. Chinese chunking with another type of spec [C]//The Third SIGHAN Workshop on Chinese Language Processing. 2004: 24-26.
- [30] Eric Brill. Transformation-based error-driven learning and natural language processing: A case study in part of speech tagging [J]. Computational Linguistics, 1995, 21(4): 543-565.
- [31] 俞士汶, 段慧明, 朱学锋, 孙斌. 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(6): 58-65.
- [32] 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, 182.
- [33] 于浚涛. 基于最大熵的汉语介词短语自动识别[D]. 大连理工大学硕士学位论文, 2006.
- [34] 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, 186.
- [35] 卢朝华, 徐好芹, 王玉芬. 基于语义分析的汉语介词短语识别方法研究[J]. 电脑与电信, 2012, 3: 46-48.
- [36] 霍亚格, 黄广君. 基于最大熵的汉语短语结构识别方法[J]. 计算机工程, 2011, 37(16): 206-208, 211.
- [37] 温苗苗, 吴云芳. 基于 SVM 融合多特征的介词结构自动识别[J]. 中文信息学报, 2009, 23(5): 19-25.
- [38] 鉴萍, 宗成庆. 基于双向标注融合的汉语最长短语识别方法[J]. 智能系统学报, 2009, 4(5): 406-413.
- [39] 朱丹浩, 王东波, 谢靖. 基于条件随机场的介宾结构自动识别[J]. 现代图书情报技术, 2010, (7/8): 79-83.
- [40] 周强, 张伟, 俞士汶. 汉语树库的构建[J]. 中文信息学报, 1997, 11(4): 42-51.
- [41] 胡思磊. 基于 CRF 模型的汉语介词短语识别[D]. 大连理工大学硕士学位论文, 2008.
- [42] 宋贵哲. 汉语介词短语识别研究[D]. 大连理工大学硕士学位论文, 2011.
- [43] 张杰. 基于多层 CRFs 的汉语介词短语识别研究[D]. 大连理工大学硕士学位论文, 2013.
- [44] 张灵. 基于层叠条件随机场的汉语介词短语识别研究[D]. 沈阳航空航天大学硕士学位论文, 2012.

- [45] Li Hongzheng and JinYaohong. A CRF Method of Identifying Prepositional Phrases in Chinese Patent Texts [C]//Proceedings of the Eighth SIGHAN Workshop on Chinese Language Processing (SIGHAN-8). 2015, 86-90.
- [46] Sun GuangLu, Huang ChangNing, Wang XiaoLong and Xu ZhiMing. Chinese Chunking Based on Maximum Entropy Markov Models [J]. Computational Linguistics and Chinese Language Processing, 2006, 11(2): 115-136.
- [47] 袁应成. 基于用法属性的现代汉语介词短语边界识别研究[D]. 郑州大学硕士学位论文, 2011.
- [48] 张坤丽, 韩英杰, 咎红英, 袁应成. 基于统计的介词短语边界识别研究[J]. 河南大学学报(自然科学版), 2011, 41(6): 636-640.
- [49] 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.
- [50] 付禾芳, 李朝霞. 介词短语识别中规则与统计方法融合的探讨[J]. 研究与开发, 2010, 11: 17-20.
- [51] 咎红英, 张腾飞, 张坤丽. 规则与统计相结合的介词用法自动识别研究[J]. 计算机工程与设计, 2013, 34(6): 2152-2157.
- [52] 郭丹丹, 由丽萍. 面向核心依存分析的介词短语自动识别[J]. 情报探索, 2014, (11): 1-3.
- [53] Charles J. Fillmore. Frame semantics and the nature of language [J]. Annals of the NY Academy of Sciences, 1976, (2): 20-32.
- [54] Charles J. Fillmore, Collin F. Baker and Hiroaki Sato. The FrameNet Database and Software Tools [C]//Proceedings of the Third International Conference on Language Resources and Evaluation, 2002, 1157-1160.
- [55] 由丽萍. 中文框架语义分析[M]. 北京: 经济科学出版社, 2013.
- [56] Hongzheng Li and Yaohong Jin. Identifying Prepositional Phrases in Chinese Patent Texts with Rule-based and CRF Methods [C]//Proceedings of 29th Pacific Asia Conference on Language, Information and Computation, 2015, 143-149.



李洪政(1990—), 通信作者, 博士研究生, 主要研究领域为机器翻译、深度学习等。
E-mail: lihongzheng@mail.bnu.edu.cn



晋耀红(1973—), 博士, 教授, 博士生导师。主要研究领域为数据挖掘、信息抽取等。
E-mail: jinyaohong@hotmail.com

方滨兴院士当选中国中文信息学会第八届理事会理事长, 李生教授担任名誉理事长

2016年12月23—24日,中国中文信息学会第八次全国会员代表大会暨学会成立35周年学术会议在北京中国科技馆隆重举行,大会通过无记名等额投票方式选举产生了中国中文信息学会第八届理事会全体成员、常务理事会全体成员以及学会领导班子成员。中国电子信息产业集团方滨兴院士当选中国中文信息学会第八届理事会理事长,北京理工大学黄河燕教授、北京语言大学李宇明教授、科大讯飞刘庆峰董事长、拓尔思施水才董事长、中科院软件所孙乐研究员、清华大学马少平教授、内蒙古大学那顺乌日图教授、格微软件张桂平董事长、百度副总裁王海峰教授、中科院自动化所宗成庆研究员当选副理事长,中科院软件所孙乐研究员当选秘书长,北京语言大学杨尔弘教授当选副秘书长。

12月14日,召开了第八届理事会第一次全体会议,专委会和工委会主任们向理事会汇报各委员会的工作,学会新任理事长方滨兴院士针对学会工作发表了一系列重要讲话,为学会今后工作指明了方向。随后还召开了学会第八届理事会第一次常务理事会,经方滨兴理事长提议,大家一致赞同李生教授担任学会名誉理事长。在学会第八届理事会第一次常务理事会党员会议上,依据中国科协关于建立学会党组织文件精神,成立了学会党组织,并已上报中国科协科技社团党委。