大规模真实文本中汉语动词语法搭配模板的自动识别*

张昱琪　周强

智能技术与系统国家重点实验室

清华大学计算机科学与技术系，北京 100084，中国

zyq@s1000e.cs.tsinghua.edu.cn　zhouq@s1000e.cs.tsinghua.edu.cn

摘要：本文通过一种规则匹配的方法，对真实语料中的动词语法搭配模板进行自动识别。该种方法在词界块边界预测信息的基础上，以最长名词短语的识别为切入点，通过规则匹配的方法，在大规模真实文本上完成了“动词+名词短语”、“动词+动词短语”、“动词+名词短语+动词短语”、“动词+名词短语+动词短语”4 类动词语法搭配模板的自动识别。初步实验结果显示这种方法对“动词+名词短语”模板的识别具有较好的效果，4 类模板的总体识别正确率为 76.55%，召回率为 61.93%。

关键词：模板，动词，语法搭配，自动识别，大规模语料

Automatic Acquisition of Chinese VerbSubcategorization Frames from Large Scale Tagged Corpora

Zhang Yuqi, Zhou Qiang

National Key Laboratory of Intelligent Technology & Systems

Department of Computer Science and Technolog

Tsinghua University, Beijing 100084, P. R. China

zyq@s1000e.cs.tsinghua.edu.cn　zhouq@s1000e.cs.tsinghua.edu.cn

Abstract: This paper presents a method for automatically acquiring Chinese verb subcategorization frames from a large corpus. The method is based on a preprocessor that ascertains the boundaries of phrases. A tagged corpus is first partially parsed to identify maximal noun phrases and then some templates are used to guess the appropriate subcategorization frames for each verb token in the corpus. In an experiment involving the identification of four fixed subcategorization frames, our current system showed more than 75% accuracy.

Keywords: subcategorization, verb, grammatical collocation, automatic acquisition, large scale corpora,

1 引言

语言中存在许多固定的、可确认的、但又不属于成语的短语和结构。这些词的组合被称作可再现的组合、固定组合或搭配。搭配作为描述词间组合能力的一种重要的词汇知识，在

*本项研究得到国家自然科学基金(项目号：69903007)和国家 973 基金(项目号：1998G030507-2)的资助，在此表示衷心的感谢。

自然语言处理中具有重要作用。Benson.M. *et al.* (1986)在 BBI 英语搭配词典中定义搭配（collocation）为“语言中任意的、可重现的词组合”，并把搭配分为语法搭配（grammatical collocation）和词汇搭配(lexical collocation)。语法搭配是由一个起支配作用的词（可以为名词、形容词或动词）与一个语法结构，例如介词结构，小句等组成的短语。考虑到动词在汉语句子中的支配地位，该项研究选择了汉语动词作为研究对象。希望利用汉语自动句法分析技术，从大规模真实文本中自动发现汉语动词的语法搭配。

近些年来，国外的研究人员在英语的语法搭配方面作了很多探索。在语法搭配类型总结方面，《The BBI combinatory Dictionary of English》[2]把英文的语法搭配被分成了 8 大类，其中在第 8 类中着重分析了有关动词的语法搭配，提出了 19 种动词搭配模板。Ushioda[3]等对英语动词语法搭配模板的自动识别进行了探索。他们把英语动词搭配模板进行了分类处理：对易于用语法规则表达的模板，采用提取词性特征书写规则，按规则进行匹配的方法。对于不宜用语法规则概括的结构用 Loglinear model[1]方法进行了研究，取得了较满意的效果（正确率达到了 83%）。不仅搭配模板的自动识别，各类模板出现的频度和具体搭配实例提取的分析器也已经开始了尝试[8]。

汉语搭配信息自动获取的研究工作还处于刚刚起步的阶段。《现代汉语语法信息词典》[5] 从语言学的角度比较详尽地描述了汉语中受到语言学界普遍认可的 5 万多词的语法属性，包括词法和句法。词条的获取主要依靠于语言学家的语法知识。但词条中没有词法和句法出现频度的信息。在汉语搭配知识的自动发现方面，白硕[4]进行了一些开拓性的探索。但在文献中还没有见到在大规模真实文本上进行过类似的自动获取研究。汉语动词搭配模板自动获取技术的开发，不仅可以作为技术应用于对语料库的处理加工和其它信息抽取工作；还可以获得动词语法搭配和词法搭配的有用信息，为语言学的研究和自然语言的排歧工作提供资源。

本文提出了一种基于模式匹配的方法，用于在大规模真实文本上自动识别各类、各层次的动词语法搭配模板。现阶段主要识别以下 4 类模板：“v+np”、“v+vp”、“v+np+vp”、“v+np+np”；随着工作的深入，能够自动识别的模板种类将会逐渐增多。该种方法在短语边界自动界定的基础上，首先自动识别出句子中的最长名词短语，然后按照事先设计好的模板类型用部分分析和模式匹配的方法进行自动识别。实验结果显示，这种方法对于“v+np”模板的识别效果较好，4 类模板的整体识别率为 76.55%，召回率为 61.93%。实验结果同时也显示了该方法对于处理动词短语和复杂名词短语效果不佳的不足之处。

2 实验基础与前提

该实验的基础语料库是 200 万汉字的汉语平衡语料库，其中大部分来源于九十年代的出版物，一小部分来源于八十年代的出版物。其体裁主要有 4 类：文学作品（约 44%）、新闻（约 30%）、学术文章（约 20%）和实际生活中的语言（约 6%）。这些语料经过了分词和词性标注两步预处理，并经过了人工校对。

词界块（word stem, WS）[7]是由句子中的词语与它的成分边界位置标记组成的结合体，简记为 $ws_i = <w_i, b_i>$，其中 b_i 可取值 0,1,2,3。

b_i=0，则表示 w_i 处于成分中间位置；b_i=1，则 w_i 处于成分左边界；b_i=2，则 w_i 处于成分右边界；b_i= 3，则 w_i 既可处于成分左边界，又可处于成分右边界，即存在歧义现象。

通过构造统计模型，可以对词界块的成分边界进行自动预测。文献[6]给出了具体的预测算法。

在动词语法搭配模板的识别中，名词短语的识别占有非常重要的地位。动词语法搭配不可避免的要涉及名词短语，如：模板‘v+np’，‘v+np+vp’等。因此名词短语的正确识别是动词语法搭配识别的基础。最长名词短语的自动识别[7]基于名词短语（np）的内部结构，采用自底向上的句法分析方法，取得了较好的效果。（正确率 85.4%，召回率 82.3%，其中长度大于 5 的最长 np 的正确率为 70.8%，召回率为 61.7%）

3 实验方法

3.1, 识别的模板种类

表 1：汉语动词语法搭配模板

5 种结构	举例	4 大类模板
v+np|n	打排球	v+np|n
v+vp|v	喜欢唱歌	v+vp|v
v+np|n+np|n	喂孩子牛奶	v+np|n+np|n
v+np|n+vp|v	告诉他下雪了	v+np|n+vp|v
v+np|n+vp|v(兼语)	请客人吃饭	

在现阶段研究中，我们主要识别表 1 中 4 大类模板，其中‘v+np|n+vp|v’模式中的兼语结构和非兼语结构，留待在下一阶段工作根据动词的属性和词汇信息进行区分。该 4 类动词模板的自动识别只是“汉语动词语法搭配模板的自动识别”这一课题的第一部分工作，随着今后研究工作的深入，能够自动识别的动词语法搭配模板类型将会增多。

3.2, 实验算法

输入：经过分词、词性标注、词界块边界预测处理的汉语句子

输出：以动词为文件名的文件，文件中包括：包含该动词的整句；识别出的该动词所带的模板类型；自动抽取出的各类模板类型的实例及实例在原句中的位置。

数据结构：栈结构

基本流程：①从左向右扫描整个句子，利用最长名词短语识别算法识别出句子中的最长名词短语。识别结果保留在栈中。

②自栈顶向栈底（对句子来说相当于从右向左）按搭配模板进行匹配：

◇ 必要条件：模板边界 => 词语块边界

◇ 在栈顶发现可能的模板右边界，其中包括：

i 词语块右边界；　ii 名词短语块及名词；　iii 简单的动词短语块及动词；

◇ 若栈顶的元素为最长名词短语块，则进入语块内部识别其中可能包含的模板

◇ 向栈底方向匹配模板，匹配成功，则进行规约，把规约结果放入栈顶；继续向左匹配模板，直到不能再找到符合模板的结构，把栈顶元素弹出栈。栈空结束。

图 1 汉语动词语法搭配模板自动识别算法

为了将动词语法搭配模板识别和最长名词短语识别较好地结合起来，我们设计了一种栈结构。在栈顶按照模板类型和词界块信息，进行移进-规约操作。基本内容如图 1。

整个算法对整个句子一共做过两遍处理。第一遍自左向右处理句子，根据名词短语的内部结构和词语块边界信息在栈顶进行移进-规约操作。识别句子中的最长名词短语。

动词搭配模板的识别是自右向左的对句子的第二遍处理：判断栈顶元素是否可能为模板右边界；自右向左匹配已知的 4 种模板。如果匹配成功，则把规约后得到的动词短语（可以认为动词语法搭配模板是一种动词短语）压入栈顶，这样可以识别嵌套的动词短语中的模板；如果栈顶元素不可能是模板的一部分，则把它弹出堆栈。此过程不断重复，直到栈为空。

4 实验结果分析

我们实验所用的语料来源于 2.1 节中提到的 200 万字的语料库。所用语料约有 3Mbytes 大小，包含约 35 万词，共有 14000 多句，平均句长 23.8 词/句。

语料中共有 7158 个不同的动词，自动识别算法从中找到的所带模板种类数>=1 的动词共有 3345 个。即另外的 3813（=7158－3345）个动词或为不及物动词或带未知类型模板。自动识别出的这 3345 个动词所带语法模板的出现频度很不均匀，大部分动词的模板出现总频度<5。为了对实验结果进行评估，我们从实验结果中模板出现频度最高的 300 个动词里随机抽取出 30 个动词，人工识别出语料库里这 30 个动词所带的模板及模板类别；然后把自动识别结果和人工标注的正确结果进行比较。表 2 中列出了随机抽取出的 30 个动词。

表 2： 测试动词表

消灭	考虑	增强	管理	负责	改善	改造	扩大	开发	深入
从事	贯彻	依靠	满足	看见	采用	维护	开展	改变	培养
获得	发挥	作出	实行	分析	存在	进入	取得	产生	形成

为了检查动词语法模板自动识别器的性能，我们设定了以下两个指标：

1） 正确率= 正确识别的模板总数 （PFT）/自动识别出的模板总数（EFT）

2） 召回率= 正确识别的模板总数 （PFT）/语料库中所有应识别出的模板总数（TFT）

表 3：动词语法搭配模板自动识别结果

	PFT	EFT	TFT	正确率	召回率
v+np|n	765	930	1216	82.26%	62.9%
v+vp|v	70	125	104	56%	67.3%
v+np|n+vp|v	23	65	66	35.38%	34.85%
v+np|n+np|n	4	6	6	66.67%	66.67%
Rest	399	677	410	58.94%	97.32%
Total(不包含 rest)	862	1126	1392	76.55%	61.93%
Total (包含 rest)	1261	1803	1802	69.94%	69.98%

表 3 显示，在 4 个模板中，模板‘v+np’的正确率比较高；4 个模板的召回率大多数都在 60%-70%之间，这说明有大量应该被提取出的模板实例被识别为‘rest’类，这也是‘rest’的“自动识别出的模板总数”比“语料库中所有应识别出的模板总数”大的原因。模板‘v+np+vp’，‘v+np+np’提取出来的实例数较少，对最终统计结果的准确性有一定影响。

‘v+np’模板和‘rest’的实例数量和其它模板相比占优势，因此这两部分识别结果对最后总体统计结果有较大的影响。

实验算法目前主要存在以下几个方面的识别难点：(注：例子中“____”是正确的结果)

1. 具有复杂结构的“的”字短语

语料中经常会出现结构“[v [A] 的[np]”，其中“A”的结构非常灵活多样，可以是名词短语，动词短语，介词短语，形容词短语等。这种复杂的“的”字短语是导致模板‘v+np’识别召回率较低的重要原因之一。由于复杂名词短语包含的字数较多，结构复杂。现在的算法在栈顶进行移进——归约操作，对较远距离的语法信息不能有效的加以利用。名词短语的识别错误直接影响到动词语法搭配模板的识别效果。

2， 并列结构的短语

语料句子中的并列结构是‘v+np’识别召回率较低的另一个重要原因。我们的算法中有专门识别并列短语的模块，但原先的认识是并列短语的各个成分之间大多数存在词性、字数等特征的对称关系。从实验结果分析看出并列连词前后的并列成分情况还是比较复杂的。有的并列成分并不存在这些特征的对称关系，如：*[消灭/v [我/rN [军/n 主力/n]和/c [“/“ [捕捉/v [中共/nO 首脑部/n]”/”]*。目前的算法对这类并列短语识别比较困难。从实验的结果中还发现相当一部分复杂的名词短语（短语包含的字数>=5）都是由于其中包含并列短语而产生的。所以并列短语的识别对长句子识别效果的提高是非常重要的。

3，“[v [vp|np]的 [np|vp]”结构

“[v [vp|np]的 [np|vp]”是一个很典型的歧义结构，根据模板右边界的不同有两种可能性。一种是模板右边界在“的”字前（[v[vp|np]），如：*[是/vC [产生/v [生命/n 和/c 力度感/n]的/u [重要/a 依据/n]*；另一种是模板右边界在“的”字后（[v [vp|np]的 [np|vp]），如：*[看见/v [拳头/nP 队长/n]的/u 拳头/n]*。处理该结构需要专门的排歧方法，目前的算法没有这种排歧功能，现在的处理是对这种结构都按照“[v [vp|np]的 [np|vp]”结构提取。因此忽略了“[v [vp|np]”结构的模板，使得‘v+np’、‘v+vp’动词模板的识别错误率都大大提高。

以上几类问题都是影响‘np’、‘vp’识别的重要因素。从表 3 易看出“v+np+vp”模板的正确率和召回率都比较低。由于该类模板既包含‘np’又包含‘vp’，因此，前两类模板的识别效果直接影响这类模板的实验结果。目前以最长名词短语识别为切入点的不利于解决 np、vp 间这种复杂的关系。

最后需要说明的是，由于我们词语块的预处理没有经过人工校正，因此存在一定的错误率（3.14%），对实验结果有一定的影响。

5 未来工作

针对“实验结果分析”部分所发现的各类问题，下一阶段将继续完善算法：在对名词短语和动词短语内部结构进一步分析的基础上，加强动词短语在识别中的地位；采用移进-规约分析机制，提高算法的效率；把实验结果建成词汇知识库，使得确定的词汇搭配信息既可以用于排歧，又可以对搭配模板的自动识别器起指导作用。

参考文献

[1][Agresti,1990]A.Agresti.“Categorical Data Analysis”.New York,NY:John Wiley and Sons,1990.

[2][BBI,1986] Benson.M.,Benson.E.and Ilson.R. “The BBI Combinatory Dictionary from Corpora”, Proc.Of 5th conference on Applied Natural Language Processing,1986.

[3][Ushioda *et al.*,1993] Akira Ushioda, David A.Evans,Ted Gibson, Alex Waibel.“The Automatic Acquisition of Frequencies of Verb Subcategorization Frames from Tagged Corpora”. 1993.

[4][白硕，1995]白硕.“语言学知识的计算机辅助发现”，科学出版社，1995.

[5][YZWZ,1998]俞士汶，朱学锋，王惠，张芸芸.“现代汉语语法信息词典”，1998.

[6][ZhouQiang,1996]周强.“汉语语料库的短语自动划分和标注研究”，博士研究生学位论文，北京大学计算机系，1996.

[7][ZhouQiang,1998]周强.“汉语句法规则的自动获取及应用”，博士后研究人员工作期满出站报告，清华大学计算机系，1998.

[8]Susanne GAHL,“Automatic extraction of subcorpora based on subcategorization frames from a part-of-speech tagged corpus”,1998.

附录 1　部分实验结果实例

格式说明：<文件名><句子序号><原始语料整句>

{<模板类型> <模板左边界在原始句中的位置><模板右边界在原始句中的位置><抽取出的模板实例>}（评价）

1. BAIKE002 41 [下面/n [简要/aD [介绍/v [几/m 个/qN][比较/dD [主要/b 的/u [民族/n 医药/n]的/u [发展/vN 和/c 特点/n]，/，[以/c [便于/v [了解/v [中国/nS [传统/a 医学/n]的/u 实质/n]。/。]

v_vp 16　22　[便于/v [了解/v [中国/nS [传统/a 医学/n]的/u 实质/n]

v_np 17　22　[了解/v [中国/nS [传统/a 医学/n]的/u 实质/n]

（评价：“介绍”后复杂的名词短语没能识别出来）

2. BAIKE009 151 [推动/v [化工/n 发展/v]的/u 动力/n][是/vC [工农业/n 生产/vN]和/c [人民/n 生活/n][对/p [化学品/n 的/u 需要/vN]，/，[它/rN [所/u 依靠/v]的/u 基础/n][是/vC[化学/n 、/、物理学/n 、/、数学/n]和/c[各/rB 种/qN][工程/n 技术/n] 。/。]

v_np 0　4　[推动/v [化工/n 发展/v] 的/u 动力/n]

（评价：“推动”在本句的模板应为'[推动/v [化工/n 发展/v]'，此类错误属于“[v [vp|np]的 [np|vp]”结构排歧问题]

3. BAIKE009 119 [天然/b 橡胶/n][仅/d [生长/v 于/p][热带/n 及/c 亚热带/n]地区/n]，/，[不/dN [产/v 橡胶/n]的/u 国家/n][考虑/v [战时/t[会/vM [受到/v 封锁/v]，/，[都/d [极其/dD [重视/v [建立/v 于/p][石油/n 化工/n]基础/n]上/f]的/u [合成橡胶/n 工业/n] 。/。]

v_np 3　8　[生长/v 于/p][热带/n 及/c 亚热带/n]地区/n]

v_n 11　12　[产/v 橡胶/n]

v_v 18　19　[受到/v 封锁/v]

v_vp 23　32　[重视/v [建立/v 于/p][石油/n 化工/n] 基础/n]上/f] 的/u [合成橡胶/n 工业/n]

v_np 24　32　[建立/v 于/p][石油/n 化工/n] 基础/n]上/f] 的/u [合成橡胶/n 工业/n]

（评价：“考虑”后的‘v_vp’由于‘考虑’和‘受到’之间的结构，使‘v_vp’模板没能识别出；“建立 于+…”作为‘v_np’的边界识别错误导致了“重视”后所接短语类型的判别错误。）

汉语标点符号每个自成一类。