

汉语短语标注标记集的确立^{*}

周强 俞士汶

(北京大学计算语言学研究所)

北京, 100871

【摘要】 本文提出了一个汉语短语标注的基本标记集, 并从句法功能和结构组成方面对不同短语的性质进行了深入的分析和探讨, 以期对汉语短语划分和标注的自动处理和人工校对提供一个统一的处理标准。

1. 引言

对汉语语料库的多级加工处理, 主要分为以下几个阶段: 切词、词料标注、短语结构标注、语义信息标注等。对于前两个阶段, 我们已进行了一些研究和探索, 提出了一种切词和词类标注相融合的汉语语料库多级加工方法, 取得了较好的处理效果([ZY93])。目前的研究重点, 开始转向汉语短语的自动划分和标注方法的探索上, 而这项工作的一个重要基础是确定合适的短语标记集。

在汉语中, 短语具有特别重要的地位。它的内部结构比较稳定, 往往作为一个整体和句子中的其他成分发生作用, 并且它的构造原则和句子的构造原则也基本一致, 朱德熙先生认为, “如果我们把各类词组的结构和功能都足够详细地描述清楚了, 那么句子的结构实际上也就描述清楚了, 因为句子不过是独立的词组而已”([Zhu85], P74)从这个意义上看, 汉语短语标注的研究具有很高的理论和实用价值。它的顺利完成, 将进一步进行动词格槽的填充、词语依存关系的确定、以及汉英机器翻译的研究打下良好的基础。

对于英语短语的划分(bracket)和标注, 比较大的研究项目有英国 Lancaster 大学 UCREL 的 Lancaster Treebank ([GLS87]) 和美国 Pennsylvania 大学的 Penn Treebank ([MSM93])。前者的标记集较大, 通过组织成不同的层次描述了详细的短语句法信息。而后者的标记集则较为简练, 只有 14 个句法标记, 但它的特点是增加了四个表明不同空元素(Null Elements)的标记。近几年来, 有关汉语句法标注的研究也逐渐开展起来, 清华大学进行了汉语依存语法的自动标注实验, 提出了一个依存语法标注体系([ZH94])。

本文通过吸收汉语层次分析研究的最新成果, 提出了一个用于汉语短语划分和标注的基本句法标记集, 希望为汉语短语标注的自动处理和人工校对提供一个统一的基本规范。在

* 本项研究受国家自然科学基金资助

下面的几节中,第二节简要地介绍了一下短语标记集的确定原则和基本组成,然后第三、四节详细地分析了不同短语的句法功能和结构特点,最后是结束语。

2. 短语标记集的确定

2.1 确定原则

1) 小标记集的思想:

初步设想将短语标记集的规模保持在十几个标记左右,形成一个小标记集,其主要包括了反映短语语法功能的 np, vp, ap 等标记。采用小标记集,一方面可以便于人工标注和校对大规模的语料,提高处理语料的正确性和一致性;另一方面也可以在数量较少的正确标注语料的基础上进行统计,得到较为丰富的基本短语分布信息,从而可以为短语自动标注,特别是基于统计的处理提供足够的统计数据。

2) 结构和功能相结合:

在层次分析中,直接成分(IC)切分的原则是结构、功能和意义相统一。而在汉语短语的划分和标注过程中,我们认为其中更为重要的是要依据结构和功能来确定那些词可以组合成短语,不同的短语应标上什么样的标记。事实上,自动处理和人工校对,在这两种信息的应用上是有不同的侧重点的:人工标注,比较容易利用句子中的句法功能信息确定不同短语的边界及其相应的标记。而自动处理,则只能利用不同短语的结构组合信息以及一些特征词(“了”、“很”等)信息,通过对一个词串的分析和排歧处理,得到较为准确的短语划分和标注结果。因此,短语标记集的确定也必须兼顾两者的不同特点。

2.2 基本标记集:

目前设想的短语标记集主要包括以下 15 个标记:

{np, nbar, vbar, vp, abar, ap, dp, pp, bp, tp, sp, mp, dj, fj, zj}

下面通过一些具体的例子对这些标记的性质作一下简要的说明:

- 1). np: 名词性短语, 如: 我们买的, 漂亮的帽子
- 2). nbar: 名词性准短语, 如: 工人们, 资本主义
- 3). vbar: 动词性准短语, 如: 看了一眼, 学过
- 4). vp: 动词性短语, 如: 给他一本书, 去看电影
- 5). abar: 形容词性准短语, 如: 高兴高兴, 红了
- 6). ap: 形容词性短语, 如: 特别安静, 更舒服一点
- 7). dp: 副词性短语, 如: 虚心地, 非常非常
- 8). pp: 介词短语, 如: 在北京, 被他的老师
- 9). bp: 区别词性短语, 如: 大型中型小型, 这件
- 10). tp: 时间词性短语, 如: 战争初期, 周末晚上
- 11). sp: 处所词性短语, 如: 村子里, 中国内地
- 12). mp: 数量短语, 如: 两三天, 这群

13). dj:单句句型,如:她态度和蔼 那时候,天气还很冷

14). fj:复句句型,如:如果他愿意,我就陪他去看看

15). zj:整句,如:祝你生日快乐! 你去不去?

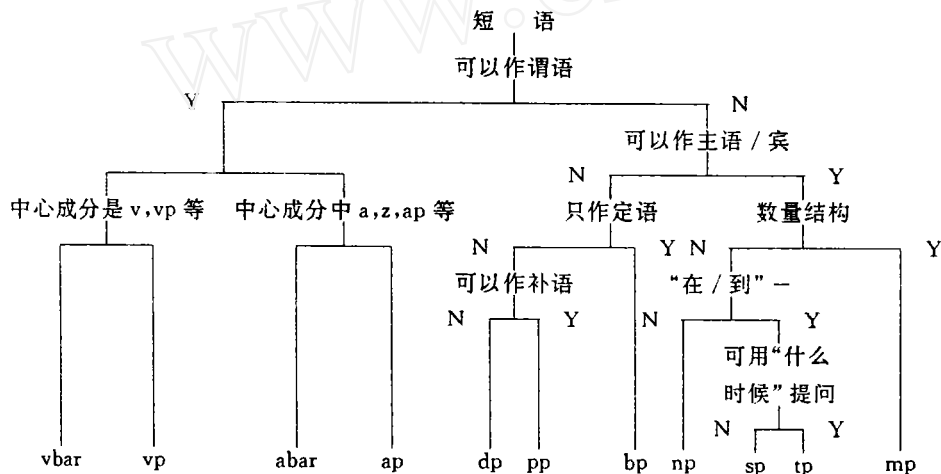
3. 短语标记的句法功能描述

3.1 功能结构分类体系:

下面的图 1 列出了区分不同短语标记的若干功能结构特征,需要特别说明的是:

1)某些短语,如:dp,bp,np,tp,sp 等,可以在功能上找到很明显的区分特征,因此可以比较容易地根据句子中的句法功能信息把它们标注出来。

2)某些短语,如:vp,vbar,ap,abar 等,在功能上没有明显的区分特征,很大程度上要依靠结构组合信息来加以区分,详见下一小节。



3.2 常见短语的基本句法功能:

1)动词性短语 vp:

在句子中主要作谓语,一般不能作补语。

2)形容词性短语 ap:

在句子中可以作谓语,也可作补语和定语。

3)副词性短语 dp:

只能作状语,不能充当其它句法成分。

4)介词短语 pp:

在句子中可以作状语、定语和补语,但不能充当其他句法成分。

5) 区别词性短语 bp:

在句子中只能作定语,不能充当其它句法成分。

6) 名词性短语 np:

在句子中可以作主语、宾语,但一般不能作谓语,不能作“在/到”的宾语。

7) 时间词性短语 tp:

在句子中可以作主语、宾语,可以作“在/到”的宾语,并能用“什么时间”提问,但一般不能作谓语。

8) 处所词性短语 sp:

在句子中可以作主语、宾语,可以作“在/到”的宾语,可以用“哪儿”提问,但不能作谓语。

9) 数量短语 mp:

在句子中可以作主语、宾语、定语,有时也可作述补结构的补语。

4. 不同短语的结构特点

本节的研究旨在从结构上找到不同短语的组成特点,为短语的自动划分和标注提供更多有用的信息。

1) 名词性短语 np:

汉语中 np 的最常见形式是以名词 $n^①$ 或名词性短语 np 为中心成分的定中结构、联合结构或同位结构短语。常见的组合有包括:

① 定中结构:

■ $a|b|f|i|j|l|m|n|r|s|t|v|z+["的"]+n^②$

■ $r|t|mp|sp+np$

■ $xp^3+“的”+n|np$

■ $dj+“的”+n|np$

② 联合结构

■ $n+\{n\}_+:$ 铅笔橡皮

■ $n+\{n\}_+c+n:$ 父母与孩子

■ $np+\{np\}_+$

■ $np+\{np\}_+c+np$

③ 同位结构

■ $n+n|r:$ 诗人李白 父亲自己

① 不同词类标记的性质,可参看附录:汉语基本词类标记集

② 结构规则符号说明:

‘|’,或操作

[X]:表示 X 是可选的

[X].:表示 X 可以重复 0 到 n 次

[X]₊:表示 X 可以重复 1 到 n 次

③ xp:表示可以取各种常见短语,主要包括 vp,np,ap,mp 等

■ r+m|n|r|mp|np:你俩 我们学生 他们三个

值得注意的是,汉语中有些形容词和动词,也能受定语修饰而组成定中结构的 np,如:

■ r+“的”+a:我的光荣

■ n+[“的”]+v:儿童教育 问题的提出

■ a|r+“的”+v:周密的调查 我们的讨论

另外,汉语的 np 还包括了两类特殊结构的短语:“的”字结构和“所”字结构

■ a|b|f|n|r|s|t|v|z+“的”

■ xp+“的”

■ dj+“的”

■ “所”+v:所得 所惧

2) 名词性准短语 nbar:

其结构组合主要包括以下几种情况:

① 名词加后缀“们”,表复数,如:工人们,学生们

② 双音节名词的紧密组合体,语法上可以认为是一个词,如:资本主义,知识分子

3) 动词性准短语 vbar:

在我们的标注体系中,vbar 短语的结构组成主要包括以下几种情况:

① 动词的不同重叠式,反映了一定的句法意义:

■ v+“了”+v:看了看

■ v+“了”+“一”+v:听了一听

■ v+“一”+v:走一走

■ v+v:运动运动(双音节动词重叠式)

■ v+“不”|“没”+v:去不去 讨论没讨论

② 单音节动词与单音节动词或单音节形容词的紧密组合体,表示一个完整的意义,如:
忍住 吃饱

③ 动词与趋向动词的组合,如:送进 捧起

④ 与助词“了、着、过”的组合,反映了句子的不同语态(aspect):

■ v+“了、着、过”:看过 学习了

■ vbar+“了、着、过”:运动运动了 忍住了

⑤ 动词前接介词“被、给”,表被动意义:

■ p+v|vbar:被打败 给晒死

⑥ 动词与介词的紧密组合体,如:来自 作用于 打在(了)

4) 动词性短语 vp:

汉语中基本的 vp 是以 v、vbar 或 vp 为中心成分的述宾结构、述补结构、状中结构、连动结构和联合结构短语。其基本组合包括:

①述宾结构:

■ $v|vbar+n|r|v|a$

■ $v|vbar|vp+xp$

■ $v|vbar+dj$

②述补结构:

■ $v+pp$: 汇集在广场上

■ $v+“得”|“不”+a|v$: 听得懂 看不见

■ $v+“得”+ap|vp$: 检查得那么彻底 气得发了疯

■ $v+“得”+dj$

③状中结构:

■ $d|dp+vp|vbar$: 正在起床 大大方方地坐着

■ $pp+vp$: 从上海来上学

■ $tp|sp+vp$

④连动结构:

■ $v+v|vp$: 来学习 去看电影

■ $vp|bar+vp|vbar$: 乘火车去北京 倒水喝了 走过去看了看(这里的 vp 一般为述宾结构)

⑤联合结构:

■ $v+\{v\}.[+c+v]$

■ $vbar+\{vbar\}.[+c+v]$

■ $vp+\{vp\}.[+c+vp]$

另外,还有一些复杂结构的 vp ,如:

①兼语结构:

■ $v+np+v|vp$: 请领导考虑 通知职工学习文件

②双宾结构:

■ $v|vbar+np+np$: 送给我们三本书

③复谓结构:

■ $v|vbar+a|z|ap$: 闻着香喷喷 看上去挺年轻

5)形容词性准短语 $abar$:

它主要包括了以下几种情况:

①双音节形容词的 ABAB 重叠式,如:高兴高兴

②形容词与某些趋向动词组成的结构,如:成熟起来

③形容词加助词“了、着、过”组成的结构,如:红了,

6)形容词性短语 ap :

汉语中常见的 ap 是以 a 、 z 或 ap 为中心的状中结构和述补结构,还有一些述宾结构及

联合结构等:

①状中结构:

- d|r+a:很安全 这么安静
- a+a|z:绝对可靠 突然通红
- dp+a:非常地可靠
- d+ap:更快一点

②述补结构:

- a+“极了”:漂亮极了
- a+“得”+“很”:机灵得很
- a+“得”+vp|ap:激动得直挥胳膊 黑得很可怕
- a+“得”+dj:热得大家都喘不过气来

③述宾结构:

- a|abar+m|q:快了点 满意些 少了许多
- a|abar+mp:交割了三天

④联合结构:

- a+{a}. [+c+a]:机智勇敢 伟大光荣和正确
- z+{z}. [+c+z]:糊里糊涂慌里慌张
- ap+{ap}+:更精密更灵活

另外,我们把“似的”结构也归入 ap 中,这里的“似的”结构是有助词“似的”(包括“般”、“一般”、“样”、“一样”等)附着在实词(或短语)后边构成的结构,常见组合有:

- n|a|v+“似的”等:石头似的 死一般
- np|vp|ap+“似的”等:象得了气管炎似的
- dj+“似的”等:箭出弦一般

7)时间词性短语 tp:

在我们目前的标注体系中,时间词性短语 tp 主要包括:

①以时间词为中心的定中结构:

- t|n+t:宋朝初期 中国古代
- vp|mp+t(“以前”|“以后”):临走以前 三天以后
- dj+t(“以前”|“以后”):他来了以后

②附加方位词(“前”、“后”、“之前”、“之后”)等的方位结构:

- mp|vp|dj+f:几天前 上班以后 她吃饭以前

③联合结构:

- t+{t}. [+c+n]:昨天今天和明天
- tp+{tp}. +c+tp

8)处所词性短语 sp:

常见结构是以处所词或方位词为中心成分的定中结构和方位结构短语:

①定中结构:

■d+f:正前方 最上面

■n+s:中国内地

②方位结构:

■n|r|np+f:树林里 他左面 木头桌子上

9)介词短语 pp:

汉语中 pp 的结构比较规范,其基本组合包括:

■p+n|r:根据原则 把我们

■p+v:被打败

■p+np|sp|tp

10)数量短语 mp:

最基本的结构是数词与量词的不同组合:

■m+q:一头 这台

■m+a+q:一大杯

■m+q+m:三米五 五十上下

■d|r+mp 大约五十岁 这三个

11)副词性短语 dp:

最常见的结构是‘地’字结构以及一些以副词为中心成分的状中结构或联合结构:

①‘地’字结构:

■a|d|i|n|v|z+“地”

■ap|mp|vp+“地”:非常仔细地 一步一步地 不停顿地

②状中结构:

■d+d|r:并没有 不怎么 不太

③联合结构:

■dp+{dp}.*[+c+dp]

12)区别词性短语 bp:

主要包括了以下两种组合情况:

①由区别词组成的联合结构,如:急性慢性 大型中型小型

②指示代词与量词组成的指量结构,如:这台 那群

★注:以上常见短语的结构描述主要参阅了([FX91],[LZY91],[PYQ92],[WL92])等资料

13) 单句句型 dj:

这是最基本的句型组合情况。它包括最为常见的主谓结构,由状语加上主谓结构形成的状中结构及包含有连词的结构。且句尾无表示语调的标点符号或语气词。其常见组合包括:

- np+np: 这个厂就一辆车
- sp+np: 桌前两三把小沙发
- np+mp: 这个学生十八岁
- np+ap: 今天的比赛太精彩了
- sp+ap: 教室里非常热闹
- vp+ap: 锻炼身体很重要
- dj+ap: 他担任组长比较合适
- np+vp: 王老师教我们体育
- sp+vp: 台上坐着主席团
- dj+vp: 他不来也可以
- np+dj: 这儿的战士嘴唇都干了
- sp+dj: 山上红旗在飘扬

★注:汉语单句的基本句型信息主要参阅了([YYXY89])

14) 复句句型 fj:

是有多个单句通过连词或标点符号连接而成的。主要包括由平行的分句组成的平行结构复句(并列、递进、选择等)以及由主从复句组成的主从结构复句(假设、因果、转折、条件等),句尾也没有标点或语气词。

15) 整句 zj:

是由词、短语、单句句型或复句句型加上语调,即句尾标点符号或语气词构成的。其结构组合主要包括:

- n|a|v[+y]+. |? |!: 证件!
- np|ap|vp[+y]+. |? |!: 多么美丽的城市啊!
- dj|fj+y: 他去了
- dj|fj+. |? |!: 他态度端正。
- zj+y: [[他去了]吗]
- zj+. |? |!: [[我明白了]。]

5. 结束语

本文提出了一个汉语短语标注的基本标记集,并从句法功能和结构组成方面对不同短

语的性质进行了深入的分析和探讨,以期为汉语短语划分和标注的自动处理和人工校对提供一个统一的处理标准。

当然,目前的标记集还只是一个基本设想,有关的信息还需要在对大规模汉语语料的短语标注实践中不断地加以补充和完善。

参考文献:

- [FX91]范晓(1991).《汉语的短语》,商务印书馆
- [GLS87]Garside. R. Leech. G. and Sampson. G. (1987). *The Computational Analysis of English*. Longman.
- [LZY91]李子云(1992).《汉语句法规则》,安徽教育出版社
- [MSM87]Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini (1993). "Building a Large Annotated Corpus of English: The Penn Treebank", *Computational Linguistics*, 19(2), 313-330
- [PYQ92]房玉清(1992).《实用汉语语法》,北京语言学院出版社
- [WL92]吴竞存,梁伯枢(1992).《现代汉语句法结构与分析》,语文出版社
- [YYXY89]北京语言学院句型研究小组编,“现代汉语基本句型”,连载于《世界汉语教学》1989(1,3,4),1990(1),1991(1)
- [Zhu85]朱德熙. (1985).《语法答问》. 商务印书馆
- [ZH94]周明,黄昌宁(1994).“面向语料库标注的汉语依存体系的探讨”,中文信息学报,8(3),35-52
- [ZY93]周强,俞士汶,(1993).“一种切词和词性标注相融合的汉语语料库多级加工法”,陈力为主编,《计算机研究与运用》,北京语言学院出版社,126-131.

附录:

汉语基本词类标记集:

形容词	a	语素字	g	数词	m	量词	q	标点符号	w	动词性语素字	Vg
区别词	b	前接成分	h	名词	n	代词	r	非语素字	x		
连词	c	成语	i	专有名词	ng	处所词	s	语气词	y		
副词	d	简称略语	j	指人的专有名词	ngp	时间词	t	状态词	z		
叹词	e	后接成分	k	象声词	o	助词	u	形容词性语素字	Ag		
方位词	f	习用语	l	介词	p	动词	v	名词性语素字	Ng		

DEFINITION OF THE TAGSET FOR ANNOTATING CHINESE PHRASE

Zhou Qiang, Yu Shiwen

Institute of Computational Linguistics, Peking University

Beijing 100871

ABSTRACT: In this paper, we propose a syntactic tagset for annotating Chinese phrase, and discuss the syntactic functions and constituent structures of the different kinds of phrase tags. We hope that this research work can be developed to become a working standard for bracketing and annotating Chinese phrase.