汉语句子的主题－主语标注

陈小荷　（北京语言学院语言信息处理研究所）
石定栩　（香港理工大学中文及双语学系）

摘要： 为了将汉语文本中的句子分成主题句、主谓句和其他句子三种类型，我们的做法是：先按标点将文本分割为一个个“准子句”，对准子句标注核心谓词以及核心谓词之前的体词性短语；然后据此重新划分复句中的各个分句，标注出各个单句或分句的句法性质，作为句子分类的基础。

本文主要介绍准子句核心谓词的标注方法以及核心谓词之前的体词性短语的标注方法。

关键词： 主题-主语　子句　主要动词　谓词前体词性短语

To Mark Topic and Subject in Chinese Sentences

Abstract: In a language typology project currently underway, it becomes necessary to mark the topic and subject of all sentences in a large corpus and then classify the sentence types accordingly. The process is to mark the clauses, the main verb of each clause and all the noun phrases in front of the main verb. The marking system for main verb and pre-verb NPs will be discussed in this paper.

Key Words: topic-subject, clause, main verb, pre-verb NPs,

一、引言

我们正在做“汉语主题句和主谓句比率统计”的研究，需要把汉语文本中的句子分成主题句、主谓句和其他句子三种类型。就单句而言，这种分类比较容易。我们可以先看句子中有无核心谓词：如果有核心谓词并且它前面有两个或两个以上的体词性短语，或者有介词结构“关于…”、“对于…”等，可认为是主题句；如果核心谓词之前只有一个体词性短语，则认为是主谓句；如果没有核心谓词或者核心谓词之前没有体词性短语，就算是其他句子。复句可划分为若干个分句，但是汉语复句中的分句怎样区分主题句和主谓句，还没有形成比较统一的看法。我们目前是这样处理的：

（一）复句的第一个分句通常是在主题链中起引介作用，可根据单句分类办法来确定它是主题句还是主谓句；

（二）复句的其他分句一般是对前面引入的主题所作的陈述，在这些分句中，主题常常承前省略，或者用代词复指，因此即使核心谓词之前没有体词性短语也可以算是主题句；但是如果后续分句的核心谓词之前出现的是另一个体词性短语，则认为是主谓句；

（三）偏正关系的两个分句算一句，根据“正”那一部分来分类。

关于主题链问题的讨论，可参看文献［1］、［2］。本文主要介绍跟该问题相关的句法自动标注方法，涉及三个方面：第一，将文本分割为一个个单句或分句；第二，标注单句或分句中的核心谓词；第三，标注出核心谓词之前的体词性短语。由于复句中分句的分割必须以核心谓词和体词性短语的标注为前提，因此我们的实际做法是：先按标点（逗号、分号、冒号、句号、问号、感叹号等）将文本分割为一个个“准子句”；接着标注准子句中的核心谓词（零个或一个）；然后标注核心谓词之前的体词性短语（零个或若干个）；
最后据此重新划分复句中的各个分句，没有核心谓词的准子句并入下一个准子句，同时标注出各个单句或分句的句法性质（有无核心谓词，核心谓词之前有多少个体词性语，单句还是分句，是第几个分句），作为句子分类的基础。例如：①

(1) SP0: [我们] 必须坚决｛贯彻｝“科学技术进步必须面向经济建设和社会进步；经济建设和社会进步必须依靠科学技术进步”的基本方针。

(2) SP1: 在经济建设中，[科技界] 必须为解决工、农业和第三产业的发展过程中所出现和遇到的关键问题 ｛提供｝技术措施，

P2: 必须大力｛推广｝先进适用的科技成果，

P3: 必须｛加快｝高技术产业的发展速度。

(3) SS1: [最年轻的军长] [45岁]，

SS2: [最年轻的师长] [38岁]，

SS3: [最年轻的团长] [32岁]。

(4) SSP1: [过去没有飞过的高难度课目]，[他] 带头｛飞｝；

SSP2: [过去没有闯过的“禁区”]，[他] 带头｛闯｝；

SSP3: [过去没有创下的纪录]，[他] 带头｛创｝。

我们所使用的是北京语言文化大学语言信息处理研究所提供的现代汉语语料，目前规模为50万字。这些语料都已经标注词性，词性标记共112种，对于词的句法功能区分比较细致，有利于句法分析。例如，在体词性短语中充当修饰语（不带“的”）或中心语的动词有特殊的标记，跟其他动词区别开来，大大减轻了识别核心谓词的负担，标注体词性短语也比较方便。

二、核心谓词标注

在实验阶段，我们先对35篇文本手工标注了核心谓词。这35篇共有准子句5008个。标注的原则是：每个准子句最多标注一个核心谓词；准子句中如果没有谓词、或虽有谓词但不是整个准子句的核心，则不标核心谓词。

国内有采用基于规则的方法来判别和确定句子的中心谓语及其相应边界的，可参考文献［3］。严格地说，只有对句子作全面的句法分析之后，才有可能准确地标注核心谓词，但是在自动句法分析取得突破性进展之前，还难以做到对句子做全面的句法分析。因此，我们退而求其次，通过对准子句作线性序列的分析来尽可能地“猜测”核心谓词。一个谓词在准子句中是否核心谓词，跟以下因素有关：

（一）谓词本身的类别，某些谓词比别的谓词更经常充当核心。在112种词性标记

中，谓词有以下１５种，后附的数字是充当核心的次数与出现次数之比，记作 HF(V)：②

a	形容词	0.060	ab	带宾语的形容词	0.314
z	状态词	0.083	vg	一般动词	0.322
va	助动词	0.780	vf	形式动词	0.559
vi	连系动词	0.814	vv	动词之前的趋向动词	0.400
vgj	带兼语的动词	0.748	vgs	带小句宾语的动词	0.794
vgv	带动词宾语的动词	0.564	vga	带形容词宾语的动词	0.467
vgn	带体词宾语的动词	0.527	vgd	带双宾语的动词	0.750
iv	谓词性成语	0.308			

（二）谓词之前状语的个数、谓词之后的结果补语（语料中标为 vc ）或动态助词的个数。经简单的统计可以看出，挂在谓词上头的这种“零碎”越多，谓词充当核心的概率越高。不过，介词结构状语可以很长，其内部层次可以很复杂，可能包含别的谓词，这时就很难确定是哪个谓词的状语了。

（三）谓词前后的结构助词“的”，谓词类别相同时，离“的”越近，充当核心的可能性就越低。当然也要看夹在中间的是什么词性标记，例如在谓词与“的”之间的 vc （结果补语）跟 ng （一般名词）就大不一样：前一种情况下，谓词不可能充当核心；后一种情况下，谓词有可能充当核心。另外， kn （名词后缀）会取消紧邻其前的谓词充当核心的可能性。

我们曾经采用规则描述方法来识别核心谓词。实验表明仅凭研究者内省而写出的规则很难覆盖真实语料中的种种复杂情况，因此改而采用统计方法。我们相信，只要语料规模足够大，所考察的谓词语境（谓词的上下文）足够宽，统计数据是能够在相当程度上反映谓词充当核心的可能性大小的。

关于谓词语境，我们只考察了谓词前后各两个词的词性，分别记作 L2 、 L1 和 R1 、 R2 ，语境不足时（例如首词或尾词是谓词），以虚设的词性标记来填补。实验表明，四个词的语境宽度，对于核心谓词标注来说是比较合适的：再宽一些，需要花费更多的时间和空间，对标注正确率的提高没有明显作用；语境太窄，标注正确率会明显下降。这里用 HF(L2,L1,V) 表示前两个词性标记跟谓词同现时该谓词充当核心的频率， 用 HF(V,R1,R2) 表示谓词跟后两个词性标记同现时该谓词充当核心的频率。 文本标注时，谓词的评分的计算公式为：

SCORE = HF(V) * HF(L2,L1,V) * HF(V,R1,R2)　　（１）

标注新语料时， HF(L2,L1,V) 或 HF(V,R1,R2) 可能会因为出现新的语境而使得其值为零，这时我们假定该谓词在新的上文或下文中充当核心的概率为 0.5 。

按公式（１）计算准子句中每一个谓词的得分，选择其中得分最高并且大于预定阀值（根据实验，阀值定为 0.1 ）的作为核心谓词。如果没有谓词，或者没有一个谓词的得分超过阀值，就不标核心谓词。

从词性序列的某些特征来判断，有的准子句几乎总是没有核心谓词，例如，首词为介词并且尾词为方位词的、以逗号结尾的准子句通常是一个介词结构。对于这种情况，我们在谓词评分之前先进行这类判断以提高标注效率。

一般用召回率和正确率两个指标来衡量标注效果，召回率反映“正本”中有百分之几的标记在“试本”中也出现，正确率则反映“试本”中有百分之几的标记可被“正本”确认为

正确。但是由于我们是每个准子句最多只标注一个核心谓词，对于没有核心谓词的准子句应该不标，不标也是一种标注，所以只需计算正确率：对于每一个准子句，“试本”与“正本”的标注完全一样就是正确的，否则就算错误。正确标注的准子句个数除以所有准子句的个数即为核心谓词标注正确率。

实验阶段之后，我们对５０万字语料分２４批进行标注，每批一般为１０个文本。第一批是先手工标注，接着进行训练，然后作封闭测试；以后各批次都是先利用前面得到的训练数据进行开放测试，接着手工改正标注错误，再进行训练，然后进行封闭测试。测得以下数据：

批次	开放测试	封闭测试	批次	开放测试	封闭测试
1		0.9695	13	0.8898	0.9391
2	0.8307	0.9673	14	0.8958	0.9560
3	0.8771	0.9632	15	0.9158	0.9730
4	0.8642	0.9214	16	0.9103	0.9552
5	0.8797	0.9669	17	0.8916	0.9511
6	0.8847	0.9446	18	0.8973	0.9593
7	0.8903	0.9490	19	0.8809	0.9259
8	0.8983	0.9700	20	0.8951	0.9338
9	0.9176	0.9753	21	0.9258	0.9543
10	0.9042	0.9617	22	0.9099	0.9585
11	0.8635	0.9412	23	0.9193	0.9529
12	0.8941	0.9523	24	0.9306	0.9563

从以上数据可以看出，开放测试与封闭测试的正确率逐渐接近，开放测试的正确率最后稳定在９０％左右。

我们也曾把谓词跟它的上文和下文综合起来考虑，即用如下计算公式：

SCORE = HF(V) * HF(L2,L1,V,R1,R2)　　（２）

其中　HF(L2,L1,V,R1,R2)　表示谓词跟前后各两个词性标记同现时充当核心的频率。实验表明，上文跟下文的相关程度不高，用公式（２）虽然可以稍稍提高封闭测试的正确率，但由于对语境限制过细，开放测试的正确率大大降低。

三、体词性短语标注

目前我们只标注了核心谓词之前的体词性短语，如果准子句中没有核心谓词，则把其中作直接成分的体词性短语都标注出来。

文献［４］报道了一项采用统计方法确定中文最长名词短语的左右边界的研究，并且承认该方法的效果不很理想。我们的工作与之相似，但由于只要标出核心谓词之前的体词性短语，因此任务相对单纯一些。

我们采用基于规则的方法和“双向逼近”的策略来标注核心谓词之前的体词性短语：第一、从准子句的开头向右逼近体词性短语的开头，这主要是要跳过句首的连词、惯用语、层

序标志等，例如：

⑸ 二 、 创新之处〔应〕是正确的。

--- →

第二、从核心谓词（若无核心谓词则从句末）开始向左逼近体词性短语的结尾，这主要是要跳过核心谓词的状语（包括副词、形容词、介词结构、“X地”状语等），例如：

⑹ 空军广大官兵 正在 用自己的智慧和汗水 为人民空军现代化建设〔创造〕着更加辉煌

← --- --- --- --- --- --- --- --- --- --- --- --- --- ---

的未来。

第一个逼近比较容易，虽然主语之前也可能出现介词结构等状语，但是这种状语往往带逗号，已被分割为一个准子句，所以不存在太大困难。第二个逼近，由单词（副词、形容词、时间词、处所词）充当状语的情况也容易处理，因为它们往往是修饰后边遇到的第一个谓词，一般不会“越位修饰”；关键是要识别出充当状语的词组。这些词组可以分为两类，一类是左边界确定，右边界开放，介词结构就是这种情况。另一类是右边界确定，左边界开放，包括以下几种：

（一）“X地”，助词“地”是右边界，如“有组织有计划地”；

（二）方位短语，方位词是右边界，如“党的十一届三中全会之后”；

（三）“X以来”，其中“以来”或“来”是右边界，如“近几年来”，“1991年以来”；

（四）时间词短语，时间词是右边界，如“八五期间”；另外，有些词和时间词功能相同或相似，语料中标作名词，如“时”和“时候”，我们把这些词作右边界形成的短语作为时间词短语来处理，例如“实现产业化时”。

这四种左边界开放的状语中，“X地”状语比较复杂。我们从语料中归纳出以下几种模式：

vgn ng vgn ng usdi	有 计划 有 组织 地	vg usdi	担心 地
nx qv mx qv usdi	一 下 一 下 地	ng usdi	民主 地
ng vgv vgb usdi	心 存 敬畏 地	dr usdi	不时 地
dd a ng usdi	最 大 限度 地	id usdi	不约而同 地
vgn ng usdi	有 意识 地	iv usdi	一如既往 地
vg kn usdi	建设 性 地	b usdi	无形 地
dd a usdi	很 好 地	a usdi	爽朗 地
a ng usdi	大 跨度 地	z usdi	疾速 地
a a usdi	及时 客观 地		

向左逼近时，一见到“地/usdi ”就向左作最大匹配以确定“X地”状语的左边界，这样只要匹配模式没有遗漏（有遗漏可以随时补充），一般不会有问题。

介词结构通常是作状语，用来修饰动词的，③ 介词结构中常常内嵌一个方位短语或时间词短语，例如：

⑺ 在 科技经济发展中

⑻ 在 会见两院院士时

这种介词结构的左边界和右边界都是确定的，处理起来比较方便。关键是要把这些边界

标记词（介词与方位词、时间词等）的共现关系都找出来，并且给出一个共现概率，以便当有若干个右边界标记发生竞争时作出恰当的选择。另外，介词与某些名词、介词与某些动词也有很高的共现概率，如果能够把这些共现关系都找出来，对于介词结构的右边界的确定是很有用处的。当然，如果有一个现成的较大规模的汉语树库（ Treebank ），那就很好办了。目前的条件下，我们的做法是：统计语料里的每个介词与其后的非邻接词（包括标点，不超出准子句范围）的共现次数。然后计算这些搭配的概率和互信息（ Mutual Information ），用共现概率与互信息的乘积来表示这些搭配的强度，结果比较理想。例如，在我们统计的５０万字语料中，跟介词“在”搭配强度最高的１２个非邻接词是：

中　的　上　，　下　。　方面　、　时　和　内　里

跟介词“为”搭配强度最高的１２个非邻接词是：

的　。　提供　服务　了　和　创造　，　、　而　奋斗　贡献

跟介词“根据”搭配强度最高的１２个非邻接词是：

的　，　情况　和　特点　需要　七　规定　精神　国情　制定　意见

这三个介词代表了三种类型：“在”是跟方位词搭配强度高，“为”是跟动词搭配强度高，“根据”是跟名词搭配强度高。

显然，搭配强度高的非邻接词中，也有一些是与预示右边界无关的，例如顿号和句号纯粹是由于它们本身出现概率高，数词“七”可能是由于语料的特殊性。又如助词“的”对介词“对”、“在”和“为”在不同程度上有预示右边界的作用，“的”字前面那个词可能是右边界，但介词结构“根据…”不可能再加助词“的”形成“的”字结构。因此有必要考虑汉语的句法规则，对每个介词的非邻接词从词性上加以过滤。例如，对以上三个介词所能搭配的词类，分别规定如下：

“在” t（时间词） s（处所词） f（方位词） n（名词）
v（动词） ur（“来”、“来说”之类的助词）

“为” v（动词）

“根据” v（动词）n（名词），（逗号）

于是，我们确定介词结构的右边界的方法是：从介词之后的第一个非邻接词开始自左向右扫描，找到属于该介词所能搭配的词类并且共现概率与互信息乘积最高的那个词，如果它是名词、时间词、处所词、方位词或“来”、“来说”之类的助词，则它本身就是右边界；如果它是动词、助词“的”或标点等，那么它左边的那个词是右边界。

现在的任务是要确定核心谓词之前的体词性短语的右边界，算法如下：

1. 假定 NPR（核心谓词之前的体词性短语的右边界）是在紧邻核心谓词的左边，当前词指针为 PI=NPR；
2. 自右向左扫描，直至 PI>NPL（核心谓词之前的体词性短语的左边界）：
   2.1 如果 PI 是单词形式的状语，则将 NPR 和 PI 同时向左移动一个位置；
   2.2 否则，如果 PI 是“地”类标记词，则利用模式匹配方法找到该状语的左边界，然后将 NPR 和 PI 同时移到该状语之前；
   2.3 否则，如果 PI 是介词，则利用上述方法找到介词结构的右边界：
      2.3.1 如果右边界跟 NPR 邻接，则表明该介词结构是修饰核心谓词的状语，将 NPR 和 PI 同时移到该介词之前；

2.3.2 否则表明该介词结构不是修饰核心谓词的状语，而是核心谓词之前的体词性短语内部的成分，中止扫描；

2.4 否则仅将 PI 向左移动一个位置；

3. 如果 NPR<NPL，表明核心谓词之前不存在体词性短语；否则可对 NPL 和 NPR 之间的部分做进一步分析，以确定其中有几个体词性短语（非并列关系的，例如大主语和小主语、主语和体词性谓语等等）。

例(6)从核心谓词开始向左跳过了三个状语（介词结构“为…”、“用…”和副词“正在”）。又如：

(9) 因机械原因发生的严重飞行事故万时率 已｛接近｝世界先进水平。

← ---

其中介词结构“因…”的右边界是“原因”，跟 NPR 所指的词项“率”不邻接，所以判定为不是修饰核心谓词的状语。

仅根据标点划分准子句，有时会割裂句法关系。在核心谓词标注后的校对中，凡遇到这种情况我们都对准子句的划分作适当调整，在此基础上再作体词性短语标注。采用这种办法，有利于正确识别核心谓词之前的非常复杂的体词性短语。

我们对九个文本进行了核心谓词之前的体词性短语标注的实验。分析实验结果时仍然只计算正确率，即跟“正本”完全一样才算正确，正确率为９２％ 。

参考文献

［１］ Dingxu Shi, Topic Chain as a Syntictic Category in Chinese, Journal of Chinese Linguistics, Vol. 17, No. 2

［２］ Dingxu Shi, Topics and Topic Comment Constructions in Chinese

［３］罗振声、孙长捷、孙才，汉语句型成分自动分析中谓语识别策略的研究，，《计算语言学进展与应用》，清华大学出版社，１９９５

［４］周明，基于语料库的中文最长名词短语的自动抽取，《计算语言学进展与应用》，清华大学出版社，１９９５

［５］陈小荷，崔永华，关于建立大规模汉语树库的设想，《计算机时代的汉语和汉字研究》，清华大学出版社，１９９６

附注

① 符号说明：“S”表示体词性短语，如果核心谓词之前有多个体词性短语，用“SS”表示，句子中体词性短语加方括号；“P”表示有核心谓词，句子中核心谓词加花括号；数字“0”表示单句，“1”表示是复句中的第一分句，其余类推。

② 但在实际标注中，kv（动词后缀）有时也标为核心谓词。

va 或 vv 后边常常或总是出现实义谓词，哪个算核心谓词，容或有不同看法；为了以后标注体词性短语的方便，我们一律把出现在前面的 va 或 vv 标作核心谓词。

③ 有几种介词结构如“对…”、“在…”可加“的”作定语。