汉语词类标注规则的获取技术*

周莉娜　郑家恒　刘开瑛
山西大学计算机科学系

摘要：本文是在手工标注 10 万语料的基础上，通过提取逻辑规则，并利用对语料的统计，总结归纳出汉语词类标注的三种规则的获取技术。用获取的规则对另外 20 万语料进行测试，效果较为满意。

一、引　言

用基于规则的方法，对英语进行词类自动标注，国外已经建立了实用系统。最典型的是七十年代美国 Brown 大学设计的 TAGGIT 词类自动标注系统。它采用了 86 种词类标记和 3300 条上下文框架规则，用来对 Brown 百万词语料进行词类标注，正确率达 77%。我国对汉语语料库自动标注的研究起步较晚。目前还处于探索研究阶段，使用规则集的方法尤其不成熟。一方面，我国还没有适合汉语词类规则集的建造方案；另一方面，用规则的方法对汉语语料库进行自动标注的正确率到底能达到多少，还未被尝试过。

从 1990 年起，我们山西大学语言信息处理技术研究组确定了《汉语语料库语法标记》，建立了词类自动标注实验系统，选取了覆盖政务、法规、科技、科普、新闻五大类 30 万语料。其中手工标注 10 万，并对其余 20 万语料进行了自动切分和自动标注。在标注系统的开发和使用过程中发现，用规则集和统计方法相结合的技术对新词和兼类词进行词类推测是必要的。

本文正是在实际标注的基础上，提取了词类标注规则，对汉语语料库进行自动标注作了初步探讨。本文重点介绍词类标注规则的获取技术。规则的获取技术大致可分为三类：1. 取自语料中与上下文无关的词语结构。2. 依据语料中的与上下文有关的语境。3. 来源于对源语料的概率统计。

通过使用上述技术获取的规则，我们对 30 万语料进行了测试。正确率为：封闭测试 95.5%、开放测试 83%。单独用规则标注的词，正确率为：封闭测试 85%、开放测试 63%。

二、按词语结构获取的规则

在对汉语进行词类自动标注时，我们把具有两种或两种以上词类的词叫兼类词，否则为非兼类词。词表中未收进的词叫新词。据统计，兼类词的词条数占总数的 13%，但它们在语料中实际出现的词次要高得多。本文提出的规则的获取技术正是为了处理兼类词和新词。

词语结构规则取自所处理语料中的现实语法现象，分为共性规则和个性规则。共性规则是针对某一词类制定的，它是靠功能来驱动。个性规则是由某一个或某几个词语的情况来确定。它是靠词语来驱动。共性规则源于语料中词语的分布知识和词本身的构造知识。所涉及的词类有方位词、状态词、趋向动词和名词等。个性规则源于某些词特殊的语法现象。按词语结构获取的规则有时间词、数字词、词缀、重叠词、固定搭配等近 200 条。下面重点介绍词缀、重叠词和固定搭配规则。

本文所使用的词类标记系统分为 25 个大类，67 个小类（不包括标点符号）。为了便于描述规则，规定如下符号：

* 国家自然科学基金资助项目

X＝X1X2…Xm ：表示当前词（Xi 表示字）　x　：表示当前词标记

Y　　：表示 X 的前一词　　y　：表示 Y 的标记

Z　　：表示 X 的后一词　　z　：表示 Z 的标记

D　　：表示直接量　　Da ：表示直接量地址

Wx　　：表示词语 W 的标记

注：直接量为词语规则中与该词相联系的某一待查的词语。

1. 词缀规则 包括前缀规则和后缀规则。前缀规则用于处理 ABB 的重叠形式的形容词 、数量词、人名中的简称等。后缀词规则主要用于处理机关、团体名称、职务、地名等。

R1 设 K1＝｛金、银、红、黄、绿、兰、白、灰、黑｝

若 m＝3，X1∈K1 且 X2X3 为 BB 的形式，则 x＝A（形容词）。

如：兰茵茵、金烂烂。

R2 设 K2＝｛一、几｝

若 m＝3，X1∈K2，X2X3 为 BB 的形式且 Bx＝QN 或 QV，

则 x＝MM（数量词）

如：“一片片、几回回”。

R3 设 K3＝｛老、大、小｝

若（X2…Xm）x＝NPFF（姓氏）且 X1∈K3，则 x＝NPF（姓名）。

如：老李、大李、小李。

R4 设 K4＝｛老、总、工｝

若（X1…Xm－1）x＝NPFF ，Xm∈K4，则 x＝NPF。

如：李老、李总、李工。

R5 设 K5＝｛赛、酸、仪、家、学、色｝

若 Xm∈K5，则 x＝NG。（名词）

如：排球赛、核糖核酸、地动仪、心理学、灰白色等。

R6 设 K6＝｛化｝，若 Xm∈K6，则 x＝VG。（动词）

如：系统化、电气化等。

R7 设 L7＝｛然｝，若 Xm∈L7，则 x＝A。

如：欣然、愕然、飘飘然等。

2. 重叠词规则

根据分词规范，动词及形容词的重叠形式 AA、AABB 一律为一个分词单位。但是，汉语中叠词的出现无法预测，且不易一一列出。对动词、形容词及量词来说尤为如此。根据语料中叠词的构成，总结出以下规则：

R1　如果 X 为 AABB 的形式且（AB）x＝A 或 VG，则相应地 x＝A 或 x＝VG。

R2　如果 X 为 AA 的形式且 Ax＝VG 或 QN（名量词）或 QV（动量词），则相应地 x＝VG 或 QN 或 QV。

例：张张（QN）笑脸　　白帆　点点（QN）

打打闹闹（VG）　　热热闹闹（A）　让　我　看看（VG）

3. 固定搭配规则

词“为”有规则如下：

R1　设直接量为“以”

如果　Da<><空>，则 x＝VI（动词）

R2 .设直接量为“所”

如果　Da<><空>，则 x＝P（介词），Dx＝US（助词）

词“到”有规则：

R1　设直接量为“从”

如果　Da<><空>，则 x＝P

例如：①以 经线 为（VI）中央经线 ，

②南极洲 绝大部分 在 南极圈 内 ，为（P）三 大洋 所（US）环绕 ，

③下面 从 热带 到（P）寒带 依次 加以 说明 。

三、与上下文有关规则的获取

这里的上下文是依赖于源语料的真实语境。鉴于所处理语料领域的不同，有的语法现象在一般语料中并不很普遍，而在某个领域的某种语境中却出现得较为频繁。对这种总的看起来颗粒度较小的语法现象，制定了上下文有关规则。这类规则分为两种：一种是基于词语的，一种是基于标记的。基于词语的主要是兼类词规则。基于标记的规则可用来处理兼类词，也可用来处理新词。这类规则总计近 2000 条。

1、兼类词规则

因为汉语缺乏形态标记的屈折变化，这就导致了词类兼类歧义的广泛存在。

1）基于词语的规则

如：“当”字可作介词与普通动词。

R.　如果：y＝<标点符号>或<空>，则：x＝P

否则：如果：Z∈｛成、作、了、过｝，则：x＝VG，否则：x＝P

例如：①当（P）蚌 、钳 等 瓣鳃类 遇到 危险 时 ，

②把　“是”　字　当（VG）成　等于 ，

2）基于标记的规则

如：对“重、高、深”制定如下规则：

R. 如果：y＝MG（概数词）或 QN 或 MX 或 z＝MX

则：x＝NG，否则：x＝A

例如：“马里亚纳 海沟 深（NG）达 11034 米 ，是 世界 上 海洋 最 深（A）的 地方。”

3）综合运用前后词的词语和标记的规则。

例如：“上”有动词、方位词和趋向动词三种词类。

主力　沿　大渡河　北　上（VG），

在　飞机　上（FS），

飞　上（VQ）　天空。

针对上述情况，我们制定了如下规则：

R1　设直接量 D 为“在”

如果：Da<><空>，则：x＝FS，Dx＝PZ（介词）

R2　如果：y＝VG，则：x＝VQ，否则：x＝VG

2、标记搭配规则

这类规则利用了位于一个词的前和紧跟其后的词的标记信息，并在两个方向上各向前看两个标记。在运用这些规则时，很可能因后一词或后二词为空，导致不满足条件而错标，降

低了规则的效率。为此，我们专门对 span 的长度进行了统计。一个 span 由 N 个相邻的兼类词及其前后的非兼类词构成。span 中兼类词的个数 N 叫做 span 的长度。结果如下：

span 长度 百分比 范围	1	2	3	4	5	6	7	8
单　独	77. 66	17. 28	3. 52	0. 85	0. 47	0. 13	0. 04	0. 04
累 计	77. 66	94. 94	98. 47	99. 32	99. 79	99. 92	99. 98	100

由表中可看出，长度为 1 的 span 占 77. 66%，这就保证了我们在大多数情况下运用规则的有效性。此外，利用已建立的标记搭配表，我们可以处理 span 长度小于等于 3 的情况，并且满足条件的 span 占到总数的 98. 4%，所以采用这种方法是完全可行的。

如：　PZ　A　x　FS　，=>　x=NG

例如："竞争"有"VG、NG"两种标记可选，利用这条规则标为："在（PZ）激烈（A）竞争（NG）中（FS），"

又如：VY　A　x　，　=>x=NG

NG　CW　x　NG　=>x=VG

D　x　US　=>x=A

四、概率统计方法获取的规则

这种方法被用于两处。一处是在对兼类词的各种兼类情况集中归纳后，对与一个词同现的各种标记情况进行了统计、排序。其中的频率最高的标记作为该兼类词的所有其它规则不满足的最终归宿。第二处用于提取 span 规则，它是对上下文部分相关标记的概率统计。span 规则是根据 span 长度到手工标注文本中提取相应于一个 span 长度的标记串。这些 span 规则按其长度不同被分开。每种长度 span 规则集中的规则都由两边的非兼类词唯一确定。冲突由概率统计解决，结果只保留一个概率最大者。该类规则总计 1000 余条。

1）span 规则

遇到长度大于 3 的 span（长 span），标记搭配表也无能为力。这种情况虽然为数不多，但是它们中每一个所涉及的标记数量却较大。对标注正确率的影响也不容忽视。所以，有必要对它们进行专门处理。

对于长 span，同样提取出相应长度的标记串，并且保留了由语境中非兼类边界标记确定的概率最大的一个。

例如：A　x　y　YE=>x=NG，y=NG

CC　x　y　VY=>x= A，y=NG

在取 span 规则时，也存在一个问题：两个非兼类词标记之间标记串的情况可能非常杂，甚至各种情况的概率相差无几。例如：在标记"NG"与"US"之间的标记可有五种情况：

```
        / D    VG    VQ\
      /   VG  NG    NG  \
NG      A    VG    NG    / US
   \    D    VA    VG  /
     \ VG  NG    VG/
```

而且每种标记串所占比率都很低。针对这种情况，我们设置了阈值（0. 5）。当小于该值时，就把新词默认为有“NG、VG”两种标记的兼类词。

2）同现标记规则

一般来说，兼类词的各种兼类情况比重不同。特别是一个词最常用的标记最能反映它的基本词类，而出现频率较少的标记很可能是由于人工失误造成。在所有其它规则都未能确定一个兼类词的标记时，就给它选择同现概率最高的一个标记。

例如：

“过”一词共有四种标记，经统计它们出现的概率为UT（67%）、VG（18%）、D（12%）、VQ（3%），产生了下句的标注结果：

“他仔细 观察 了 从来 未 有 人 注意 过（UT）的大教堂 里 吊灯 的 晃动 情况。”

又如：对词“和”更明显，它的三种标记的出现概率为：CW（97. 8%）、P（2%）、NG（0. 36%），所以有：

“海峡 在 交通 和（CW） 战略 上 往往 具有 重要意义。”

结 束 语

用本文介绍的技术获取的规则建立的实验模型进行词类标注，已取得了较为满意的效果。我们尽量获取较全面的知识，力争在一条规则中解决一个语法功能；并希望将来建立规则词典，用统一的控制策略来处理。然而我们发现：标注的平均正确率很难进一步有较大提高，因为，一方面，自然语言非常复杂，大文本中将不断出现未经处理的语法现象；另一方面，输入中可能有噪声词，不能被正确处理。这可通过把更多的个性特征引入规则得到一些改善，但大量残留错误须通过对模型进行句法分析改正，少量的必须依赖语义信息。

参考文献

【1】R. Garside , G. Leech and G. Sampson . The computationalAnalysis of English Chapter 4：A Corpus— Based Approch，London：Longman , 1987.

【2】Johansson，S. and K. Hofland，Frequency Analysis ofEnglish Vocabulary and Grammar Based on the LOB Corpus，Vol，1 ：Tag Frequencies and Word Frequencies，Oxford：Clarendon Press，1989

【3】Steven J. DeRose，〃 Grammatical Categary Disambiguation by Statistical Optimmization〃，IJCAI'91 960—965

【4】朱德熙，语法讲义，商务印书馆，1984

【5】刘倬，论机器翻译规则系统的编制方法，语言与计算机，中国社会科学出版社，1990

【6】黄昌宁，国外语料库述评，机器翻译研究进展，电子工业出版社

【7】刘开瑛、郭炳炎，自然语言处理，科学出版社，1991

【8】汉语语料库词类标记，山西大学计算机应用研究所，1991

Acquisition Techniques of Rules for Gramatical Tagging for Chinese Corpus

Zhou Lina Zheng Jiaheng Liu Kaiying
(Dept. of Computer Science ,Shanxi University,Taiyuan 030006)

ABSTRACT

Based on gramatical tagging of 100,000 Chinese words corpus manually , by drawing logical rules combined with using statistics on corpus,this paper presents 3 kinds of techniquesof acquiring rules for automatically tagging Chinese corpus. And the results of testing on other 200,000 words corpus is satisfactory.