

自然语言理解

(7)

宗成庆

中科院自动化研究所
模式识别国家重点实验室

cqzong@nlpr.ia.ac.cn

<http://www.nlpr.ia.ac.cn/English/cip/cqzong.htm>

No.95, Zhongguancun East Road
Beijing 100080, China



<http://www.ia.ac.cn>
Tel. No.: +86-10-6255 4263

第七章 词性标注

7.1 概述

词性或称词类（Part-of-Speech, POS）是词汇最重要的特性，是连接词汇到句法的桥梁。

词性标注（POS tagging）的主要任务是消除词性兼类歧义。词性兼类在任何一种自然语言中都普遍存在。

例如，在英语中：

- 1) Time **flies like** an arrow.
- 2) I want you to **web** our annual report.

对 Brown 语料库的统计，55%词次兼类。汉语中常用词兼类现象严重，《现代汉语八百词》兼类占 22.5%。

7.1 概述

◆ 汉语中的词性兼类现象： [赵铁军，2001]

- 1) 形同音不同，如：“好 (hao3, 形容词)、好 (hao4, 动词)”
- 2) 同形、同音，但意义毫不相干，如：“会 (会议, 名词)、会 (能够、动词)”
- 3) 具有典型意义的兼类词，如：“典型 (名词或形容词)”
- 4) 上述情况的组合，如：“行 (xing2, 动词/形容词; hang2, 名词/量词)”



7.2 词性标注集的定义

□ 标注集确定原则：

不同语言的语法中，词性划分基本上已经约定俗成。
自然语言处理中对词性标记要求相对细致。

◆ 一般原则：

- 标准性：普遍使用和认可的分类标准和符号体系；
- 兼容性：与已建立的资源标记尽量一致，或可转换；
- 可扩展性：扩充或修改；

7.2 词性标注集的定义

◆ UPenn Treebank 的词性标注集确定原则：

- 可恢复性（recoverability）：从标注语料能恢复原词汇或者借助于句法信息能够区分不同的词类；
- 一致性（consistency）：功能相同的词应该属于同一类；
- 不明确性（indeterminacy）：为了避免标注者在不明确的条件下任意决定标注类型，允许标注者给出多个标记（限于一些特殊情况）。

- [Marcus et al., 1993]

7.2 词性标注集的定义

◆ UPenn Treebank 中的英语词性标注集：

NN 名词（单数或不可数）

NNS 名词复数

NNP 专用名词（单数）

NNPS 专用名词（复数）

VB 动词

... ..

共计：30种标记符号

7.2 词性标注集的定义

◆ 汉语词性标注集：

北大：名词n、时间词t、处所词s 39 种标记符号

山西大学：基本词类17类

... ..

百花齐放

7.3 词性标注方法

□ 基于规则的汉语词性标注方法

◆ 规则消歧法

➤ 非兼类词典

➤ 兼类词典

- 词性可能出现的概率高低排列

➤ 构造兼类词识别规则 - [刘开瑛, 2000]

- 并列鉴别规则

如：体现了人民的要求（N/V？）和愿望（N，非兼类）。

7.3 词性标注方法

- 同境鉴别规则

如：一个优秀的企业必须具备一流的产品（名词，非兼类）、一流的管理（N/V？）和一流的服务（N/V？）。

- 区别词鉴别规则（区别词只能直接修饰名词）

如：他们搞的这次大型（鉴别词，非兼类）调查（V/N？）历时半年。

- 唯名形容词鉴别规则（有些形容词只能直接修饰名词）

如：重大（唯名形容词）损失（N/V？）

巨大（唯名形容词）影响（N/V？）

7.3 词性标注方法

➤ 根据词语的结构建立词性标注规则

• 词缀（前缀、后缀）规则

- 形容词：蓝茵茵，绿油油，金灿灿，...
- 数量词：一片片，一次次，一回回，...
- 人名简称：李总，张工，刘老，...
- 其它：年轻化，知识化，...{化}
 篮球赛，足球赛，...{赛}

... ..

• 重叠词规则

- 看看，瞧瞧，高高兴兴，热热闹闹，...



7.3 词性标注方法

◆ 基于错误驱动的词性标注方法

- 初始词性赋值
- 对比正确标注的句子，自动学习结构转换规则
- 利用转换规则调整初始赋值

- [Brill, 1992]

7.3 词性标注方法

□ 基于统计模型的汉语词性标注方法

◆ 基于 n -gram 的语言模型

汉字串 : $Text$

词性标注符号串 : $Tags$

$$\hat{Tags} = \arg \max_{Tags} P(Tags | Text) = \arg \max_{Tags} P(Tags)$$

应用系统 : 1) 1983年 Mashall 提出的 LOB 语料库的标注系统:
CLAWS (Constituent-Likelihood Automatic Word-tagging System)
2) DeRose 对 CLAWS 改进后 VOLSUNGA 系统 (bi-gram)。

7.3 词性标注方法

◆ 基于 HMM 的词性标注方法

- 状态集
- 输出符号
- 初始状态概率
- 状态转移概率
- 符号输出概率

- [Manning, 2001] pp. 357-359:

- . Jelink's Method
- . Kupier's Method

7.3 词性标注方法

□ 规则和统计相结合的汉语词性标注方法

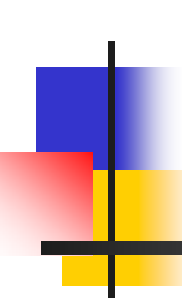
规则消歧，统计概率引导

或者统计方法赋初值，规则消歧

- [周强，1995；张民，1998，荀恩东，1999]

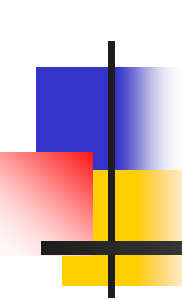
小结

- ❑ 词性标注的意义和复杂性
- ❑ 词性标注集确定的基本原则
- ❑ 词性标注的基本方法
 - ◆ 规则方法
 - ◆ 统计方法
 - ◆ 规则和统计相结合的方法



习题

1. 了解目前常见的几种汉语词性标注集，比较它们的差异，并阐述你个人的观点。
2. 掌握各种词性标注方法的要点，了解目前汉语词性标注的几种主要方法。



Thanks

谢谢!