计算所汉语词性标记集

Version 3.0

制订人：刘群 张华平 张浩

0. 说明

计算所汉语词性标记集主要用于中国科学院计算技术研究所研制的汉语词法分析器、句法分析器和汉英机器翻译系统。本标记集主要参考了以下词性标记集：

1. 北大《人民日报》语料库词性标记集；
2. 北大2002新版词性标记集（草稿）；
3. 清华大学汉语树库词性标记集；
4. 教育部语用所词性标记集（国家推荐标准草案2002版）；
5. 美国宾州大学中文树库（ChinesePennTreeBank）词性标记集；

由于计算所的汉语词法分析器主要采用北大《人民日报》语料库进行参数训练，因此本词性标记集主要以北大《人民日报》语料库的词性标记集为蓝本，并参考了北大《汉语语法信息词典》中给出的汉语词的语法信息。

本标记集在制定过程中主要考虑了以下几方面的因素：

1. 有助于提高汉语词法分析器的切分和标注正确率；
2. 有助于提高汉语句法分析器的正确率；
3. 有助于汉英机器翻译系统进行翻译；
4. 易于从北大《人民日报》语料库词性标记集进行转换；
5. 对于语法功能不同的词，在不造成词法分析和句法分析歧义区分困难的情况下，尽可能细分子类。

基于以上考虑，我们在标注过程中尽量避免那些容易出错的词性标记，而采用那些不容易出错、而对提高汉语词法句法分析正确率有明显作用的标记。例如，在动词的子类中，我们参考了宾州大学中文树库的做法，把汉语动词“是”和“有”分别做成单独的标记，而没有采用“系动词”的标记。因为同样是“是”这个动词，其句法功能很多，作“系动词”只是其中一种功能，而要区分这些功能是非常困难的，会导致词法分析的正确率下降。

在名词子类中，我们区分了“汉语人名”、“日语人名”和“翻译人名”，这不仅仅是因为这三种人名要采用不同的参数进行训练与识别，而且在汉英机器翻译中也要采用不同的分析算法进行翻译。又如，我们把表示时间的“数词＋‘年’”（如“1995年”）合并成一个时间词，而表示年头的“数词＋‘年’”分别标注为“数词”和“量词”，这是因为我们通过实验发现这种区分在词法分析阶段通过统计方法可以达到较高的正确率，而且这种区分对于后续的句法分析和机器翻译有非常重要的作用。

对于某些词类（助词和标点符号），基本上是一个封闭集，而这些词类中各个词的语法功能相差很大，在这种情况下，我们尽可能地细分其子类。

另外，与其他词性标记集类似，在我们的标记体系中，小类只是大类中一些有必要区分的一些特例，但小类的划分不满足完备性。

1. 名词

名词分为以下子类：

- n 名词
  - nr 人名
    - nr1 汉语姓氏
    - nr2 汉语名字
    - nrj 日语人名
    - nrf 音译人名
  - ns 地名
    - nsf 音译地名
  - nt 机构团体名
  - nz 其它专名
  - nl 名词性惯用语
  - ng 名词性语素

2. 时间词

- t 时间词
  - tg 时间词性语素

3. 处所词

- s 处所词

4. 方位词

- f 方位词

5. 动词

- v 动词
  - vd 副动词
  - vn 名动词
  - vshi 动词“是”
  - vyou 动词“有”
  - vf 趋向动词
  - vx 形式动词
  - vi 不及物动词（内动词）
  - vl 动词性惯用语
  - vg 动词性语素

6. 形容词

- a 形容词
  - ad 副形词
  - an 名形词

ag 形容词性语素
al 形容词性惯用语

7. 区别词

b 区别词
bg 区别词性语素
bl 区别词性惯用语

8. 状态词

z 状态词

9. 代词

r 代词
rr 人称代词
rz 指示代词
rzt 时间指示代词
rzs 处所指示代词
rzv 谓词性指示代词
ry 疑问代词
ryt 时间疑问代词
rys 处所疑问代词
ryv 谓词性疑问代词
rg 代词性语素

10. 数词

m 数词
mq 数量词

11. 量词

q 量词
qv 动量词
qt 时量词

12. 副词

d 副词

13. 介词

p 介词
pba 介词“把”
pbei 介词“被”

14. 连词

c 连词
cc 并列连词

15. 助词

u 助词
- uzhe 着
- ule 了 喽
- uguo 过
- ude1 的 底
- ude2 地
- ude3 得
- usuo 所
- udeng 等 等等 云云
- uyy 一样 一般 似的 般
- udh 的话
- uls 来讲 来说 而言 说来
- ujl 极了
- uzhi 之
- ulian 连 （“连小学生都会”）
- uqj 起见

16. 叹词

e 叹词

17. 语气词

y 语气词

18. 拟声词

o 拟声词

19. 前缀

h 前缀

20. 后缀

k 后缀

21. 字符串

x 字符串
- xx 非语素字
- xu 网址 URL

22. 标点符号

w 标点符号
- wkz 左括号，全角：（〔 ［ ｛ 《【 〖〈 半角：([{<
- wky 右括号，全角：）〕 ］ ｝》 】〗〉 半角：)]{>
- wyb 半角引号，半角："'

wyz 左引号，全角：“ ‘ 『
wyy 右引号，全角：” ’ 』
wj 句号，全角：。
ww 问号，全角：？ 半角：?
wt 叹号，全角：！ 半角：!
wd 逗号，全角：， 半角：,
wf 分号，全角：； 半角：;
wn 顿号，全角：、
wm 冒号，全角：： 半角：:
ws 省略号，全角：…… …
wp 破折号，全角：—— －－ ——－ 半角：--- ----
wb 百分号千分号，全角：％ ‰ 半角：%
wh 单位符号，全角：￥ $ ￡ ° ℃ 半角：$