这周主要了解自然语言处理基础工作中词性标注、命名实体识别、中文组块分析的一般标记方法。其次，在dblp上继续寻找事件抽取相关的好论文。

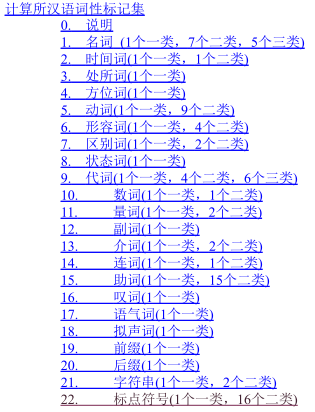
1. 词性标注集

哈工大LTP采用的863词性标注集

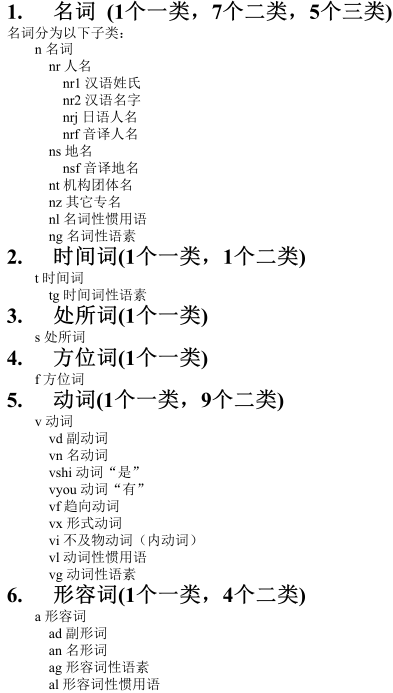


计算所汉语词性标记集（共计99个，22个一类，66个二类，11个三类）主要用于中国科学院计算技术研究所研制的汉语词法分析器、句法分析器和汉英机器翻译系统。

详细列表如下：



部分标注符号如下：



1. NER标注

BIO标记：

B stands for 'beginning' (signifies beginning of an NE)

I stands for 'inside' (signifies that the word is inside an NE)O stands for 'outside' (signifies that the word is just a regular word outside of an NE)

一些实例：

* **B-NP**: beginning of a noun phrase chunk
* **I-NP**: inside of a noun phrase chunk
* **O**: outside of a noun phrase chunk
* **B-PERS, B-DATE,...**:  beginning of a mention of a person/date...
* **I-PERS, B-DATE,...** : inside of a mention of a person/date...
* **O**: outside of any mention of a named entity

1. Chunk标注（来自The Chinese Penn Treebank Tag Set中文宾州树库标记）

|  |  |
| --- | --- |
| 组块类型 | 标记说明 |
| ADJP | 形容词性短语 |
| ADVP | 副词性短语 |
| CLP | 分类型短语 |
| DNP | 的字短语 |
| DP | 限定短语 |
| DVP | 地字短语 |
| LCP | 时间/地点限定短语 |
| LST | 列表型短语 |
| NP | 名词短语 |
| PP | 介词短语 |
| QP | 数量词短语 |
| VP | 动词短语 |