所有模型都在人民日报和BDT上测试，其中，2in1模型在人民日报语料上训练，其他模型都有在额外的语料上训练。

**总体表现**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 人民日报 | **上位词性标注正确率（按词）** | **下位词性标注正确率（按词）** | **上位词性标注正确率（按句）** | **下位词性标注正确率（按句）** |
| **2in1** | 0.956003 | 0.939139 | 0.432234 | 0.321123 |
| **jieba** | 0.583615 | 0.481636 | 0.015873 | 0.00854701 |
| **pkuseg** | 0.912774 | 0.899148 | 0.288767 | 0.229548 |
| **thulac** | 0.932302 | 0.867962 | 0.319902 | 0.106227 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| BDT | **上位词性标注正确率（按词）** | **下位词性标注正确率（按词）** | **上位词性标注正确率（按句）** | **下位词性标注正确率（按句）** |
| **2in1** | 0.851978 | 0.747082 | 0.049 | 0.005 |
| **jieba** | 0.617212 | 0.478968 | 0 | 0 |
| **pkuseg** | 0.851144 | 0.739048 | 0.081 | 0.015 |
| **thulac** | 0.852888 | 0.73632 | 0.051 | 0.005 |

在人民日报上，2in1效果最好，原因有二：

（1）2in1单纯在人民日报上训练并在上面测试；

（2）2in1和gold的词性标记是一样的，而其他模型都有POS集的差异。在BDT上，2in1的效果和pkuseg、thulac差不多，这是因为POS集的优势抵消了训练集的劣势。

**错误分析**

首先是2in1的错误分析，这个错误分析更有价值，因为训练集和测试集的POS集合是一致的。通过简单的归类，我们大致可以划分出以下主要的词性标注错误。

1. **分词问题，倾向于把带有前缀、后缀的词切开，然后把其中一个标注为整个词的词性**

最/d 高/a || 最高/a

最/d 新/a || 最新/a

只/d 能/v || 只能/v

只/d 是/v || 只是/c

这/r 次/q || 这次/r

这/r 个/q || 这个/r

1. **数量词问题，同样倾向于把数量词切分开，然后把标注为/m和/q**

１１/m 点/q || １１点/t

７/m 点/q || ７点/t

几/m 度/q || 几度/m

两/m 节/q || 两节/j

三/m 低/a || 三低/j

一/m 点/q || 一点/m

1. **词性活用判断错误**

演出/v || 演出/vn

邀请/v || 邀请/vn

要求/n || 要求/vn

销售/vn || 销售/v

修订/v || 修订/vn

修路/v || 修路/vn

稳定/an || 稳定/a

灵活/a || 灵活/ad

1. **兼类词判断错误，此类错误介词和动词较多**

靠/v || 靠/p

借/v || 借/p

给/p || 给/u

点/q || 点/n

与/c || 与/p

然后是jieba的错误分析，在排除了因标注集合导致的错误之后，我们可以简要归纳出以下错误类型。

1. **典型的数词+量词倾向于合并为一个整体/m**

一副/m || 一/m 副/q

一勺/m || 一/m 勺/q

三十六位/m || 三十六/m 位/q

三卷/m || 三/m 卷/q

千家/m || 千/m 家/q

1. **连续出现的词倾向于合并**

广泛传播/nr || 广泛/a 传播/vn

广大听众/i || 广大/b 听众/n

应邀出席/nr || 应邀/v 出席/v

干部职工/n || 干部/n 职工/n

中国国际广播电台/nt || 中国/ns 国际/n 广播/vn 电台/n

香港特别行政区/nt || 香港/ns 特别/a 行政区/n

1. **词类活用出现错误，多出现于动词活用于名词**

验收/v || 验收/vn 1

骤增/v || 骤增/vn 1

鸣响/vn || 鸣响/v

驱使/v || 驱使/vn

1. **其他奇怪的分类错误，多涉及到对文本整体的语义理解，难度较大**

马克思列宁主义/ns || 马克思列宁主义/n

马哲史/nr || 马哲史/j

马德里/nr || 马德里/ns（姓马的人）

马虎/nr || 马虎/an

反倒/v || 反倒/d

其次是pkuseg的错误分析。我们同样排除了因POS集不同导致的错误。

1. **时间词和量词分类错误**

４３年/t || ４３/m 年/q

３年/t || ３/m 年/q

１１亿多/m || １１亿/m 多/m

百年/m || 百/m 年/q

过去/t || 过去/v

近来/t || 近来/d

1. **把数词+量词合并为一个词，尤其是在“一+…”**

一体/n || 一/m 体/Ng

一口气/i || 一/m 口/q 气/n

一句话/l || 一/m 句/q 话/n

1. **成语、习语、缩略语分类错误**

解放思想/v || 解放思想/i

规章制度/n || 规章制度/l

莺啼燕唱/n || 莺啼燕唱/l

亚太/j 地区/n || 亚太地区/j

发达国家/n || 发达国家/l

1. **人名分类错误**

刘/nr 也刚/d || 刘也刚/nr

刘/nr 学春/n || 刘学春/nr

刘/nr 宝珍/n || 刘宝珍/nr

刘/nr 工践/n || 刘工践/nr

1. **带有前缀分类错误**

剩余/vn 价值/n || 剩余价值/l

副书记/n || 副/b 书记/n

副市长/n || 副/b 市长/n

小/a 企业/n || 小企业/n

高性能/v || 高/a 性能/n

1. **词类活用错误**

快乐/an || 快乐/a

感想/vn || 感想/n

感觉/n || 感觉/vn

积极/ad || 积极/a

积累/v || 积累/vn

最后是thulac的错误分析。

1. **“不X”类词分类错误，把单个词拆开了**

不/d 久前/a || 不久前/t

不/d 仅/d || 不仅/c

不/d 住/v || 不住/d

不/d 再/d || 不再/d

不/d 准/a || 不准/v

不/d 利/a || 不利/a

不/d 可/v 避免/v || 不可避免/l

1. **词类活用、兼类词错误**

严厉/a || 严厉/ad

严密/a || 严密/ad

享受/v || 享受/vn

亲切/a || 亲切/ad

任课/v || 任课/vn

自信/a || 自信/vn

领导/v || 领导/n

过去/v || 过去/vn

过去/v || 过去/vn

1. **带有前缀的词分类错误**

阿/j 姨/n || 阿姨/n

阿/j 里/f || 阿里/ns

那/r 么/g || 那么/r

那/r 种/q || 那种/r

1. **复杂的外国人名分类错误**

谢尔/np 盖/v 耶/g 夫/g || 谢尔盖耶夫/nr

萨/j 哈/j 夫/g || 萨哈夫/nr

查理斯/np ·/w 科雷/n || 查理斯·科雷/nr

综上所述，上面几个词性标注工具都有以下的错误类型：

1. 数量词错误，以“一X”为代表；
2. 词类活用、兼类错误，因此这类词的各种词性都会出现在训练集中，因此在测试的时候，模型难以判断到底该用哪一种，这是错误最多的一类；
3. 副词+X的错误，以“不X”为代表，其他还有“只X”，“还X”等。

除了以上共性，各工具还有一些特性：

2in1：因为该模型是在人民日报上训练的，领域相同，而且POS集也相同，错误较少。

Jieba：倾向于把连续的词合并，在一些专有名词上也有错误。

Pkuseg：时间词错误，成语习语等错误，人名错误，这可能是分词标准不同导致的。

Thulac：总的来说错误较少，主要体现在复杂专有名词的错误。