中文信息处理

陶

2024 年 3 月 1 日

目录

1 规则派Vs统计派

1.1 普通话中，“二”与“两”的区别？如何让机器正确填写以下片段？

规则派：

rule1：基数末尾、十位用二，不用两。十二、二十三、*十两、*两十五。

rule2：基数词百位以上，都可以用。

rule3：单个序数，同rule1。第二、*第两、初二、*初两。

rule4：数量名结构，度量衡以外，用“两”，不用“二”。两张桌子、*二张桌子。

rule5：度量衡做量词时，可以用“两”，也可以用“二”。两米、二米

优点：

如果规则完备，可生成所有合格的片段，避免所有不合格的片段。

缺点：

需要语言学专业人士参与总结。

统计派：

观察“二”、“两”在所有已知文本中的搭配，然后照抄。

优点：无需语言学专业人士参与。让机器做匹配即可。

缺点：从理论上说，出错的可能性总是有的。

互联网带来的大量数据是统计派占上风的根源

2 文本的处理

2.1 语言学单位角度

2.1.1 汉字

汉字编码

字库建设

汉字输入

汉字显示

2.1.2 词

中文分词

词库建设

词性标注

命名实体识别

2.1.3 句子

句法成分
论元结构
配价
语义特征
歧义结构分析

2.1.4 篇章

篇章衔接
篇章连贯
篇章标注

2.2 应用角度

中文分词
语料库建设
信息检索
问答系统
自动文摘
信息抽取
机器翻译

3 单字繁转简

给机器一个繁体字，让机器给出它的对应的简体字。
如果一个字不出现在对照表中，系统会报错:
解决方法：
1.完善对照表，添加上新的字。
2.忽略这次错误，改成不做处理，照抄，或其他方案。
事实上繁体字转简体字也存在一对多的情况

4 单字简转繁

导致的结果：搜索到多个匹配项
解决方法：从匹配到的多个结果中选择最适合的结果
先分词，然后根据上下文判断。

导航

在这里，你可以添加一些导航链接，如链接到、子节、第二节等。