**CRF++工具用于命名实体识别**

|  |  |
| --- | --- |
| 学号: | 1130310128 |
| 姓名: | 杨尚斌 |
| 邮箱: | [rccoder@foxmail.com](mailto:rccoder@foxmail.com) |
| 指导老师: | 赵铁军 |

1. **实验环境设置**
2. 硬件条件

* Windows 10 x64 RAM-8G CPU-Intel(R) Core(TM) i5-3230M @ 2.60GHz
* CRF++ 0.53 (Windows版)

1. 训练语料

从 1998 年人民日报选择出的 3000 句和 5000 句两种规模的语料

1. 测试语料

给定的 49996 条测试语料

1. **实验设计**

基本特征选择： 当前词 + 当前词性。特征模板采用工具给定的格式

多种的特征组合：

1. **特征组合1**

# Unigram

U00:%x[-2,0]

U01:%x[-1,0]

U02:%x[0,0]

U03:%x[1,0]

U04:%x[2,0]

U05:%x[-2,0]/%x[-1,0]/%x[0,0]

U06:%x[-1,0]/%x[0,0]/%x[1,0]

U07:%x[0,0]/%x[1,0]/%x[2,0]

U08:%x[-1,0]/%x[0,0]

U09:%x[0,0]/%x[1,0]

1. **特征组合2**

# Unigram

U00:%x[-2,0]

U01:%x[-1,0]

U02:%x[0,0]

U03:%x[1,0]

U04:%x[2,0]

U05:%x[-1,0]/%x[0,0]

U06:%x[0,0]/%x[1,0] ‘

U10:%x[-2,1]

U11:%x[-1,1]

U12:%x[0,1]

U13:%x[1,1]

U14:%x[2,1]

U15:%x[-2,1]/%x[-1,1]

U16:%x[-1,1]/%x[0,1]

U17:%x[0,1]/%x[1,1]

U18:%x[1,1]/%x[2,1]

U20:%x[-2,1]/%x[-1,1]/%x[0,1]

U21:%x[-1,1]/%x[0,1]/%x[1,1]

U22:%x[0,1]/%x[1,1]/%x[2,1]

1. **特征组合3**

# Unigram

U00:%x[-2,0]

U01:%x[-1,0]

U02:%x[0,0]

U03:%x[1,0]

U04:%x[2,0]

U05:%x[-1,0]/%x[0,0]

U06:%x[0,0]/%x[1,0]

U10:%x[-2,1]

U11:%x[-1,1]

U12:%x[0,1]

U13:%x[1,1]

U14:%x[2,1]

U15:%x[-2,1]/%x[-1,1]

U16:%x[-1,1]/%x[0,1]

U17:%x[0,1]/%x[1,1]

U18:%x[1,1]/%x[2,1]

U20:%x[-2,1]/%x[-1,1]/%x[0,1]

U21:%x[-1,1]/%x[0,1]/%x[1,1]

U22:%x[0,1]/%x[1,1]/%x[2,1]

U23:%x[0,1]

1. **实验结果**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **实验序号** | **训练语料规模（词数或句子数）** | **特征模板形式** | **精确率（%）** | **召回率（%）** | **F值** |
| **1** | 3000 | 1 | 99.2911 | 34.6933 | 0.514199 |
| **2** | 3000 | 2 | 93.7125 | 70.5814 | 0.805186 |
| **3** | 3000 | 3 | 93.5953 | 74.9526 | 0.83243 |
| **4** | 5000 | 1 | 93.6477 | 77.5463 | 0.8448398 |
| **5** | 5000 | 2 | 93.6347 | 75.4481 | 0.835633 |
| **6** | 5000 | 3 | 93.6312 | 76.3235 | 0.846432 |

1. **实验说明**
2. 在比较小的训练语料下，精确度随着特征模板的增多而减小，召回率和F值（正确率和召回率的调和平均值）增大；在比较大的训练语料下，三者的大小与特征模板的关系不是很大。
3. 在同样的特征组合下，语料越大，精确率下降，召回率和F值增大。