**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 | 91.2240234083 | 97.6810135846 | 94.3421648426 |
| **2** | 3000 | 2 | 96.168406082 | 99.9521770319 | 98.023791251 |
| **3** | 3000 | 3 | 96.5616045845 | 99.9523716741 | 98.22773499 |
| **4** | 5000 | 1 | 91.3939547987 | 98.5744311432 | 94.8484881791 |
| **5** | 5000 | 2 | 96.6235722355 | 99.9286038814 | 98.2483009476 |
| **6** | 5000 | 3 | 96.9788898885 | 99.9180875601 | 98.426551153 |

1. **实验说明**
2. 精确度、召回率、F 值随着训练语料规模的增多而增大
3. 精确度、召回率、F 值随着特征模板的增多而增大