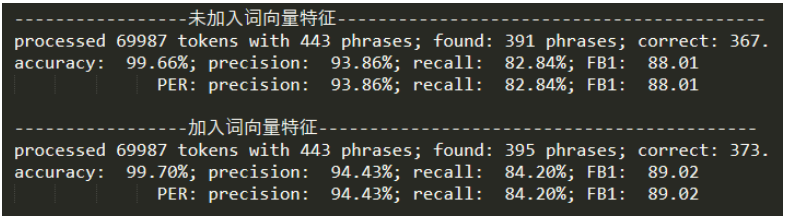
20190530

1. 基于词向量得藏文命名实体识别——4月底之前完成
   1. 飞飞：语料规模88%
   2. 孙朋：不加入词向量的F值是88.01%，加入词向量特征后的F值是89.02%。
2. 基于主动学习的藏文命名实体识别
3. 基于自学习的藏文命名实体识别
4. 基于融合方法的藏文命名实体识别

20190606

1. 1. 基于词向量的藏文命名实体识别——4月底之前完成

a)飞飞：语料规模88%

系统修改、（这周完成，6月9日）、显示

b)孙朋：不加入词向量的F值是88.01%，加入词向量特征后的F值是89.02%。

结果（excel）三列——今天完成、得结论

基模型的特征表示——下周完成——6月14号完成、效果好

3-4页（6月14号完成）

1. 基于主动学习的藏文命名实体识别——2019年8月底
   1. 开题报告
2. 基于自学习的藏文命名实体识别——特征相似度——2019年6月底、8月底
   1. 句子特征——1天（6月14号看）
   2. 特征相似度计算——？？（6月21号看）
   3. 选择策略（top5、动态）
   4. 停止（性能收敛）
3. 基于融合方法的藏文命名实体识别——2019年12月底
4. 论文撰写
   1. 小论文3篇（2019年2篇、2020年1篇）
   2. 毕业论文：2020年1月底

20190619

**语料格式说明**

# column: 0, feature: 藏文音节

# column: 1, feature: 位置特征

# column: 2, feature: 格助词

# column: 3, feature: 藏族人名高频音节

# column: 4, feature: 汉族姓氏高频音节

# column: 5, feature: 汉族人名高频音节

# column: 6, feature: 人名称谓词

# column: 7, feature: 词向量聚类类别

# column: 8, feature: 词向量相似词

# column: 9, feature: 藏文音节标签

**第一组实验（6月16日）**

**【数据】**

|  |  |  |
| --- | --- | --- |
|  | 训练集 | 测试集 |
| 文本数量 | 1271 | 140 |
| 人名数量 | 4532 | 443 |
| 文件大小（带特征标签） | 31.80 MB | 3.38 MB |

**【结果】**

|  |  |  |  |
| --- | --- | --- | --- |
| **特征组合方式** | **精确率** | **召回率** | **F1值** |
| F0（CRF 窗口5） | 71.74% | 74.49% | 73.09 |
| F0+位置特征 | 94.49% | 81.26% | 87.38 |
| F0+位置特征+格助词 | 94.13% | 83.30% | 88.38 |
| F0+位置特征+格助词+藏族人名高频音节 | 94.85% | 83.07% | 88.57 |
| F0+位置特征+格助词+藏族人名高频音节+汉族姓氏 | 94.12% | 83.07% | 88.25 |
| F0+位置特征+格助词+藏族人名高频音节+汉族姓氏+汉族人名高频音节 | 94.42% | 83.97% | 88.89 |
| F0+位置特征+格助词+藏族人名高频音节+汉族姓氏+汉族人名高频音节+人名称谓词 | 94.18% | 83.97% | 88.78 |
| F0+位置特征+格助词+藏族人名高频音节+汉族姓氏+汉族人名高频音节+人名称谓词+词向量聚类类别 | 93.22% | 83.75% | 88.23 |
| F0+位置特征+格助词+藏族人名高频音节+汉族姓氏+汉族人名高频音节+人名称谓词+词向量聚类类别+词向量相似词 | 93.67% | 83.52% | 88.31 |

**第二组实验（6月18日）**

【数据】

|  |  |  |
| --- | --- | --- |
|  | 训练集 | 测试集 |
| 文本数量 | 1032 | 239 |
| 人名数量 | 3446 | 1086 |
| 文件大小（带特征标签） | 24.70 MB | 7.74 MB |

**【结果】**

|  |  |  |  |
| --- | --- | --- | --- |
| **特征组合方式** | **精确率** | **召回率** | **F1值** |
| F0 | 70.85% | 44.75% | 54.85 |
| F0+藏族人名高频音节 | 69.74% | 47.33% | 56.39 |
| F0+藏族人名高频音节+汉族姓氏高频音节 | 69.07% | 47.70% | 56.43 |
| F0+藏族人名高频音节+汉族姓氏高频音节+汉族人名高频音节 | 70.74% | 46.96% | 56.45 |
| F0+藏族人名高频音节+汉族姓氏高频音节+汉族人名高频音节+人名称谓词 | 71.23% | 47.88% | 57.27 |
| F0+藏族人名高频音节+汉族姓氏高频音节+汉族人名高频音节+人名称谓词+词向量聚类类别 | 70.11% | 51.20% | 59.18 |
| F0+藏族人名高频音节+汉族姓氏高频音节+汉族人名高频音节+人名称谓词+词向量聚类类别+词向量相似词 | 69.29% | 51.10% | 58.82 |

1. 发以下文件 （今晚发）

* 藏族人名高频音节
* 汉族姓氏高频音节
* 汉族人名高频音节
* 人名称谓词

2. 命名实体在句子中的位置特征（明天讨论）

3. 说明词向量聚类类别和词向量相似词，并用实例说明目前的结果是否符合你的要求。（今晚发）

4. 6月16日和18日的实验结果差异很大（不应该有这么大的差异，另外18日的结果不应该这么差），找出原因。（明天讨论）

20190626

Baseline

* 88%~？？？（为什么）
* 88%（飞飞降下来 80%）

任务

1. ,完成之前的问题（今天）
2. 60个错误的分析
   1. 哪些提高
3. 测试语料
   1. 领域广（新闻、体育、娱乐、文学、宗教）
   2. 藏族人名、汉族人名、外国人名（英文、阿拉伯、俄文）