- 主题词检测模块模型图及标注过程
- 实验训练及实验结果
- 主题词库的构建计划，以及情感分析模块的后续工作

- 主题词检测模块模型图及标注过程

标注：

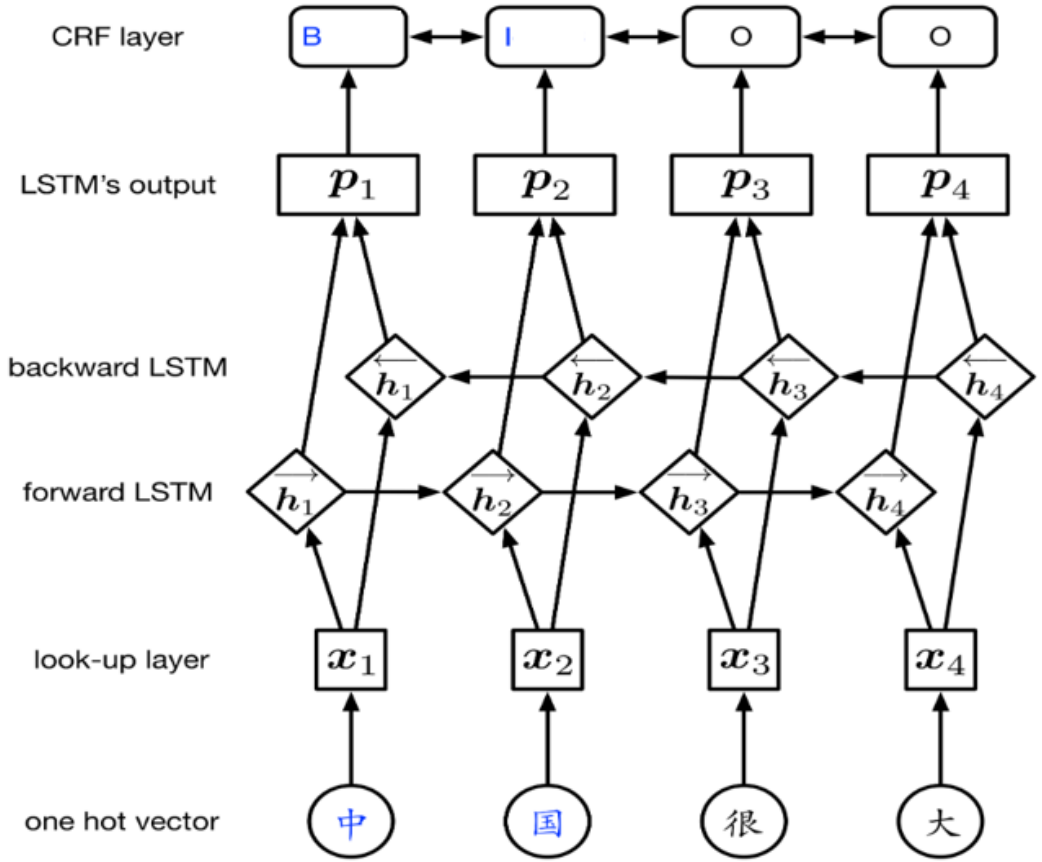
语料：Wiki百科三元组知识库2500多万条

统计词频，词性标注

筛选出：词频>30 名词性词语

词库总共是：38993个词汇

标注过程：对句子分词，词语在词库中或者是名词性词语 都标注。



模型图

- 实验训练及实验结果

- Train data: 107054条
- Validation data: 11895条 **9:1**
- Test data:80847条

超级马里奥之争！巴神垫底。巴神你可要加油啊！
在六月的雨季，瞻望喜马拉雅雪山。感恩！
楼层够高的童鞋可以参考参考。。
北京某中学地理课堂，绝对震撼，丢尽北京学生的脸！
我对皮皮虾已经产生一种敬畏了。
来来来，有啥事先站称上说，我就不信你的称不坏
生命中一个确定的必然是凋亡。
一年到底有几季？小故事，大道理。
到了晚上才想起来今天是植树节，赶忙上艹榴下了一批种子
本来是带着休闲心态看球结果要写稿郁闷啊
我妈心疼地说，小撒跟章子怡在一起怎么瘦了这么多。
真正的平静，不是避开车马喧嚣，而是在心中修篱种菊。
中国有哪些比较厉害的女风险投资人呢？
专业摄影师的定力，让人自叹弗如。
艺术家的世界，你们懂了吗？
一张网友制作的资产负债表的专业解释图
三年前陪在你身边的人现在在哪？
墨西哥总统卫队的一成员晕倒在欢迎仪式上
用你的笑容去看待这个世界，别让这个世界夺走了你的笑容。
我们每个人，都是某人，一生的至爱。
感动！免费接人的车已成洪流！好几十辆！
美丽滴午后，吃货为您奉上一杯蓝天白云。
本地化与技术写作经典书籍推荐
刚从北大百年讲堂讲座回来，有点累

```
processed 1552199 tokens with 177895 phrases; found: 175512 phrases; correct: 132879.
accuracy:  92.16%; precision:  75.71%; recall:  74.70%; FB1:  75.20
                 : precision:  75.71%; recall:  74.70%; FB1:  75.20  175512
                 : precision:  75.71%; recall:  74.70%; FB1:  75.20  175512
```

- 主题词库的构建

  - 利用wiki三元知识库，有关联关系的并成一个库。
    效果不是很好，再找新的数据或方法　　？
    考虑训练词向量 用 k-means 聚类方法试试 ？

- 情感极性分析模块

  粒度等级上可分为：

  词级别：识别一个词的倾向性。
  特征级别(Aspect Level)：识别一个Aspect的倾向性。如价格方面。
  句子级别：识别一个句子的观点倾向性。
  文档级别：识别一篇文本整体的倾向性。