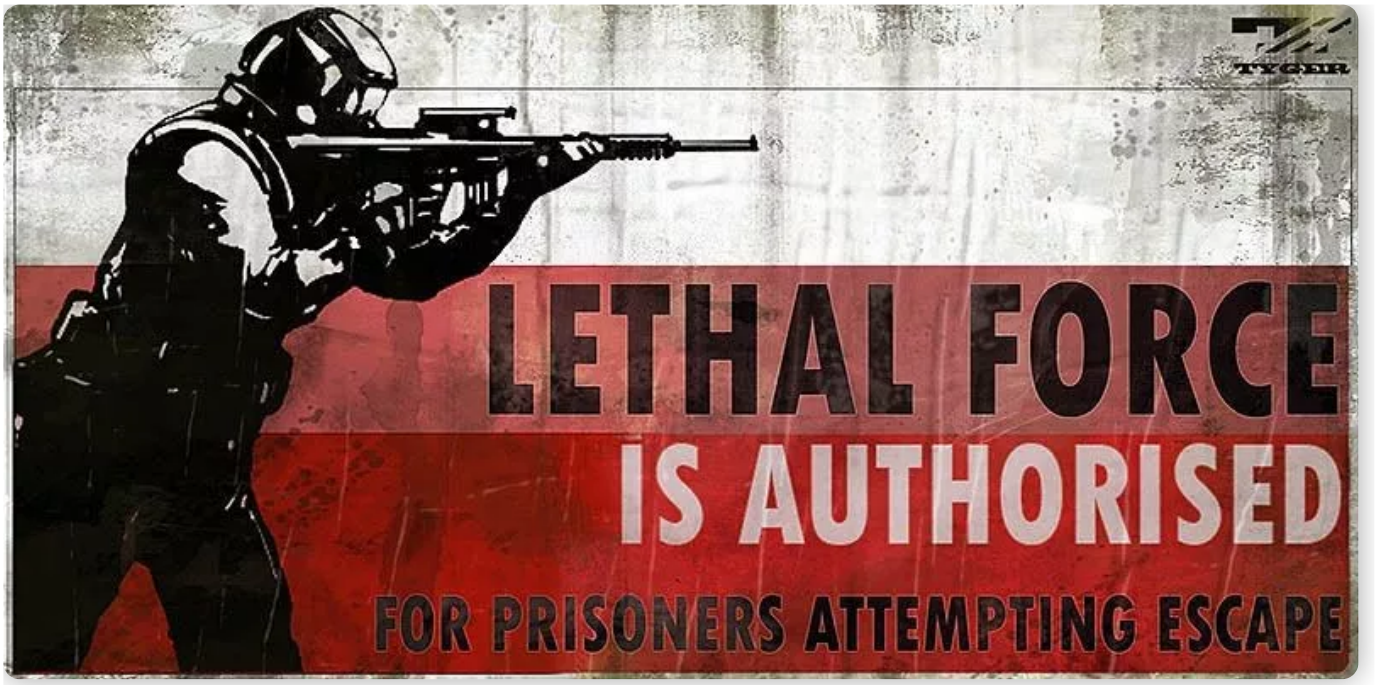
基于Bert-NER构建特定领域中文信息抽取框架

逸立学院AI Lab Python中文社区 2019-08-05

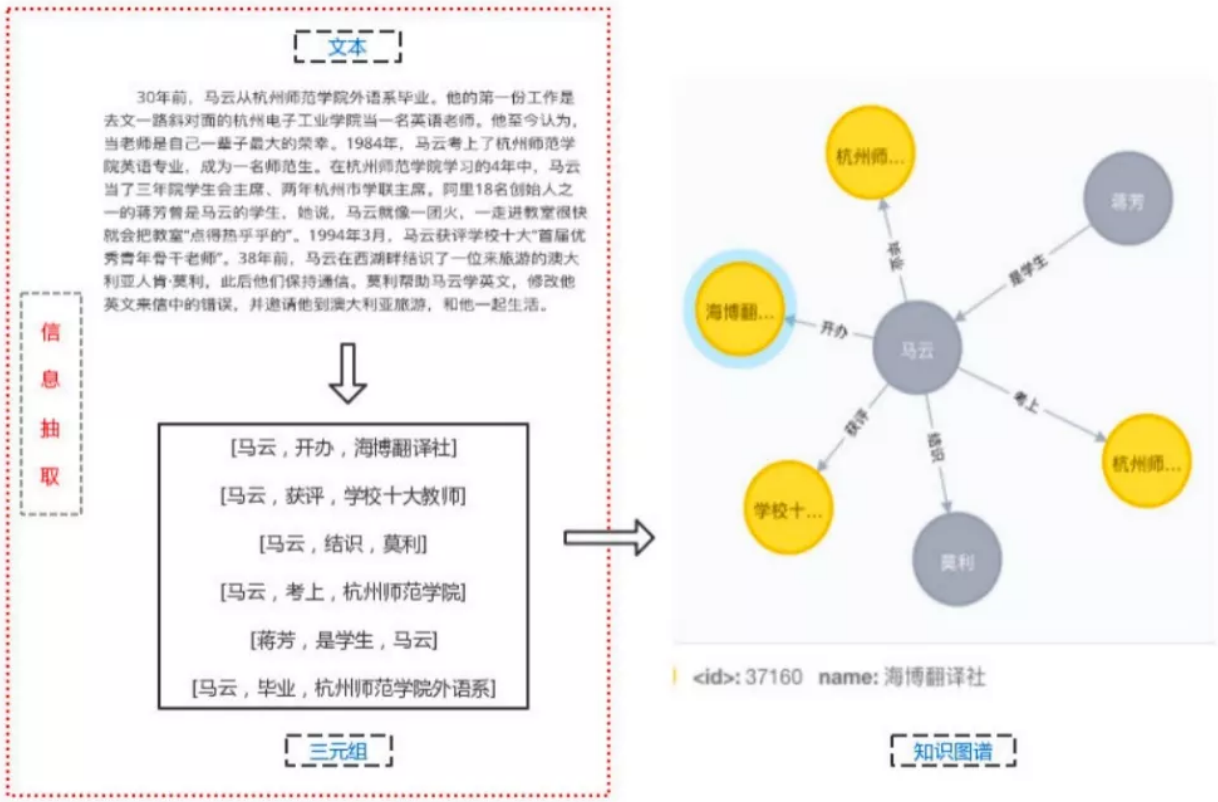


作者：朱展锋、李秋建、金立达，来自英国帝国理工学院，逸立学院AI Lab投稿。研究方向：自然语言处理、信息抽取、知识图谱

导语：

知识图谱（Knowledge Graph）主要由实体、关系和属性构成，而信息抽取（Information Extraction）作为构建知识图谱最重要的一个环节，目的就是从文本当中抽取出三元组信息，包括“实体-关系-实体”以及“实体-属性-实体”两类。然后将抽取后的多个三元组信息储存到关系型数据库（neo4j）中，便可得到一个简单的知识图谱。

本文通过多个实验的对比发现，结合Bert-NER和特定的分词、词性标注等中文语言处理方式，获得更高的准确率和更好的效果，能在特定领域的中文信息抽取任务中取得优异的效果。



1 信息抽取和知识图谱

目录

1 命名实体识别

- Bert-BiLSTM-CRF命名实体识别模型
- NeuroNER和BertNER的中文NER对比
- Bert-NER在小数据集下训练的表现

2 中文分词与词性标注

- （Jieba、Pyltp、PkuSeg、THULAC）中文分词和词性标注工具性能对比
- 分词工具与BertNER结合使用的性能

3 中文指代消解

- 基于Stanford coreNLP的指代消解模型
- 基于BertNER的中文指代消解框架

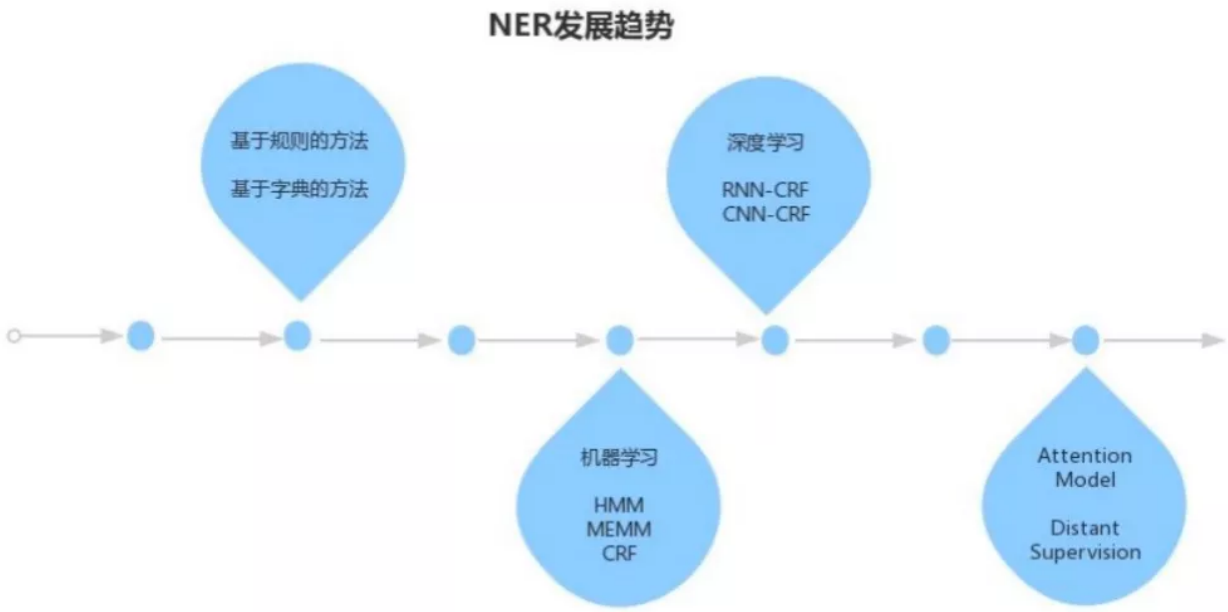
4 中文信息提取系统

- 中文信息抽取框架测试结果

一、命名实体识别

1.1 综述：

命名实体识别（Name Entity Recognition）是获取三元组中的实体的关键。命名实体指的是文本中具有特定意义或者指代性强的实体，常见的包括人名、地名、组织名、时间、专有名词等。就目前来说，使用序列标注的方法能够在NER任务中获得比较优异的效果，相对来说比较成熟。



2 NER发展趋势图

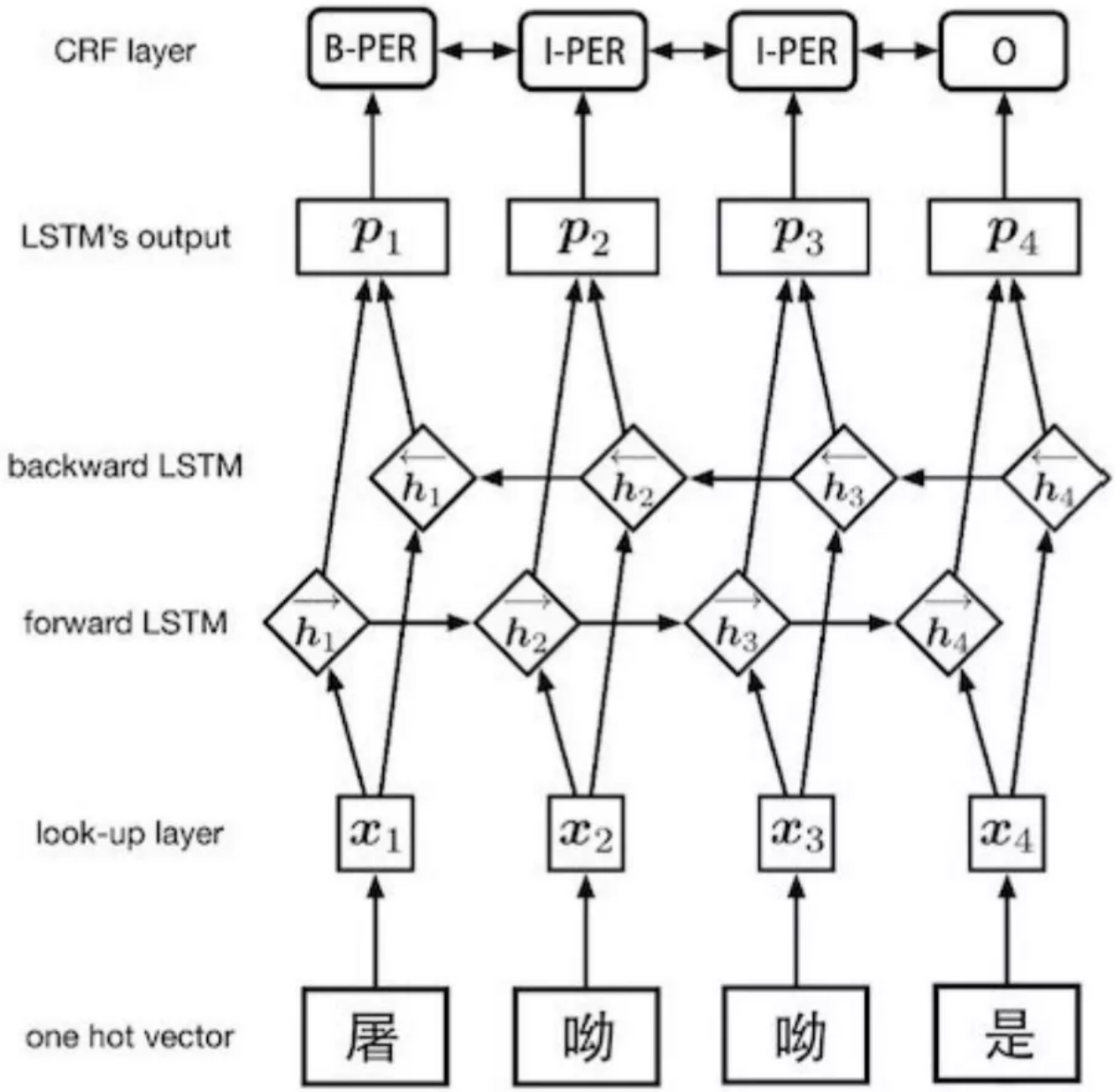
序列标注任务，即在给定的文本序列上预测序列中需要作出标注的标签。处理方式可简单概括为：先将token从离散one-hot表示映射到低维空间中成为稠密的embedding，随后将句子的embedding序列输入到RNN中，使用神经网络自动提取特征以及Softmax来预测每个token的标签。

本文对比了基于Bert的命名实体识别框架和普通的序列标注框架在模型训练、实体预测等方面的效果，并对基于小数据集的训练效果做出实验验证。

1.2模型：

1.2.1 Word Embedding-BiLSTM-CRF：

众多实验表明，该结构属于命名实体识别中最主流的模型，代表的工具有：NeuroNER。它主要由Embedding层（主要有词向量，字向量以及一些额外特征）、双向LSTM层、以及最后的CRF层构成，而本文将分析该模型在中文NER任务中的表现。



3 “词向量+BiLSTM+CRF”三层模型构造图

注：NER任务需要得到实体词的输出，所以使用字向量作为输入。

1.2.2 Bert-BiLSTM-CRF：

随着Bert语言模型在NLP领域横扫了11项任务的最优结果，将其在中文命名实体识别中Fine-tune必然成为趋势。它主要是使用bert模型替换了原来网络的word2vec部分，从

而构成Embedding层，同样使用双向LSTM层以及最后的CRF层来完成序列预测。详细的使用方法可参考：基于BERT预训练的中文NER（https://blog.csdn.net/macanv/article/details/85684284）

1.3 NeuroNER和BertNER的中文NER实验

1.3.1实验数据

1.3.1.1数据来源：

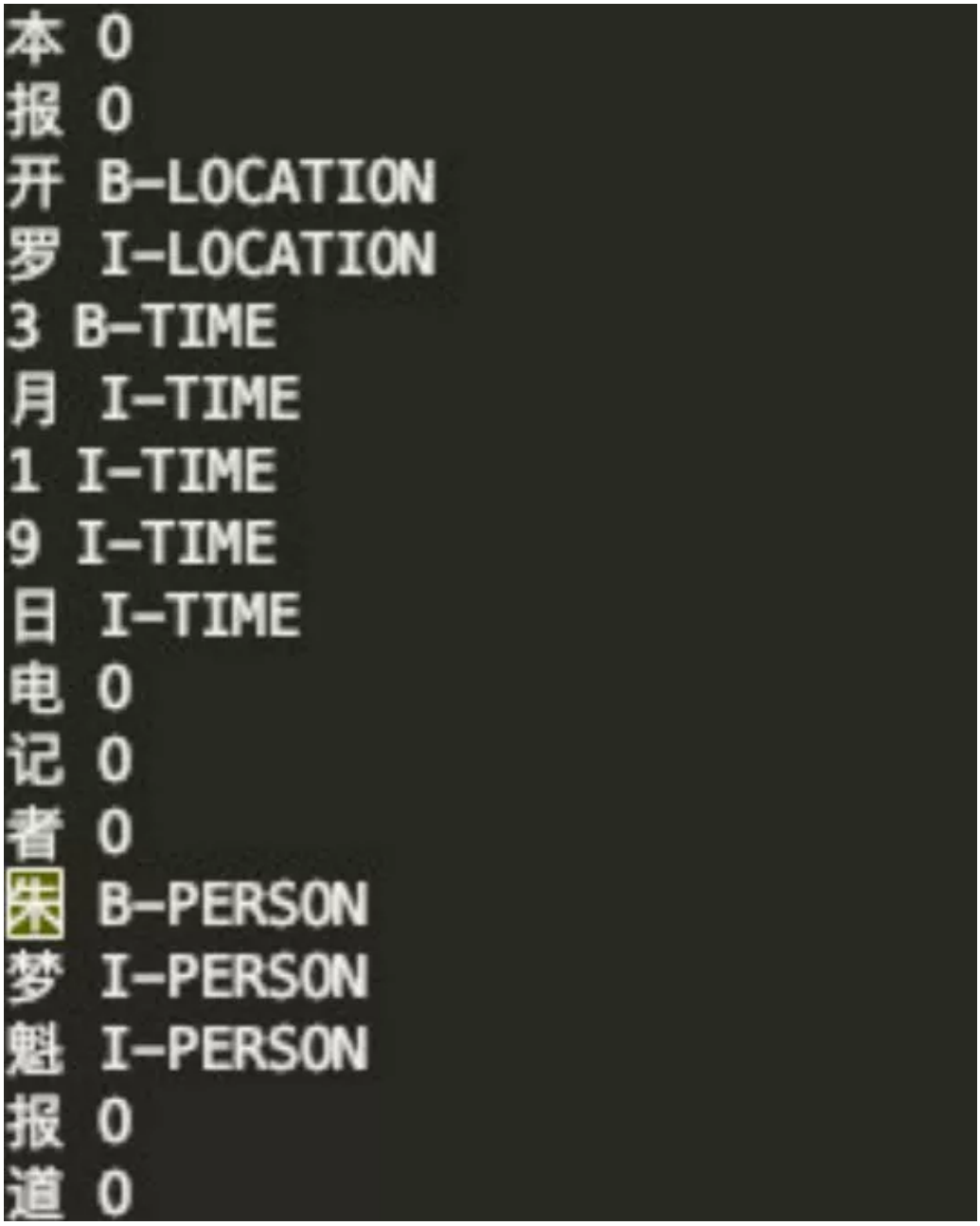
本文的NER实验数据是来自于人民网的将近7万句（250万字）中文新闻语料。

1	本报开罗3月19日电记者朱梦魁报道:埃及总统穆巴拉克今天下午会见了正在开罗访问的联合国秘书长安南,
2	就当前的一些国际和地区问题交换了意见,
3	并着重讨论了中东和平进程的最新形势、非洲事务以及前不久发生的伊拉克危机等问题。
4	快餐企业模式,
5	在传统与现代两大类下有了多种形态。
6	新落成的现代快餐企业时有所闻,
7	传统快餐业也在改造中逐步改变面貌。
8	快餐业在配送中心与连锁经营方面的实践与探索,
9	取得了一些经验。
10	特别是中外合资、外商独资和民营快餐企业,
11	在经营机制、管理体制和营销观念上的鲜明特色,
12	正影响着整个中国快餐业的建设和发展。
13	安南表示,
14	卢旺达的惨剧是一个世界性的悲剧。

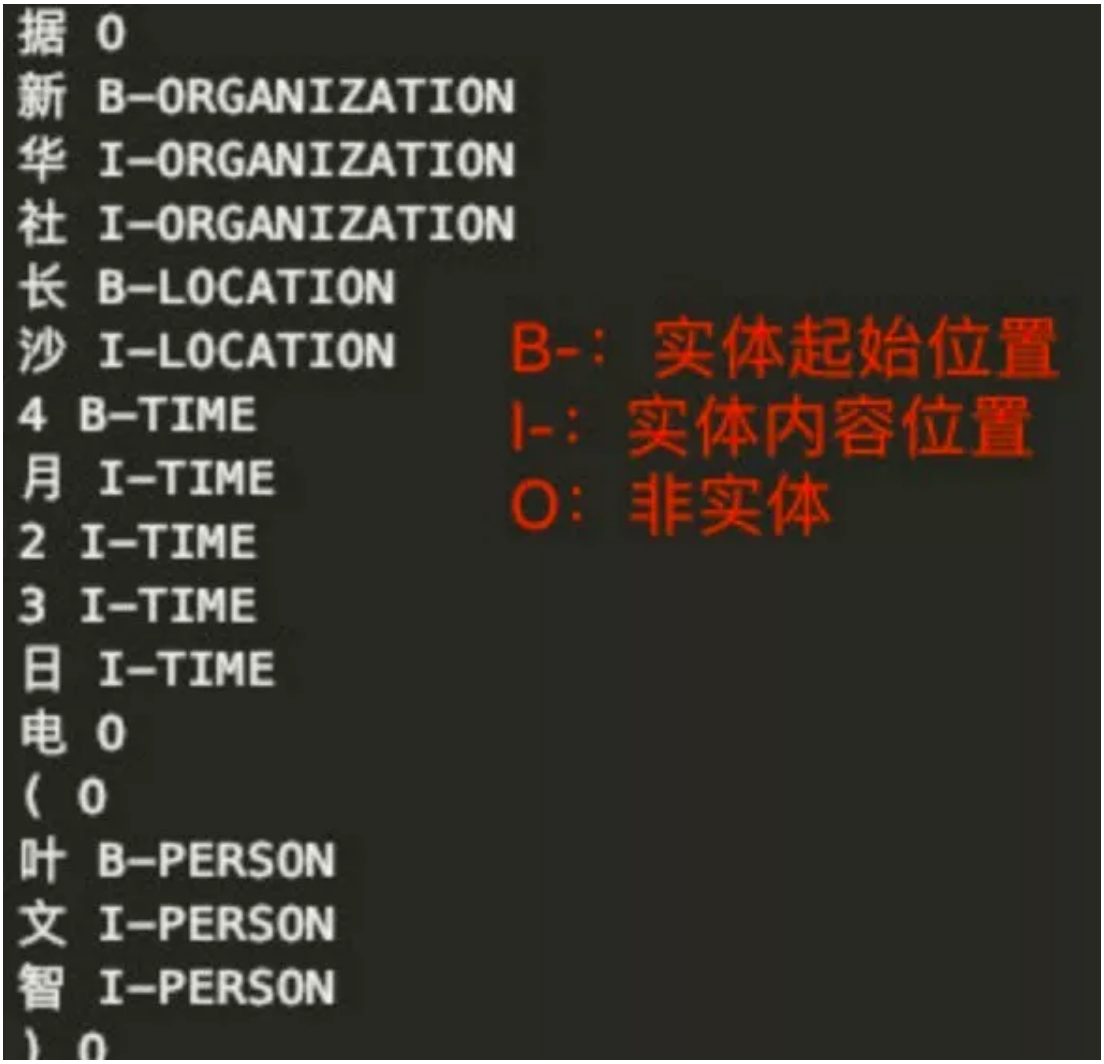
4 CSV格式的原始数据

1.3.1.2 数据样式：

本文选用BIO标注法，其中”B“表示实体起始位置，”I“表示实体内容位置，”O“表示非实体。将7万条数据样本经过清洗后，按字进行分割，使用BIO标注形式标注四类命名实体，包括人名（PERSON）、地名（LOCATION）、组织机构名（ORGANIAZATION）以及时间（TIME），构成中文命名实体识别语料库。


本 O
报 O
开 B-LOCATION
罗 I-LOCATION
3 B-TIME
月 I-TIME
1 I-TIME
9 I-TIME
日 I-TIME
电 O
记 O
者 O
朱 B-PERSON
梦 I-PERSON
魁 I-PERSON
报 O
道 O

5



6 数据标注样式图

1.3.1.3数据划分：

训练集、验证集、测试集以“7:1:2”的比例划分。其中训练集达到49600条的样本数，标注实体共88192个；验证集为7000条，包含12420个标注实体；测试集为14000条，标注实体共25780个。

数据集	样本数（句）	标注实体（个）
训练集	49600	88192
验证集	7000	12420
测试集	14000	25780

7

1.3.1.4命名实体识别结果展示：

- 展示用例：屠呦呦，女，汉族，中共党员，药学家。1930年12月30日生于浙江宁波，1951年考入北京大学，在医学院药学系生药专业学习。1955年，毕业于北京医学院（今北京大学医学部）。
- 展示用例抽取结果：[['PERSON', '屠呦呦'], ['TIME', '1930年12月30日'], ['LOCATION', '浙江宁波'], ['TIME', '1951年'], ['ORGANIZATION', '北京大学'], ['ORGANIZATION', '医学院药学系'], ['TIME', '1955年'], ['ORGANIZATION', '北京医学院'], ['ORGANIZATION', '北京大学医学部']]

1.3.1.5实验结果：

工具分类	epoch	训练时长	测试集F1值	模型加载速度	预测速度
BertNER	30	5h 49m	96.18	2.8s	80ms
NeuroNER	30	14h 19m	91.93	23s	2s
NeuroNER	100	2d 7h 10m	92.33	23s	2s

8

注：实验配置为11G Nvidia RTX2080Ti、Intel(R) Core(TM) i7-8700K CPU @ 3.70GHz、16G内存、2T硬盘

1.3.2结论：

a.实验表明，两者在相同的迭代次数训练后，测试集的F1值上BertNER比NeuroNER高出超过4个百分点。即使NeuroNER迭代epoch增加到100，仍然是BertNER的识别效果更优。
b.Bert NER在训练时长、模型加载速度、预测速度上都占据了很大的优势，达到工业级的水平，更适合应用在生产环境当中。
c.综上所述，Bert-BiLSTM-CRF模型在中文命名实体识别的任务中完成度更高。

1.4 Bert-NER在小数据集下训练的表现：

1.4.1实验数据：

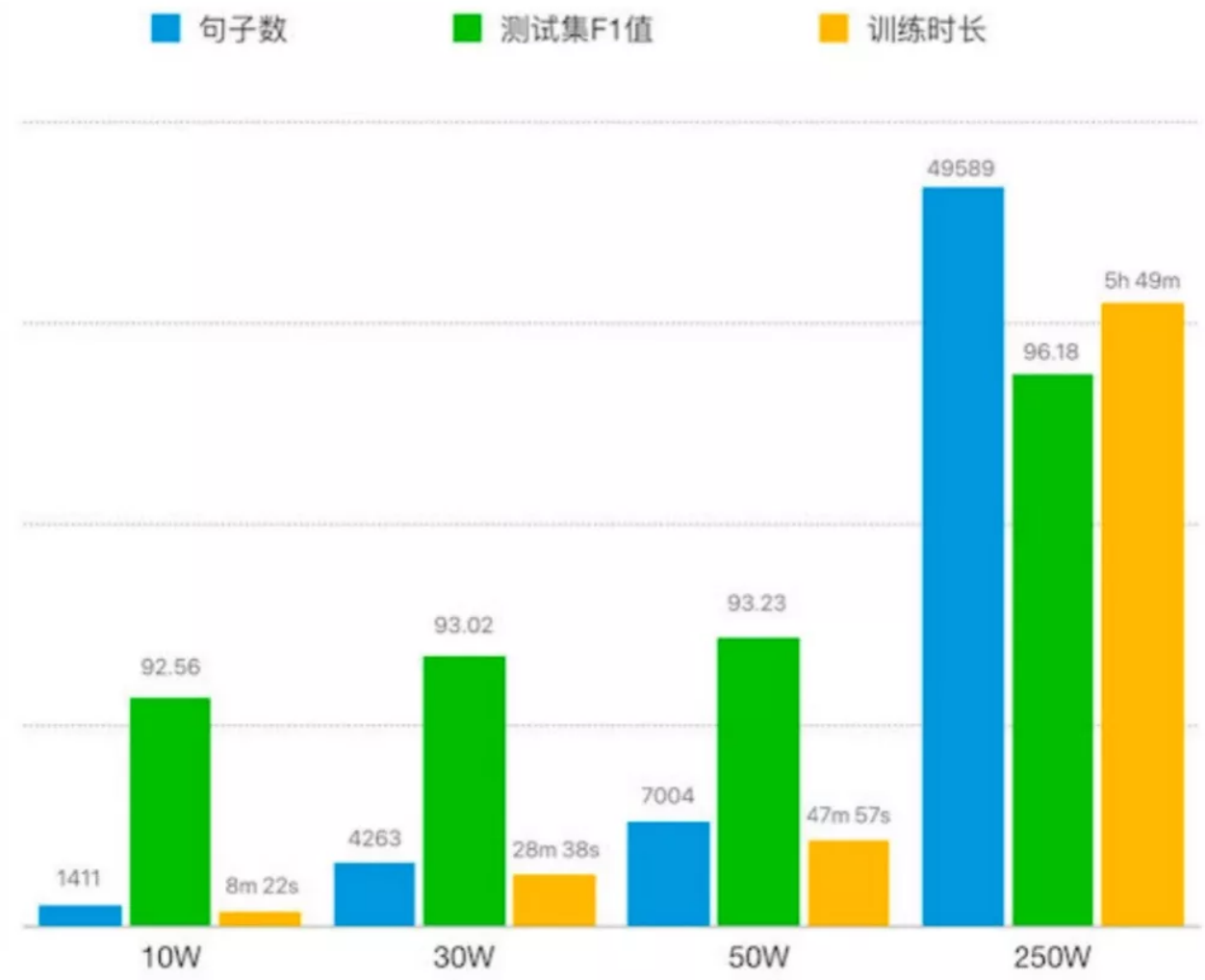
从5万句（250万字）的中文新闻语料中按文本数据的字数（万字为单位）划分出10W、30W、50W的小数据集，同样以“7:1:2”的比例得到对应的训练集、验证集、测试集。

1.4.2命名实体识别结果展示：

- 展示用例：屠呦呦，女，汉族，中共党员，药学家。1930年12月30日生于浙江宁波，1951年考入北京大学，在医学院药学系生药专业学习。1955年，毕业于北京医学院（今北京大学医学部）。
- 展示用例抽取结果：[['PERSON', '屠呦呦'], ['TIME', '1930年12月30日'], ['LOCATION', '浙江宁波'], ['TIME', '1951年'], ['ORGANIZATION', '北京大学'], ['ORGANIZATION', '医学院药学'], ['TIME', '1955年'], ['ORGANIZATION', '北京医学院'], ['ORGANIZATION', '北京大学医学部']]

1.4.3实验结果：

在相同实验配置下，四种数据集经过30个epoch的迭代训练，将句子数、训练市场、测试集F1值三个维度的实验结果进行归一化处理后，最终得到以下实验结果图表：



9 实验结果图

- 效能分析：本文将以10W的数据集实验结果作为基础，探讨在30W、50W和250W三种数据集训练，每当数据量增长一倍（即每增长10W的数据量），所带来的训练时长增长和模型提升比例：

数据集	训练时长增长比例	F1 值提升比例
30W	1.2 倍	0.25%
50W	1.2 倍	0.17%
250W	1.2 倍	0.11%

10 效能对比表

1.4.4结论：

1) BertNER在小数据集甚至极小数据集的情况下，测试集F1值均能达到92以上的水平，证明其也能在常见的文本命名实体识别任务中达到同样优秀的效果。
2) 实验结果证明，利用小数据集训练，可以大大降低人工标注成本的同时，训练时长也越少，也将极大地提高模型迭代的能力，有利于更多实体类型的NER模型构建。
3) 经过效能分析可以看出，数据量往上增加的同时，训练时长以相同的比例增加，而F1值提升的幅度在逐渐下降。因此，我们在扩充实体类别的时候，可以参考此效能比例，从而衡量所要投入的资源以及所能达到的模型效果。

二、中文分词和词性标注

2.1综述：

分词：
语言通常是需要用词来描述事物、表达情感、阐述观点等，可是在词法结构上中文与英文有较大的区别。其中最大的不同是英文将词组以空格的形式区分开来，较为容易被自动化抽取出来，而中文的词组往往需要由两个以上的字来组成，则需要通过分词工具来将语句拆分，以便进一步分析内容和意图。

词性标注：
对分词后的单词在用法上进行分类，为句法分析、信息抽取等工作打下基础。常见的词性包括名词、动词、形容词、代词、副词等。

2.2 分词和词性标注工具对比：

分词和词性标注往往是一同完成的。本文选取了主流的四款中文自然语言处理工具包括：Jieba、Pyltp、PkuSeg、 THULAC 。

工具分类	平均速度	用户自定义词典	集成程度	来源
Jieba	6ms	**支持，不与模型加载**	分词可单独使用	百度员工
Pyltp	2ms	作为特征加载到模型中	**分词、词性标注均可单独使用**	哈工大
PkuSeg	1.4s	支持，与模型加载	分词可单独使用	北京大学
THULAC	1.6s	支持，与模型加载	分词可单独使用	清华大学

对比测试了它们分词和词性标注上的效果、速度、功能以及集成程度等。其中速度方面的测试，使用了百度百科上100位科技人物的首句人物介绍，经过预测得到每句文本的平均计算。

注：实验配置为11G Nvidia RTX2080Ti、Intel(R) Core(TM) i7-8700K CPU @ 3.70GHz、16G内存、2T硬盘

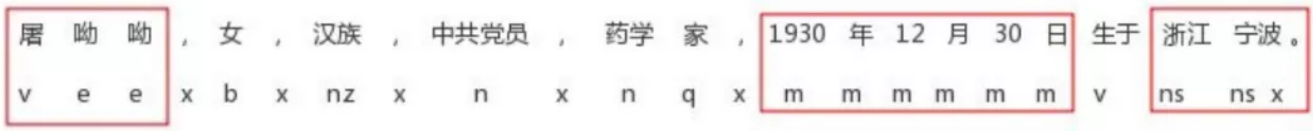
2.2.1测试文本：

屠呦呦，女，中共党员，药学家，1930年12月30日生于浙江宁波。

12

2.2.2效果对比：

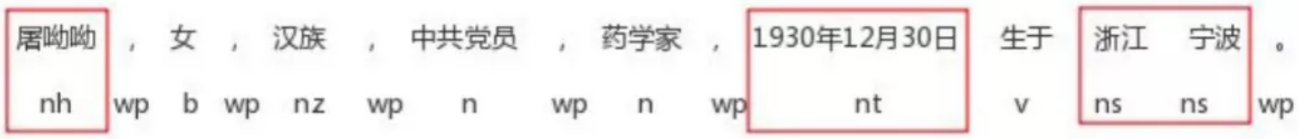
- Jieba：



13

注：v（动词）、e（叹词）、b（区别词）、n（名词）、ns（地名）、nz（其他专名）、q（量词）、m（数词）、x（非语素字）

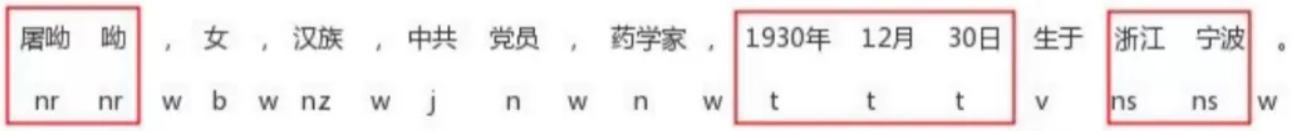
- Pyltp：



14

注：nh（人名）、n（名词）、ns（地名）、nt（时间名词）、nz（其他专名）、b（区别词）、wp（标点符号）

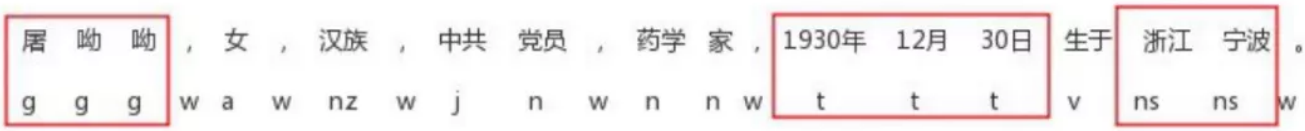
- PkuSeg:



15

注：nr（人名）、ns（地名）、nz（其他专名）、t（时间词）、b（区别词）、j（简称）、w（标点符号）

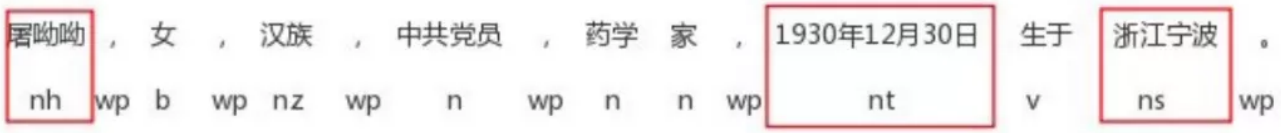
- THULAC:



16

注：g（语素词根）、ns（地名）、nz（其他专名）、t（时间词）、a（形容词）、j（简称）、w（标点符号）

- Jieba分词 + Bert-NER + Pyltp词性标注:



17

注：nh（人名）、n（名词）、ns（地名）、nt（时间名词）、nz（其他专名）、b（区别词）、wp（标点符号）

2.3结论:

a. 经过NER、分词、词性标注的对比测试后发现，Jieba分词同时具有速度快和支持用户自定义词典的两大优点，Pyltp具有单独使用词性标注的灵活性。因此，使用“Jieba分词 + BertNER作自定义词典 + Pyltp词性标注”的组合策略后，可以弥补Jieba分词在实体识别的缺点，保证较高的准确率和产品速度。

b. PkuSeg和THULAC：初始化模型就需要很长时间，导致分词和词性标注的模型预测速度慢，同时部分人名的命名实体识别有所缺失。

c. Pyltp：分词效果太过于细化，而且实际上是无法用到用户自定义词典的。因为LTP的分词模块并非采用词典匹配的策略，而是外部词典以特征方式加入机器学习算法当中，并不能保证所有的词都是按照词典里的方式进行切分。

三、中文指代消解

3.1综述：

指代消解（Coreference Resolution），即在文本中确定代词指向哪个名词短语，解决多个指称对应同一实体对象的问题。

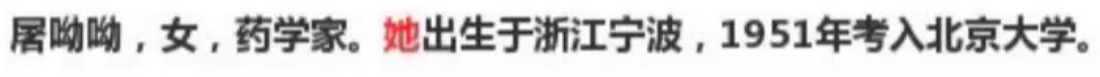
常见用于实现指代消解的工具包：NeuralCoref、Stanford coreNLP、AllenNLP等。

大部分工具包都是基于语义结构中的词和句的规则来实现指代消解，而且都是在英文的语言结构当中实现了不错的效果，NeuralCoref和AllenNLP不支持中文，而Stanford coreNLP 是具有多种语言模型，其中包括了中文模型，但Stanford coreNLP 的指代消解在中文的表现并不理想。目前而言，基于深度学习的端到端指代消解模型还达不到生产应用的要求。

3.2基于Stanford coreNLP的指代消解模型：

3.2.1系统架构：
运用Stanford coreNLP中文模型的词性标注、实体识别和句法依存功能模块+规则来构成一个中文指代消解系统。

3.2.2输入：



3.2.3结果：

分类	结果
主语	屠呦呦
词性标注	[('屠', 'NR'), ('呦呦', 'NR')]
命名实体识别	[('屠', 'O'), ('呦呦', 'O')]

19

主语"屠呦呦"被拆分为两个元素，这也直接导致了主语识别成了呦呦。最后的结果为：

屠呦呦，女，药学家。呦呦出生于浙江宁波，1951年考入北京大学。

20

3.3基于BertNER的中文指代消解框架：

本文选取Pyltp中文工具包中的依存句法分析模块，结合“Jieba分词 + BertNER作自定义词典 + Pyltp词性标注”的词性标注和BertNER实体识别模块，以确定输入文本段落的主语和实体，从而将文本中出现的代词指代到对应的实体上。并且还实现了对缺失主语的部分文本进行主语补齐。

3.3.1实验结果：

21

3.3.2经过反复的实验表明，基于BertNER的中文指代消解框架比基于Stanford coreNLP的指代消解模型在中文上获得更高的准确率和更好的效果，同时实现了主语补齐的功能，有助于抽取更多的有用三元组信息。

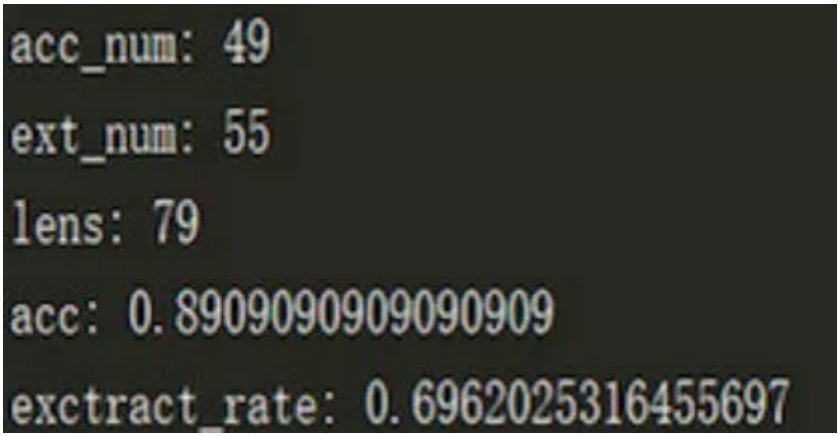
四、中文信息抽取系统

以下是基于Bert-NER的中文信息抽取系统的最终实验结果。

4.1中文信息抽取框架测试结果：

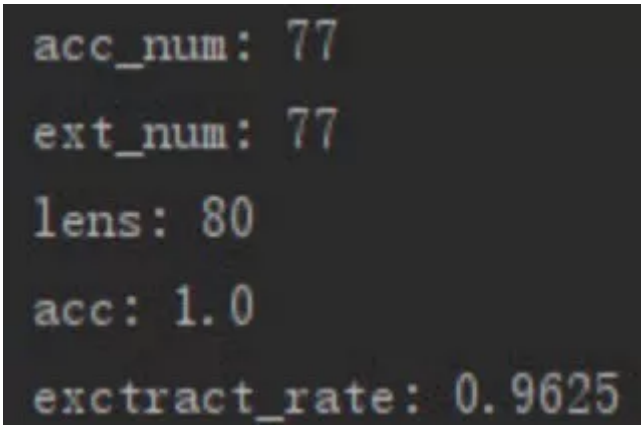
目前的规则配置文档定义了五类关系：出生于，配偶，毕业于，工作在，父（母）子。

4.1.1基于80条百度百科人物介绍，使用StanfordCoreNLP提取三元组的效果如下图所示。五类的关系抽取三元组准确率为0.89，抽取率达到0.69。



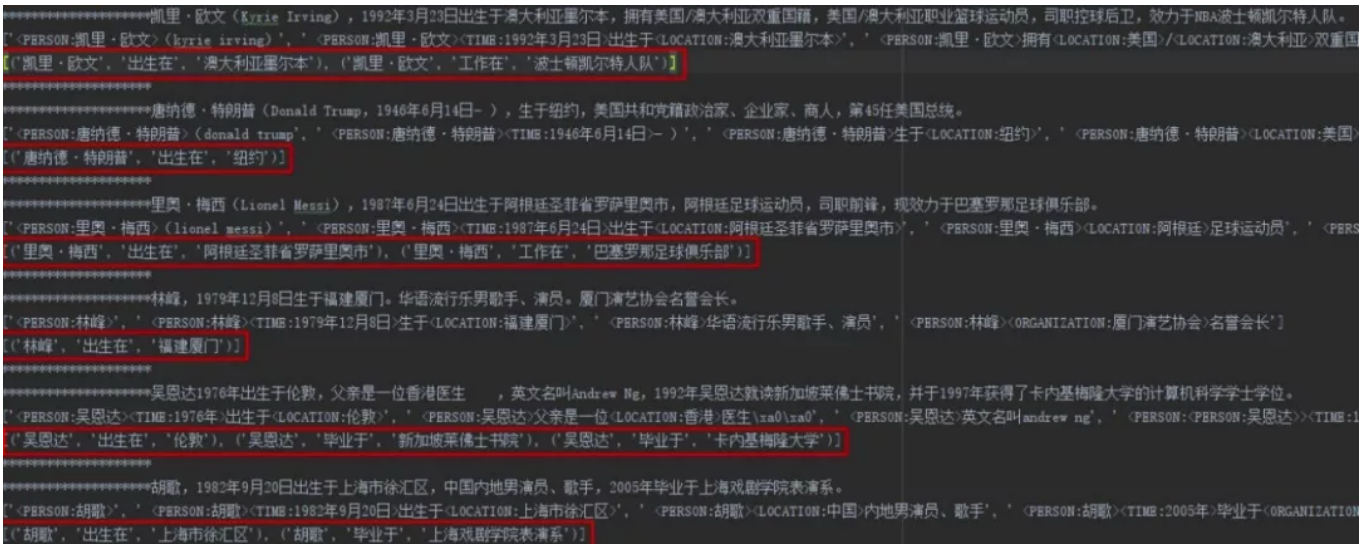
22

4.1.2基于80条百度百科人物介绍，使用本文中文抽取模型，取得较为明显的改进，五类的关系抽取三元组准确率达到0.99，抽取率达到0.96。



23

4.1.3测试用例结果展示：



24

本文实验代码：

中文命名实体识别：https://github.com/EOA-AILab/NER-Chinese

中文分词与词性标注：https://github.com/EOA-AILab/Seg_Pos