本周报告

> 2020/03/27 1652792 罗吉皓

本周主要做了数据集的搭建，分为以下几个部分：

1. modecai库相关尝试
2. spaCy库相关尝试
3. stanfordNlp库相关尝试
4. 动词-地名树搭建
5. 人工数据标注

事件地理位置定义

首先我们对于事件地点做一个定义。在modecai相关论文Geolocating Political Events in Text中，对于抽取文本事件中的核心任务定义为事件核心动词的寻找，这也是目前比较普遍的事件定义方式。本毕业设计针对事件地理位置的查询，借鉴之前论文的观点，将事件地理位置定义为与事件核心联系最为紧密的地理位置。目前实验的进展皆以此前提为基础。

实验目标

数据集的搭建主要分为两部分，第一部分从语料中提取所有的事件地点，第二部分是从预料中提取与事件最为相关的地点作为事件地点。

实验过程

0. 数据过滤

由于初始txt文本中存在较多的冗余信息，因此在实验前我提取了所有txt中的标题和第一段文字作为实验素材，总共素材共12w个。

1. modecai库相关尝试

mordecai库相关介绍在上次的文档中已经有所涉及，modecai是openeventdata官方推出的一款全文本地理分析python库，mordecai将文本信息处理主要分为3部分：

1. mordecai使用spaCy的命名实体识别从文本中提取地名。spaCy库是目前一个非常新的nlp python库，带有预训练的统计模型和单词向量，目前支持50多种语言的标记化。它具有最新的速度，用于标记，解析和命名实体识别的卷积神经网络模型，并且易于进行深度学习集成。经过spaCy处理之后，文本中相应的地理名词已经被提取出来了。
2. mordecai采用了geonames的地名索引找到提取的地名的潜在坐标。由于geonames地名索引文件非常大，mordecai将其放入Elasticsearch数据库中，通过索引起到快速寻找的功能。
3. mordecai最后使用 Keras 实现的神经网络，并对新的标注了 Prodigy的数据进行训练，以推断正确的国家和正确的地名条目，从而对每个地名进行正确的分类。

由于modecai库的训练样本为英语的缘故，做中文事件提取的时候我选择将其通过有道词典翻译的方式转换成英语，通过modecai得到事件地点后，再反向将其转化为中文。在整个过程中还是遇到很多不同的问题：

1. 有道api的翻译次数限制，导致ip被封。最后通过搭建ip代理池完成了翻译的过程
2. modecai库搭建问题。文件基本都在外网下载速度非常缓慢。

在克服以上困难后，我进行了一定的实验。这个方法虽然可行，但是存在了一定的缺陷：

1. 存在一定的翻译误差，比如：

```
原句：
中国抵达苏瓦港，对斐济进行为期8天的友好访问并提供人道主义医疗服务

翻译：
China has arrived in suva port for an 8-day goodwill visit to Fiji and humanitarian medical services

结果
[{'word': 'China', 'spans': [{'start': 2, 'end': 7}], 'country_predicted': 'CHN', 'country_conf': 0.9999591, 'geo': {'admin1': 'NA', 'lat': '35', 'lon': '105', 'country_code3': 'CHN', 'geonameid': '1814991', 'place_name': 'People's Republic of China', 'feature_class': 'A', 'feature_code': 'PCLI'}}, {'word': 'Fiji', 'spans': [{'start': 61, 'end': 65}], 'country_predicted': 'FJI', 'country_conf': 0.9986525, 'geo': {'admin1': 'NA', 'lat': '-18', 'lon': '178', 'country_code3': 'FJI', 'geonameid': '2205218', 'place_name': 'Republic of Fiji', 'feature_class': 'A', 'feature_code': 'PCLI'}}]
```

可以看到在其中苏瓦港这个核心地理位置并没有被准确的翻译，以至于在之后的判断中没有办法对于苏瓦港这个地理名词进行识别，最后结果中也没有出现这个地理位置。因此翻译中存在的误差对于最后的结果还是很大的。

2. 相关地理位置定位无法回溯。

以上述句子为例，最后结果为中国以及斐济两个地理名词。modecai库本身是有词义定位的功能的，但是由于输入算法的是英语文本，其定位也是英语文本。英语文本的语序和中文文本有一定差距，最后转换成中文后，对应的是哪一个地理位置比较难定位，尤其是一句话内存在多个地理名词指代一个地理位置，更加难以判断，比如：

```
中国日报网环球在线报道：据英国媒体7月9日报道，俄罗斯军事杂志刊登的最新一份调查报告显示，2008年俄格战争期间俄罗斯损失的战机中有一半是被自己的防空系统击落的，俄军方对此予以坚决否认。
```

句子中，俄罗斯出现了三次，事件核心地理位置确实是俄罗斯，但是哪一个俄罗斯才是真正的事件地理位置，在一个模糊的标注范围内就比较难以寻找。

3. 算法本身准确率

modecai论文中提及其正确率可以得到80%以上。实际上对于事件地点的理解，定义的不同会导致最后结果存在一定的出入。modecai中更多的是将所有可能是事件地点的位置都标注出来，比如：

```
原句：
中国警告美国必须深刻思考对华加征关税的后果。
结果：
[{'word': 'China', 'spans': [{'start': 2, 'end': 7}],
'country_predicted': 'CHN', 'country_conf': 0.99992335, 'geo': {'admin1':
'NA', 'lat': '35', 'lon': '105', 'country_code3': 'CHN', 'geonameid':
'1814991', 'place_name': 'People's Republic of China', 'feature_class':
'A', 'feature_code': 'PCLI'}}, {'word': 'United States', 'spans':
[{'start': 19, 'end': 32}], 'country_predicted': 'VIR', 'country_conf':
0.36183852}, {'word': 'China', 'spans': [{'start': 91, 'end': 96}],
'country_predicted': 'CHN', 'country_conf': 0.99992335, 'geo': {'admin1':
'NA', 'lat': '35', 'lon': '105', 'country_code3': 'CHN', 'geonameid':
'1814991', 'place_name': 'People's Republic of China', 'feature_class':
'A', 'feature_code': 'PCLI'}}]
```

在以上例子中，modecai中将所有的可能事件地理位置都进行了提取。哪一个才是核心这个问题还是存在一定的疑问的。

4. modecai本身的封装

modecai将三个步骤封装在一起。三个过程中分别得到的结果并没有办法很好的了解到。因此基于该算法源代码直接进行修改调整的工作非常困难。

基于以上的问题，使用modecai直接作为数据集构建的方式以失败告终。但是其思想还是可以被我们所参考。

2.spaCy库相关尝试

在经历modecai尝试后，我试图将其过程拆分，即原来的三个步骤，分别进行探索，来完成中文词典库的识别。其中发现 Prodigy这个标注是一个商业服务，其价格非常昂贵。因此先行进行了spaCy库的调研。

spaCy库是一个工业级的nlp分析库，但是其存在的问题与modecai一致。或者说modecai存在的问题便是spaCy库的问题：官方并没有推出中文库。但是spaCy支持引用外部数据集的方式来实现中文nlp的支持。因此我在github上寻找了一个针对spaCy开发的中文库，地址：https://github.com/howl-anderson/Chinese_models_for_SpaCy。通过应用这个库，最后实现句子的分析可以如下图所示：

	text	lemma_	pos_	tag_	ent_type_	dep_	dep
0	中国	中国	X	NNP	GPE	nsubj	抵达
1	抵达	抵达	VERB	VV		ROOT	抵达
2	苏瓦	苏瓦	X	NNP	GPE	compound	港
3	港	港	X	SFN	GPE	obj	抵达
4	,	,	X	,		punct	进行
5	对	对	X	IN		case	斐济
6	斐济	斐济	X	NNP	GPE	nsubj	进行
7	进行	进行	VERB	VV		advcl	提供
8	为期	为期	VERB	VV		acl:relcl	友好访问
9	8	8	NUM	CD	DATE	nummod	天
10	天	天	X	NNB	DATE	obj	为期
11	的	的	PART	DEC		mark:relcl	为期
12	友好访问	友好访问	X	NNP		obj	进行
13	并	并	X	RB		mark	提供
14	提供	提供	VERB	VV		ccomp	抵达
15	人道主义	人道主义	NOUN	NN		nmod	服务
16	医疗	医疗	NOUN	NN		nmod	服务
17	服务	服务	NOUN	NN		obj	提供

其中地理位置相关的信息用GPE标注了出来，还有相关的依赖的标注。但是这个语料库目前还在测试阶段，其完整性其实并不能很好的保证。在学长的建议下，我改用StanfordCoreNlp来实现这个任务。

P.S. 其实可以通过spaCy+stanford中文库来实现，但这是后来再发现的，便还没有进行过尝试。

3.StanfordNlp库相关尝试

本次试验中主要使用的是StanfordCoreNlp库中分词，词性依赖，词性标注以及语义标注工具。

```
tokenize = nlp.word_tokenize(sen)    #分词
depend = nlp.dependency_parse(sen)   #依赖
tag = nlp.pos_tag(sen)               #词性标注
ner = nlp.ner(sen)                   #命名实体标注
```

最后汇总得到矩阵如下：

```
原句：
贝尔格莱德报道台报道，多达100名阿尔巴尼亚族人在科索沃举行示威。

矩阵：
[[0, '贝尔格莱德', 'NR', 'CITY', 1], [1, '报道', 'VV', 'O', -1], [2, '台', 'NR', 'O', 3], [3, '报道', 'VV', 'O', 1], [4, '，', 'PU', 'O', 3], [5, '多达', 'AD', 'O', 6], [6, '100', 'CD', 'NUMBER', 'M', 'O', 6], [8, '阿尔巴尼亚', 'NR', 'NATIONALITY', 9], [9, '族人', 'NN', 'NATIONALITY', 12], [10, '在', 'P', 'O', 11], [11, '科索沃', 'NR', 'GPE', 12], [12, '举行', 'VV', 'O', 3], [13, '示 12], [14, '。', 'PU', 'O', 1]]

矩阵解释：
[序号，分词，词性（依存句法），命名实体标注，依赖] 当标注为root即句子核心的时候，数值为-1
```

在对于词义标注进行分析后，大致分类如下：

Named Entity Type	Examples
PERSON	President Obama, Franz Beckenbauer
ORGANIZATION	WHO, ISRO, FC Bayern
LOCATION	Germany, India, USA, Mt. Everest
DATE	December, 2016-12-25
TIME	12:30:00 AM, one thirty pm
MONEY	Twenty dollars, Rs. 50, 100 GBP
PERCENT	20%, forty five percent
FACILITY	Stonehenge, Taj Mahal, Washington Monument
GPE	Asia, Europe, Germany, North America

StanfordCoreNlp在以上基础上，地理位置还补充了标签：

“COUNTRY" "CITY" "STATE_OR_PROVINCE" "LOCATION"

在所有地理信息提取的过程中，我根据标签获取了所有地理位置的信息，并以csv的形式进行存储。

4.动词-地名树搭建

在论文“基于地名树的最佳空间尺度新闻事件地点提取方法”中提出了地名树的概念。通过提取不同的地点信息以及添加虚拟父节点的方式，搭建完成地名树，通过概率传递的方式完成地名的推测。这种方式给我留下了比较深刻的印象，论文中地名的提取，赋值是基于地名出现在文章中的不同位置进行处理的，如果在我们实验中也能搭建一棵类似的地名树，通过概率的方式传递相关参数，其最后的结果便也是我们所需要的事件地理位置，为了实现这个目标，我想到了将依存句法树与地名树结合的方式。

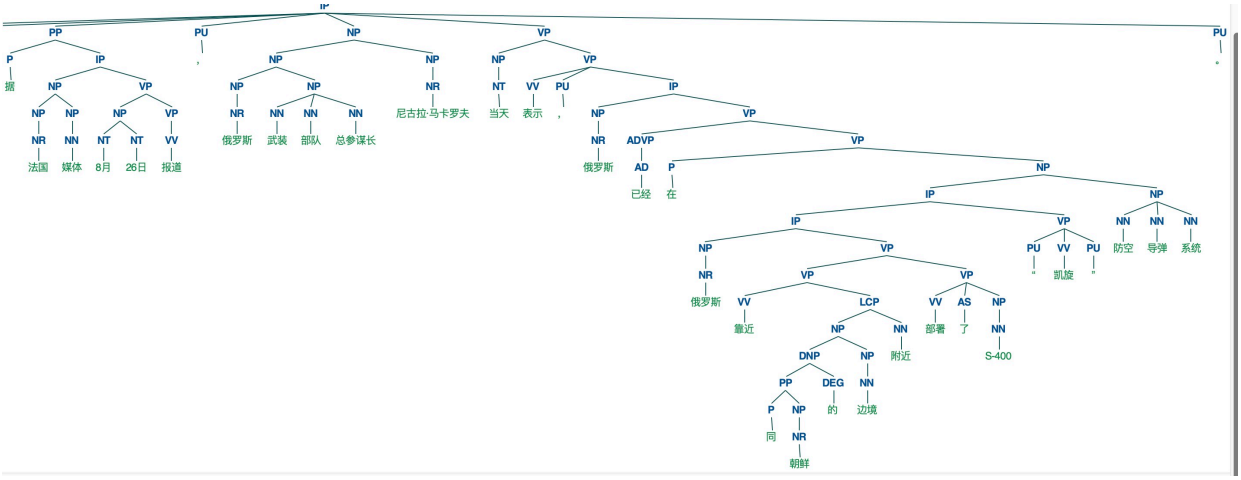
首先是对于依存句法树的相关属性进行了学习，大致类别在附录中记录。

在StanfordNlp实验的最后我通过nltk库对于整体进行可视化展示，以如下为例：

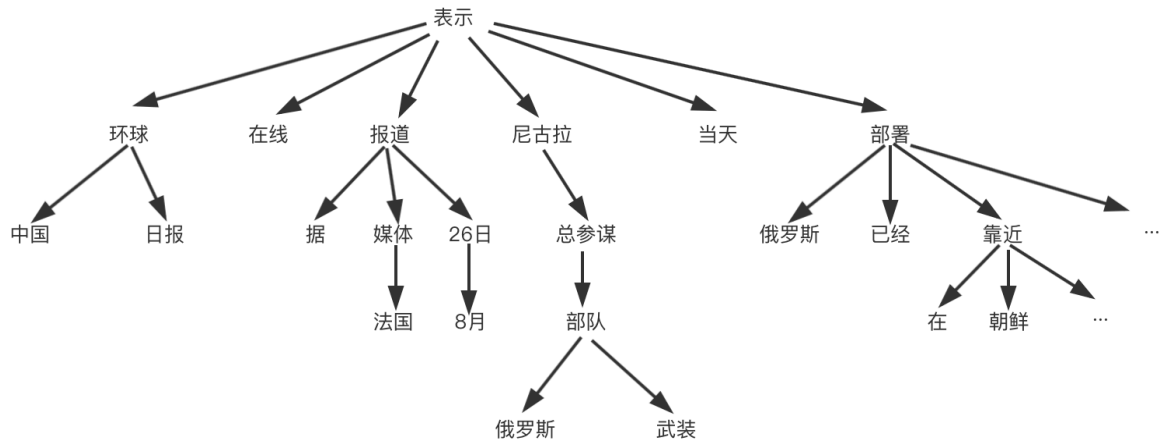
例句：
中国日报网环球在线报道：据法国媒体8月26日报道，俄罗斯武装部队总参谋长尼古拉·马卡罗夫当天表示，俄罗斯已经在俄罗斯靠近同朝鲜的边境附近部署了S-400“凯旋”防空导弹系统。

相关图表

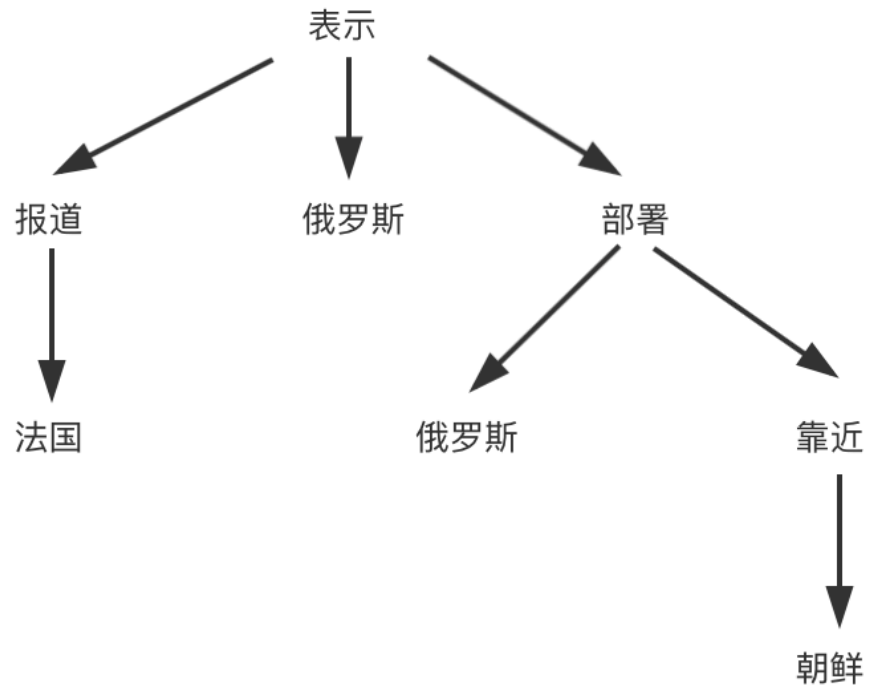
依存句法树：



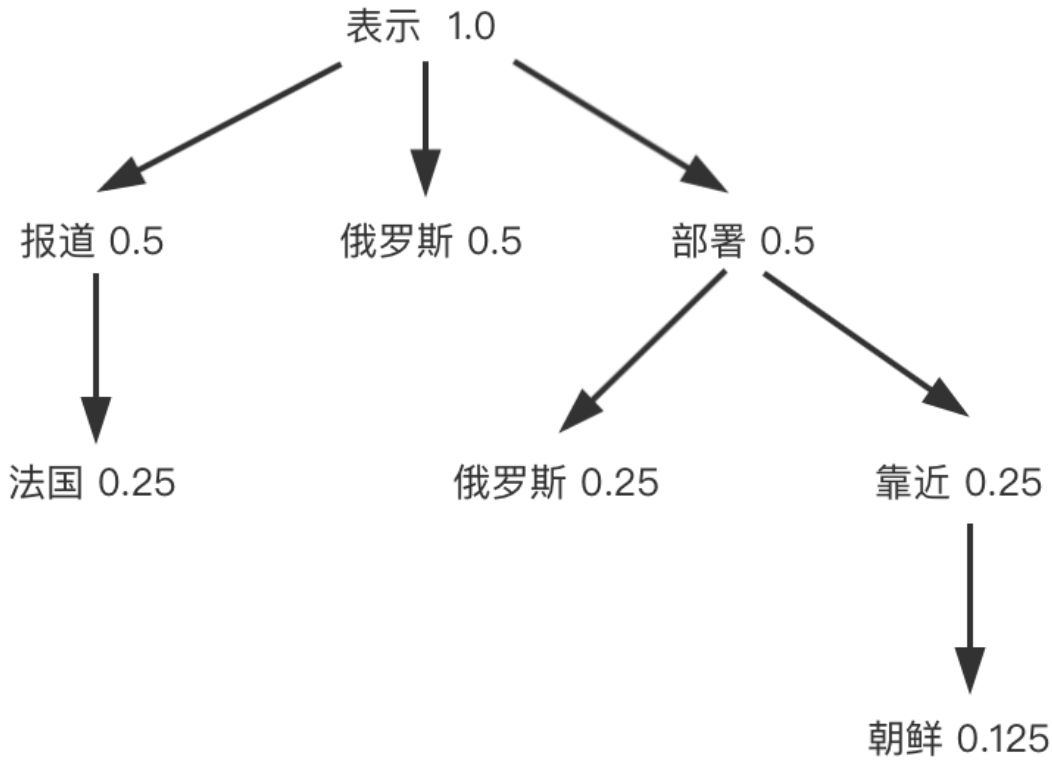
依赖关系图：



由于之前我们对于事件地理概念的定义为与事件地理核心动词最为相关的地理位置，在依存句法树以及依赖树中，相互存在依赖的地理位置以及动词便是很好的"相关"的数学表示。因此在依赖树的基础上，我去除了经过处理之后，我们可以得到所谓的动词-地名树。



打分机制目前设置的比较简单，以根节点为1分，每向下一层分值减半，如下图所示：



在赋值完成后，我建立了一个简易的地名实体消歧库，对于所有名称相同的地名节点进行合并，并记录最大数值点：

地名实体消歧库（部分）：

CF	中非共和国			
TD	乍得			
CL	智利			
CN	中国	中方	中	华
CO	哥伦比亚			
CG	刚果			
CK	库克群岛			

最后结果：

```
[['法国', 0.25, 12, 0.25, 7], ['俄罗斯', 0.75, 6, 0.75, 13], ['朝鲜', 0.125, 8, 0.125, 27]]
格式说明：
[地名，权值，位置，相同地名最大数值点，相同地名最大数值点离root距离]
```

相关问题

在实验中，我发现了一系列的问题：

1. 问题：root节点一定为动词吗?

   答： 不一定，经过统计，在我们新闻这个场景下，root节点为动词的概率大致为84%，为名词的概率大致为12%，剩下的为宾语及其他成分。其中名字做主语的情况下，地名作为主语的概率在90%。在算法中我的处理方式是如果地理名词直接为root，那这个地理名词直接为该事件的核心地理位置。如果为动词，则按照动词-地理树的形式进行计算比较。

2. 问题：报道这个词是一个非常明显的干扰词，在很多句子里面都存在以报道为root的情况。但是如此出来的地理位置不一定为核心地理位置。

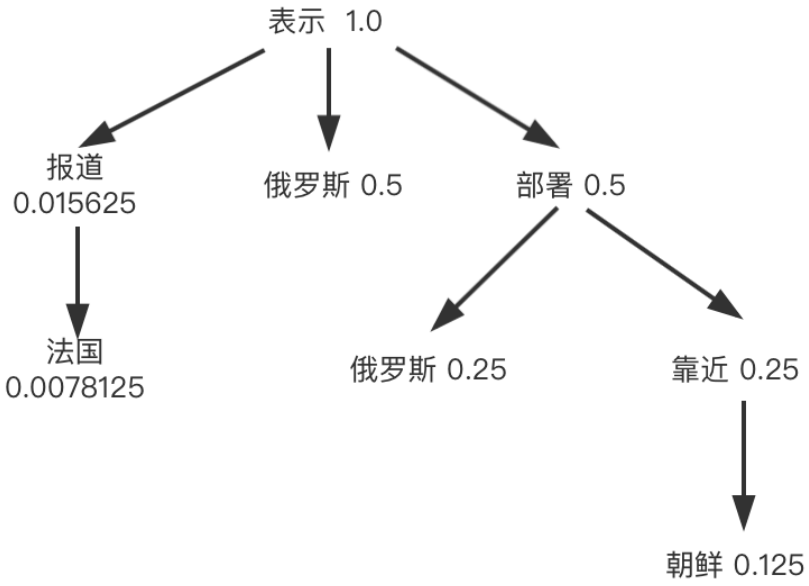
   答：我对于报道这个词进行了以下的处理：

   1. 首先我以全文本为数据集搭建了Word2Vec模型来训练词语相似度。由于有些文本采用报道，有些采用讯，电等词语，我用判断文本相似度的方式获取所有相关的词语进行统一处理，经过比较之后，将相似度标准定在0.4。
   2. 之所以不把报道进行一刀切的删除是因为在很多文本中，报道是唯一动词，报道前的地理名词确实是整段话的核心位置。因此我对于每个文本进行动词数量统计：
      1. 如果句中只存在报道这一个动词，则不予处理
      2. 如果句中不止报道一个动词且报道这个动词为root，则删除报道这个动词后，重新进

行句子成分判断

3. 如果句中报道这个动词不为root，将报道这个点的权重缩小至原来的1/32倍来起到反干扰的作用

因此最后的树为：



3. 问题：如果最后发现两个国家的权值相同怎么办

   答： 在modecai的论文中提出了一种概念叫做词间距，即动词与他相关的地理名词的距离会相对比较小。我也采用了这种思想，选取最近的地理位置进行存储。

实验结果

目前进行了所有数据的提取，抽取了前2000条进行目测，感觉准确率还行。正式的数值分析需要在人工完成标记后再进行比对。

部分实验结果：

中新网7月24日报道 据法新社报道，法国军演在马赛引发山火，当地时间23日晚，火势已蔓延约1300公顷，并造成10余间房屋和50辆汽车被毁。	11	马赛	0.5
美国反导系统测试成功击落短程目标导弹	0	美国	0.5
中新网7月31日报道 据美联社报道，美国于当地时间30日晚在夏威夷海域成功进行导弹防御系统测试。报道称，军方试射拦截导弹，将一枚短程目标导弹击落。	8	美国	0.5
印度拟再添125艘舰艇欲成世界第三大军事强国	0	印度	0.5
中新网8月2日报道 印度计划在未来十年内为其海军增加125艘战舰和潜水艇，并努力实现海军现代化和发展低成本造船实力。	4	印度	0.5

相关缺陷

这种方法还是偏向于规则制定的方式。对于数值的要求比较高，还需要针对不同的情况进行不断调整。

人工标注

目前的数据制造方法的准确率还是没有保障，希望通过人工标注数据的方式来达到最优效果以及检验。辛苦两位学弟了！

目前存在的相关问题

1. 其中一个比较核心的问题是，目前所做的一切包括未来需要做的机器学习都是基于分词以及一系列的词性标注为基础的。如果词性标注的存在偏差，最后的结果也会有偏差，比如说：

```
中新网7月31日报道 据美联社报道，美国于当地时间30日晚在夏威夷海域成功进行导弹防御系统测试。报道称，军方试射拦截导弹，将一枚短程目标导弹击落。
```

这个例子中，其事件地理位置应该为夏威夷海域，但实际结果却是美国。经过分析后我发现：

```
[[7, '，', 'PU', 'O', 11], [8, '美国', 'NR', 'COUNTRY', 11], [9, '于', 'P', 'O', 6], [10, '当地', 'NN', 'O', 4], [11, '时间', 'NN', 'O', 6], [12, '30日', 'NT', 'DATE', 6], [13, '晚', 'NT', 'TIM14, '在', 'P', 'O', 9], [15, '夏威夷', 'NR', 'FACILITY', 9], [16, '海域', 'NN', 'FACILITY', 11], [17, '成功', 'AD', 'O', 11], [18, '进行', 'VV', 'O', -1], [19, '导弹', 'NN', 'MISC', 14], [20, 'MISC', 14], [21, '系统', 'NN', 'MISC', 15], [22, '测试', 'NN', 'O', 11], [23, '。', 'PU', 'O', 11], [24, '新闻', 'NN', 'O', 18], [25, '称', 'VV', 'O', 11], [26, '，', 'PU', 'O', 18], [27, '军方', 21], [28, '试射', 'VV', 'O', 18], [29, '拦截', 'VV', 'O', 21], [30, '导弹', 'NN', 'O', 22], [31, '，', 'PU', 'O', 21], [32, '将', 'BA', 'O', 31], [33, '一', 'CD', 'NUMBER', 30], [34, '枚', '6], [35, '短程', 'JJ', 'O', 30], [36, '目标', 'NN', 'O', 30], [37, '导弹', 'NN', 'O', 31], [38, '击落', 'VV', 'O', 21], [39, '。', 'PU', 'O', 11]]
```

夏威夷海域被标注为Facility，最后的结果也自然没有他。

2. 对于事件地理位置的定义可能还需要打磨。比如如下几句话的事件地理位置该如何标注：

```
印度拟再添125艘舰艇欲成世界第三大军事强国
中国日报网环球在线报道：据英国媒体7月9日报道，俄罗斯军事杂志刊登的最新一份调查报告显示，2008年俄格战争期间俄罗斯损失的战机中有一半是被自己的防空系统击落的，俄军方对此予以坚决否认。
美军击落失控无人机防其飞入塔吉克斯坦或中国
```

时间安排

序号	日期	事项
1	3.16-3.19	modecai库相关尝试
2	3.20-3.22	spaCy库相关尝试以及中文语料库查询
3	3.23-3.24	stanfordNlp库相关尝试以及依存句法树理解
4	3.25-3.27	动词-地名树搭建以及代码编撰
5	3.26-3.27	人工数据标注检验以及算法调整以及文档编写

附录

依存句法树相关词性标注及分类

动词，形容词（4种）：VA，VC，VE，VV

1、谓词性形容词：VA

谓词性形容词大致上相当于英语中的形容词和中文语法中、文学作品里的静态动词。我们的谓词性形容词包括两类：

第一类：没有宾语且能被“很”修饰的谓语。

第二类：源自第一类的、通过重叠（如红彤彤）或者通过名词加形容词模式意味着“像N一样A”（如雪白）的谓语。这个类型的谓词性形容词没有宾语，但是有一些不能被“很”修饰，因为这些词的强调意思已经内嵌在词内了。

注意：当集合（VA）中的一个词修饰名词但没有用“的”，那么它被标注为JJ（名作定）或是一个名词，而不是VA。当集合（VA）中的一个词有一个宾语，那么它被标注为VV，而不是VA。譬如，这 项/M 活动 丰富/VV 了/AS 他 的/DEG 生活。

2、系动词：VC

“是”和“为”被标记为VC。如果“非”的意思是“不是”并且句子里没有其他动词时，“非”也被标注为VC。

“是”有几种用法：

·连接两个名词短语或者主语：他 是/VC 学生。

·在分裂句中：他 是/VC 昨天 来 的/SP。

·为了强调：他 是/VC 喜欢 看 书。

现在，在所有这些情况中，“是”被标注为VC。

3 、“有”作为主要动词：VE

只有当“有，没{有}”和“无”作为主要动词时（包括占有的“有”和表存在的“有”等等），被标注为VE。

4、其他动词：VV

VV包括其他动词，诸如情态动词，提升谓词（如“可能”），控制动词（如“要”、“想”），行为动词（如“走”），心理动词（如“喜欢”、“了解”、“怨恨”），等等。

名词（3种）：NR，NT，NN

1、专有名词：NR

专有名词是名词的子集。一个专有名词可以是一个特定的人名，政治或地理上定义的地方（城市、国家、河流、山脉等），或者是一种组织（企业、政府或其他组织实体）。一个专有名词通常是独一无二，并且不能被Det+M所修饰的。

·以下名字是专有名词：

地区/国家/村庄/城市，山脉/河流，报纸/杂志， 组织/公司，学校/联盟/基金会，个人/家庭。

·以下名字不是专有名词：

国籍（如中国人），种族（如白人），职称（如教授），疾病，职业，器官（如肺），乐器（如钢琴），游戏（如足球），花（如玫瑰），等等。

2、时间名词：NT

时间名词可以是介词的宾语，譬如在、从、到、等到。它们可以被问及，如“这个时候”，也可以被用以提问“什么时候”。它们也可以直接修饰VP（动词短语）或者S（主语）。像其他名词一样，时间名词可以是某些动词的论元。

时间名词可以是时间的名称（如1990年、一月、汉朝）或是由“PN+LC，N+LC，DT+N”等结构组成。

例子：一月、汉朝、当今、何时、今后

3、其他名词：NN

其他名词包括所有其他名词。其他名词NN，除了地方名词，一般不能修饰动词短语（有“地/DEV”没“地/DEV”）。

定位（1）：LC

方位词：LC

很多名词单独使用时不能作为介词如“在”、“到”的论元，也不能直接修饰VP（动词短语）或者S（主语）。方位词的一个功能是连接前述的名词短语或者主语，从而使整个短语可以作为这些介词的论元或者来修饰动词短语或主语。

一些方位词可以独立使用作为介词或动词的论元。一些方位词可以被“最”修饰。方位词不能被Det+M所修饰。

方位词分为两类：

·方位词：这类方位词表示方向、位置等。它们来自名词。一些可以单独使用作为介词或动词的论元。一些可以被“最”修饰。它们不能被Det+M所修饰。

—单音节方位词：如：前，后，里，外，内，北，东，边，侧，底，间，末，旁。

—双音节方位词：它们由以下部分组成：

*单音节方位词加上诸如“以、之”等的语素。 例子：之间，以北。

*两个单音节方位词。 例子：前后，左右，上下，东北。

·其他：我们把以下情况标注为LC。

. 为止：到 目前 为止。

. 开始：从 四月 开始。

. 来：5年 来。

. 以来： 1998年 以来。

. 起： 一九九三年 起。

. 在内：包括 他 在内。

代词（1种）：PN

代词的功能是作为名词短语的替代物或者表示事先详细说明的或者从上下文可知晓的被叫的人或事。它们一般不受Det+M或者形容词性短语修饰。

代词包括人称代词（如我、你），当作为名词短语单独使用时为指示代词（如这、那），所有格代名词（如其）以及反身代词（如我自己、自己）。

限定词和数词（3种）：DT，CD，OD

1、限定词：DT

限定词包括指示词（如这、那、该）和诸如“每、各、前、后”等词。限定词不包括基数词和序列词。

参见限定词部分。

2、基数词：CD

CD包括基数词并随意与一些概数词连用，如“来、多、好几”和诸如“好些、若干、半、许多、很多（如很多 学生）”等词。

例子：1245，一百。

3、序列词：OD

序列词被标注为OD。我们把第+CD看做一个词，并标注它为OD。

例子：第一百。

度量词（1）：M

度量词跟在数字后形成Det+M结构修饰名词或动词，包括类词（如“个”），表示一群的度量词，如“群”，以及公里、升等度量词。

一些度量词可以被有限的形容词（如一/CD**小/JJ**瓶/M水/NN），临时量词可以被名词和形容词修饰（如：一/CD**铁/NN**箱子/M书/NN）。

副词（1）：AD

副词包括情态副词、频率副词、程度副词、连接副词等，大部分副词的功能是修饰动词短语或主语。

如：仍然、很、最、大大、又、约

介词（1）：P

介词可以把名词短语或从句作为论元。

注释：把和被不标注为P，详见2.11部分。

如：从、对

连词（2）：CC，CS

1、并列连接词：CC

CC的主要模式是：XP{，}，CC XP。

如：与、和、或、或者、还是（or）

2、从属连词：CS

从属连词连接两个句子，一个句子从属于另一个，这样的连词标记为CS。CS模式是：CS S1，S2和S2 CS，S1。

如：如果/CS，……就/AD……

助词（8）：DEC，DEG，DER，DEV，SP，AS，ETC，SP，MSP

1、的作为补语标记/名词化标记：DEC（的，之）

如：吃的DEC

模式是：S/VP DEC{NP}

注：的还有其他标记

·DEC 他的/DEG车

·SP 他是/VC一定要来的/SP。

·AS 他是/VC在这里下的/AS车。

2、“的”作为关联标记或所有格标记：DEG

模式：NP/PP/JJ/DT DEG{NP}。

3、 补语短语 得：DER

在V-得-R和V-得结构中，“得”标记为DER。

注：有些以“得”结尾的搭配不是V-得结构，如记得，获得是动词。

4、方式“地”：DEV

当“地”出现在“XP地VP”，XP修饰VP。在一些古典文学中，“的”也用于这种情景，此时“的”也标注为DEV。

5、动态助词：AS

动态助词仅包括“着，了，过，的”。

6、句末助词：SP

SP经常出现在句末，如：他好吧[SP]?

有时，句末助词用于表停顿，如：他吧[SP]，人很好。

如：了，呢，吧，啊，呀，吗

7、 ETC

ETC用于标注等，等等。

8、其他助词：MSP

“所，以，来，而”，当它们出现在VP前时，标注为MSP。

所：他所[MSP]需要的/DEC

以或来：用……以/MSP（或来）维持

而：为……而[MSP]奋斗

其他（8）：IJ，ON，PU，JJ，FW，LB，SB，BA

1、感叹词：IJ

出现在句首位置的感叹词，如：啊。

2、拟声词：ON

① 修饰“ON地V”中的VP：雨哗哗[ON]地[DEV]下了[AS]一夜

② 修饰“ON中的N”中的NP：砰[ON]的/DEG一声！

③ 自行成句：砰砰[ON]！

④ 一般不能被副词修饰，如：哗啦啦，咯吱。

3、长“被”结构：LB

仅包括“被，叫，给，为（口语中）”，当它们出现在被字结构NP0+LB+NP1+VP中

如：他被/LB 我训了/AS 一顿/M .

注：当叫作为兼语动词时，“叫”标注为VV。

如：他叫/VV你去。

4、短“被”结构：SB（仅包括口语中的“被，给”）

NP0+SB+VP，他被/SB 训了/AS一顿/M。

注：“给”有其他标记：LB，VV和P。

如：你给/P他写封/M信。

5、把字结构：BA

仅包括“把，将”，当它们出现在把字结构中（NP0+BA+NP1+VP）。

如：他把/BA你骗了/AS。

注：“将”有其他标记：AD和VV，如：他将/VV了[AS]我的[DEG]军。

6、其他名词修饰语：JJ

包括三种类型：

①区别词 只修饰模式JJ+的+{N}或JJ+N中的名词，且一定要有“的”，它们不能被程度副词修饰。

如：共同/JJ的/DEG目标/NN，她是[VC]女/JJ的/DEG。

②带有连字符的复合词

通常为双音节词 JJ+N 如留美/JJ学者/NN

③形容词：新/JJ消息/NN

模式：JJ+N

注：当“的/DEC”在形容词和名词中间时，形容词标记为VA。

7、外来词：FW

FW仅被用于：当词性标注标记在上下文中不是很清楚时。外来词不包括外来词的翻译，不包括混合中文的词（如卡拉OK/NN，A型/NN），不包括词义和词性在文中都是清楚的词。

8、标点：PU

当标点是词的一部分时，不用标注为PU，如123,456/CD。