本周报告

> 2020/03/05 1652792 罗吉皓

本周主要做了毕业设计的事件提取调研以及地理信息算法调研，分为以下几个部分：

1. 事件提取相关资料调研
2. 地理信息Leetura算法及相关模型调研
3. 中文触发词字典构建调研

事件提取相关资料调研

这一块的工作主要是在看坚果云上的资料。核心了解内容为两方面：

1. Petrarch字典调研，了解了actor字典，verb字典的组成部分
2. Petrarch程序代码调研，了解了整个时间提取的流程，从文本分词，到标注不同词性，最后完成不同组成部分信息的提取。

这一块主要是了解了实验室目前项目的进度，学习内容在这里便不多过多赘述。

地理信息Leetura算法及相关模型调研

现有模型分析

目前许多现有的开源地理位置事件数据库对于地理信息位置的标注都没有明确地链接事件和地点。最简单的地理信息位置的标注是使用受过训练的人工编辑人员来阅读每个文档（或有时只是其标题或摘要）并手动分配地理标签。这种传统方法根本无法适应现代档案的规模和速度，因为现代档案可能包含数千亿个页面以及数百万个页面。此外，手动地理标签只能索引文档的主要焦点：对于人类索引器来说，索引一本600页的书中每个位置的索引根本不可行。因此，自动化的软件算法正在取代人类索引器，成为大型文本档案进行地理索引的主要机制。

诸如文本工程通用体系结构（GATE）之类的工具包使用的最常见方法是简单地执行关键字搜索，以搜索国家，其首都和主要城市的名称。尽管这种方法非常容易实现且运行速度很快，但由于多种原因，它还是存在一系列问题的。最大的限制之一是许多新闻文章都假定其预期的读者有有关位置的常识。例如，伊利诺伊州尚佩恩市的一家当地报纸在提及邻近城市厄巴纳时，只提到“ Urbana”，而不是“美国伊利诺伊州Urbana”。有研究显示，美国主流媒体报道中提到的美国城市中有68%没有在附近添加州名或其他上下文信息。即使找到匹配项，最终的地理索引也仅允许国家/地区级别的搜索，而缺乏相关更为详细的信息。另外，并非所有国家/地区名称都是明确的：“格鲁吉亚当局”既可以指美国州，也可以指欧洲国家。

基于现有模型的一系列缺陷，全文地理编码的新兴领域“地理信息检索”（GIR）诞生了。全文地理编码的核心是扫描文本主体以识别潜在的地理参考，然后使用外部知识库（称为“地名词典”）和文档上下文来消除歧义并将参考转换为地理空间形式。GIR实际上是指从文本中提取地理对象，将其索引为空间索引并允许使用空间信息对语料库进行空间搜索的整个流程。

mordecai模型分析

mordecai模型介绍

mordecai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。

mordecai与其他现有模型相比，最大的优势是它将事件与事件发生的地点联系在一起。spaCy分词与Prodigy词性标注很好的帮助mordecai将事件，动词，地理位置等信息标注出来，而后通过机器学习的算法将其训练联系在一起。mordecai算法在设计的过程中运用了三个假设：

1. 假设事件的动词提供的信息足够。换句话说我们可以用动词来定位事件，这是语义角色标签中的一个常见假设；
2. 假设我们可以通过合适的方式来将事件中的词汇进行降维分析；
3. 假设事件中的词语标签与事件中的地理位置表浅是相互独立的。

假设一为机器学习通过事件的动词与事件位置之间的绑定打下了基础。mordecai作者利用一组8000个带有手工标记的句子，训练了一个递归神经网络，该神经网络利用丰富的语言功能集为文本序列添加标签，以标识该单词是否是与指定动词相对应的位置词。

假设二与假设三将事件分类预测的工作进行了简化，方便训练模型的搭建即：

假设二：

$$\Phi(X) = \{\phi(w_1), \ldots, \phi(w_n)\}.$$

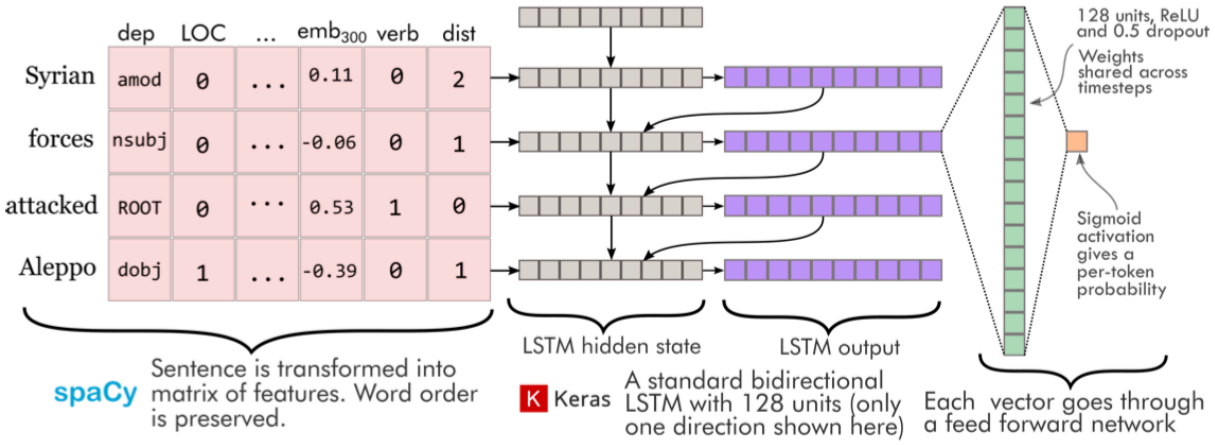
$$\hat{\boldsymbol{y}}^{(k)} = \hat{f}(\Phi(\boldsymbol{X}), v_k).$$

假设三：

$$\hat{\boldsymbol{y}}^{(k)} = \hat{f}(\Phi(\boldsymbol{X}), v_k)$$
$$= \{\hat{f}(\phi(w_1), v_k), \ldots, \hat{f}(\phi(w_n), v_k)\}$$

在机器学习模型的搭建上，作者选择了lstm网络， LSTM以隐藏矢量的形式在输入数据的每个步骤中存储内部状态。与普通RNN相比，LSTM可以学习何时从其当前输入步骤向隐藏状态添加信息，以及何时从隐藏状态“获取”信息。从理论上讲，这使LSTM学习的关系比他们原本所能学习的要长得多。当模型可以访问“未来”，并且为每个输入步骤计算两个状态向量时，双向LSTM是LSTM的标准扩展：左起一个，右起一个。将这两个向量连接起来，并用作模型其余部分的输入。

相关网络结构图如下：



mordecai最后的回归率可以达到0.83左右，相比于其他的方法，准确性有更好的保证。

Model	Prec	Rec	F1	Sentence
Baseline	0.29	0.25	0.27	0.28
Profile	0.54	0.29	0.37	0.51
PropBank[8]	*0.61*	*0.39*	*0.47*	-
CNN	0.70	0.54	0.61	-
Chung et al.[9]	*0.74*	*0.62*	*0.62*	-
Annotator	**0.88**	0.65	0.74	0.73
LSTM	0.85	**0.83**	**0.84**	**0.77**

Table 1: Per-token precision, recall, and F1 scores, and full-sentence accuracy for the word distance baseline model, expected human performance, existing results from the literature, and new model-based approaches.

具体实例如下：

```
原句：
He was speaking a day after Ankara launched an offensive in the Syrian towns of Jarablus

经过spaCy：
He was speaking a day after Ankara [launched VERB] an offensive in the Syr- ian towns of [Jarablus EVENT LOC]

最后结果：
[{'word': 'Ankara', 'spans': [{'start': 28, 'end': 34}], 'country_predicted': 'TUR', 'country_conf': 0.3727025},
{'word': 'Jarablus', 'spans': [{'start': 80, 'end': 88}], 'country_predicted': 'SYR', 'country_conf': 0.99723154,
'geo': {'admin1': 'Aleppo', 'lat': '36.8175', 'lon': '38.01111', 'country_code3': 'SYR', 'geonameid': '169179', 'place_name': 'Jarābulus', 'feature_class': 'P', 'feature_code': 'PPLA2'}}]
```

mordecai使用方法

第一步 翻译

由于mordecai训练集与测试集皆为英语文本，我对此做了数据预处理。

我使用了有道词典的api来实现中文文本与英语文本之间的转换

测试例句如下：

```
“前锋-2009A”陆空联合火力打击演习将在济南军区某合同战术训练基地开演，来自七十多个国家的一百五十多名外军留学生以及英国、以色列、土耳其等国军事观察员将实地观摩。演习总导演、济南军区某集团军参谋长胡修斌少将二十日透露，这次演习的特色和亮点就是“联合作战”和“精确作战”，体现信息化条件下作战联合制胜和联合作战的思想，并且着力使这一思想由战役军团级以上单位向战术兵团乃至分队行动延伸和发展。

翻译结果如下：
"Forward - 2009 - a" aviation joint fire drill will be a contract tactical training base in jinan military region, more than one hundred and fifty foreign students from more than seventy countries and the United Kingdom, Israel, Turkey and other countries, military observers will field view, director of exercises, jinan military area command an army chief of staff Hu Xiubin general 20, according to the characteristics of the drill and bright spot is the "joint operations" and "accurate", embodied under the condition of informatization combat united winning and the thought of joint operations, and strive to make this idea by battle legion magnitude unit to tactical corps and team action extension and development.
```

本程序调用有道词典的API进行翻译，可达到以下效果：
外文-->中文
中文-->英文
请输入你想要翻译的词或句：“前锋-2009A”陆空联合火力打击演习将在济南军区某合同战术训练基地开演，来自七十多个国家的一百五十多名外军留学生以及英国、以色列、土耳其等国军事观察员将实地观摩，演习总导演、济南军区某集团军参谋长胡修斌少将二十日透露，这次演习的特色和亮点就是“联合作战”和“精确作战”，体现信息化条件下作战联合制胜和联合作战的思想，并且着力使这一思想由战役军团级以上单位向战术兵团乃至分队行动延伸和发展。
输入的词为：“前锋-2009A”陆空联合火力打击演习将在济南军区某合同战术训练基地开演，来自七十多个国家的一百五十多名外军留学生以及英国、以色列、土耳其等国军事观察员将实地观摩，演习总导演、济南军区某集团军参谋长胡修斌少将二十日透露，这次演习的特色和亮点就是“联合作战”和“精确作战”，体现信息化条件下作战联合制胜和联合作战的思想，并且着力使这一思想由战役军团级以上单位向战术兵团乃至分队行动延伸和发展。
翻译结果为："Forward - 2009 - a" aviation joint fire drill will be a contract tactical training base in jinan military region, more than one hundred and fifty foreign students from more than seventy countries and the United Kingdom, Israel, Turkey and other countries, military observers will field view, director of exercises, jinan military area command an army chief of staff Hu Xiubin general 20, according to the characteristics of the drill and bright spot is the "joint operations" and "accurate", embodied under the condition of informatization combat united winning and the thought of joint operations, and strive to make this idea by battle legion magnitude unit to tactical corps and team action extension and development.

第二步 安装mordecai

1. 安装mordecai

```
pip install mordecai
```

2. 下载 spaCy NLP model:

```
python -m spacy download en_core_web_lg
```

3. docker获取Elasticsearch镜像，将geoname的索引目录导入Elasticsearch.

```
docker pull elasticsearch:5.5.2
wget https://s3.amazonaws.com/ahalterman-geo/geonames_index.tar.gz --output-file=wget_log.txt
tar -xzf geonames_index.tar.gz
docker run -d -p 127.0.0.1:9200:9200 -v $(pwd)/geonames_index/:/usr/share/elasticsearch/data elasticsearch:5.5.2
```

遭遇的坑

以上安装的所有的库，文件都特别大，除了docker拉取镜像还算顺利以外，其他基本下载不下来。最后我基本采用的都是先下载轮子，最后再加入本地库的方法。其中特别需要注意的一点是en_core_web_lg-2.0.0的引入。

在下载完成解压文件后，找到"path\en_core_web_lg-2.0.0\en_core_web_lg\en_core_web_lg-2.0.0\tokenizer",它的上一次即为模型目录

我们可以直接加载：

```
nlp = spacy.load("path\en_core_web_lg-2.0.0\en_core_web_lg\en_core_web_lg-2.0.0")
```

或者创建用户快捷方式

> python -m spacy link [package name or path][shortcut] [--force]

PS：上述指令将会在spacy/data目录下创建模型的快捷方式。第一个参数是模型名词（模型已经通过pip安装），或模型存放目录。第二个参数是你想使用的内部名词。设置--force标记将会强制覆盖已存在的连接。举例如下：

> # 为已经安装的模型设置快捷方式"en_default"
>
> python -m spacy link en_core_web_md en_default
>
> # 为本地存储的模型设置快捷方式 "my_amazing_model"
>
> python -m spacy link /Users/you/model my_amazing_model

相关结果展示：

```
Using TensorFlow backend.
本程序调用有道词典的API进行翻译，可达到以下效果：
外文--->中文
中文--->英文
请输入你想要翻译的词或句：“前锋-2009A”陆空联合火力打击演习将在济南军区某合同战术训练基地开演，来自七十多个国家的一百五十多名外军留学生以及英国、以色列、土耳其等国军事观察员将实地观摩，演习总导演、济南军区某集团军参谋长胡修斌少将二十日透露，这次演习的特色和亮点就是“联合作战”和“精确作战”，体现信息化条件下作战联合制胜和联合作战的思想，并且着力使这一思想由战役军团级以上单位向战术兵团乃至分队行动延伸和发展。
输入的词为：“前锋-2009A”陆空联合火力打击演习将在济南军区某合同战术训练基地开演，来自七十多个国家的一百五十多名外军留学生以及英国、以色列、土耳其等国军事观察员将实地观摩，演习总导演、济南军区某集团军参谋长胡修斌少将二十日透露，这次演习的特色和亮点就是“联合作战”和“精确作战”，体现信息化条件下作战联合制胜和联合作战的思想，并且着力使这一思想由战役军团级以上单位向战术兵团乃至分队行动延伸和发展。
翻译结果为："Forward - 2009 - a" aviation joint fire drill will be a contract tactical training base in jinan military region, more than one hundred and fifty foreign students from more than seventy countries and the United Kingdom, Israel, Turkey and other countries, military observers will field view, director of exercises, jinan military area command an army chief of staff Hu Xiubin general 20, according to the characteristics of the drill and bright spot is the "joint operations" and "accurate", embodied under the condition of informatization combat united winning and the thought of joint operations, and strive to make this idea by battle legion magnitude unit to tactical corps and team action extension and development.
WARNING:tensorflow:From /Users/logan/venv/common/lib/python3.7/site-packages/tensorflow/python/framework/op_def_library.py:263: colocate_with (from tensorflow.python.framework.ops) is deprecated and will be removed in a future version.
Instructions for updating:
Colocations handled automatically by placer.
WARNING:tensorflow:From /Users/logan/venv/common/lib/python3.7/site-packages/keras/backend/tensorflow_backend.py:3445: calling dropout (from tensorflow.python.ops.nn_ops) with keep_prob is deprecated and will be removed in a future version.
Instructions for updating:
Please use `rate` instead of `keep_prob`. Rate should be set to `rate = 1 - keep_prob`.
2020-03-05 17:35:25.088982: I tensorflow/core/platform/cpu_feature_guard.cc:141] Your CPU supports instructions that this TensorFlow binary was not compiled to use: AVX2 FMA
WARNING:tensorflow:From /Users/logan/venv/common/lib/python3.7/site-packages/tensorflow/python/ops/math_ops.py:3066: to_int32 (from tensorflow.python.ops.math_ops) is deprecated and will be removed in a future version.
Instructions for updating:
Use tf.cast instead.
[{'word': 'jinan', 'spans': [{'start': 92, 'end': 97}], 'country_predicted': 'CHN', 'country_conf': 0.9886789, 'geo': {'admin1': 'Shandong', 'lat': '36.66833', 'lon': '116.99722', 'country_code3': 'CHN', 'geonameid': '1805753', 'place_name': 'Jinan', 'feature_class': 'P', 'feature_code': 'PPLA'}}, {'word': 'United Kingdom', 'spans': [{'start': 205, 'end': 219}], 'country_predicted': 'GBR', 'country_conf': 0.9886789, 'geo': {'admin1': 'NA', 'lat': '54.75844', 'lon': '-2.69531', 'country_code3': 'GBR', 'geonameid': '2635167', 'place_name': 'United Kingdom of Great Britain and Northern Ireland', 'feature_class': 'A', 'feature_code': 'PCLI'}}, {'word': 'Israel', 'spans': [{'start': 221, 'end': 227}], 'country_predicted': 'ISR', 'country_conf': 0.99996746, 'geo': {'admin1': 'NA', 'lat': '31.5', 'lon': '34.75', 'country_code3': 'ISR', 'geonameid': '294640', 'place_name': 'State of Israel', 'feature_class': 'A', 'feature_code': 'PCLI'}}, {'word': 'Turkey', 'spans': [{'start': 229, 'end': 235}], 'country_predicted': 'TUR', 'country_conf': 0.9986525, 'geo': {'admin1': 'NA', 'lat': '39', 'lon': '35', 'country_code3': 'TUR', 'geonameid': '298795', 'place_name': 'Republic of Turkey', 'feature_class': 'A', 'feature_code': 'PCLI'}}, {'word': 'jinan', 'spans': [{'start': 316, 'end': 321}], 'country_predicted': 'CHN', 'country_conf': 0.9886789, 'geo': {'admin1': 'Shandong', 'lat': '36.66833', 'lon': '116.99722', 'country_code3': 'CHN', 'geonameid': '1805753', 'place_name': 'Jinan', 'feature_class': 'P', 'feature_code': 'PPLA'}}]
```

```
[{'word': 'jinan', 'spans': [{'start': 92, 'end': 97}],
'country_predicted': 'CHN', 'country_conf': 0.9886789, 'geo': {'admin1':
'Shandong', 'lat': '36.66833', 'lon': '116.99722', 'country_code3':
'CHN', 'geonameid': '1805753', 'place_name': 'Jinan', 'feature_class':
'P', 'feature_code': 'PPLA'}},
{'word': 'United Kingdom', 'spans': [{'start': 205, 'end': 219}],
'country_predicted': 'GBR', 'country_conf': 0.9886789, 'geo': {'admin1':
'NA', 'lat': '54.75844', 'lon': '-2.69531', 'country_code3': 'GBR',
'geonameid': '2635167', 'place_name': 'United Kingdom of Great Britain
and Northern Ireland', 'feature_class': 'A', 'feature_code': 'PCLI'}},
{'word': 'Israel', 'spans': [{'start': 221, 'end': 227}],
'country_predicted': 'ISR', 'country_conf': 0.99996746, 'geo': {'admin1':
'NA', 'lat': '31.5', 'lon': '34.75', 'country_code3': 'ISR', 'geonameid':
'294640', 'place_name': 'State of Israel', 'feature_class': 'A',
'feature_code': 'PCLI'}},
{'word': 'Turkey', 'spans': [{'start': 229, 'end': 235}],
'country_predicted': 'TUR', 'country_conf': 0.9986525, 'geo': {'admin1':
'NA', 'lat': '39', 'lon': '35', 'country_code3': 'TUR', 'geonameid':
'298795', 'place_name': 'Republic of Turkey', 'feature_class': 'A',
'feature_code': 'PCLI'}},
{'word': 'jinan', 'spans': [{'start': 316, 'end': 321}],
'country_predicted': 'CHN', 'country_conf': 0.9886789, 'geo': {'admin1':
'Shandong', 'lat': '36.66833', 'lon': '116.99722', 'country_code3':
'CHN', 'geonameid': '1805753', 'place_name': 'Jinan', 'feature_class':
'P', 'feature_code': 'PPLA'}}]
```

可以看到其中提出了济南，英国，以色列，土耳其等地理信息，与人工查找的信息基本相同。

相关问题

mordecai在一定程度上实现了事件地理信息的提取，不过这种事件与地理位置绑定的方式还是存在一定的缺陷的：

1. 其地理信息依赖于geoname库，geoname中不存在的词条不能够显示，比如：

```
He was speaking a day after Ankara launched an offensive in the Syrian
towns of Jarablus and Kobane
```

Kobane这个词在geoname中便没有收录，最后即使分词的时候指定了Kobane为Event location也没有办法最后展示出来。

2. mordecai无法进行全文信息的归整。比如埃及开罗与美国伊利诺伊州开罗市，mordecai默认就是埃及开罗
3. 由于其分词机制，一些以地区作为形容词的表述将会被忽略。比如“据美国媒体报道”，美国这个词就会被忽略。这一块的信息其实还是有一定意义的，可以被收集起来

综上，mordecai在一定程度上能够实现事件地理信息的提取，在分词，事件绑定等方面做的很出色，不过最后信息对应直接从geoname索引中直接获取有点草率，如果能综合全体文本的信息对于多含义信息进行适当调整就更好了。

Leetura

Leetura算法是Leetura教授在2012年提出的一种新型的全文地理编码器，其核心思想如下：

确定潜在的候选人

全文地理编码器的第一步是解析文档文本，以识别可能是地理参考的单词和短语。最基本的方法是提取所有大写的词组，然后加上诸如“ in”或“ near”之类的“触发词”。许多位置名称还包括常见的单词，例如形容词，动词，代词和其他词性，作为名称的一部分，例如泰国的“ Matsayit Nu Run I Man”，这可能会导致标记者将名称拆分成多个部分。此外，由非英语母语人士撰写或翻译的文档也会在语音标记过程中引起错误。最后，部分语音标记器仅适用于少数几种语言，从而将这种方法限制于英语和西班牙语等主要语言。

相比之下，Leetura使用的算法则相反：不是从文本中提取潜在的地理名称，而是消除了所有不能作为位置的文本。为此，将几个标准的Linux拼写检查器字典组合在一起，以生成所有常见英语单词的列表。将它们与地理地名词典进行了比较，以删除所有也是位置名称或位置名称一部分的单词。最终结果是一个非地理单词列表。首先将源文档文本与此列表进行比较，并将所有匹配的单词转换为一个句点，以使“昨天在加拿大到商店去的加拿大”这样的句子变成“加拿大……”。

一些常见的英语单词也可以出现在位置中，例如单词“ run”。为了解决这些问题，使用了英语词典中出现的所有单词的列表以及一个或多个地理名称的所有单词的列表来编制模糊单词的列表。然后搜索英文版的Wikipedia，以汇编出每个单词的所有匹配项的列表，并计算出大写和小写形式出现的百分比。这存储在下一阶段使用的概率表中。

检查候选人的潜在匹配

一旦潜在的地名列表被编译，下一步就是搜索那些地名，例如某些国家名，首都和主要城市。对于“华盛顿”（可能是美国州或该州的首都）或“乔治亚州”（可能是美国州或欧洲的国家）等含糊不清的引用，系统会将这两种可能性都记录为匹配项，然后在地理编码过程结束时返回以确定是否从其中一种可能的匹配项中找到了任何城市，如果没有找到，则将其删除。如果一篇文章提到佐治亚州，然后又提到亚特兰大，系统将把比赛保存到美国州，并放弃国家比赛。

如果未发现某个候选人是某个国家，首都或主要城市，则将检查该候选人是否在全球范围内有该名称的地点。如果该名称有任何位置，则将候选人保存为可能的匹配项并转发到下一个阶段。如果未找到匹配项，则系统使用较早的概率查找表来检查候选者的第一个单词和最后一个单词。在Wikipedia中出现比例最高的单词为小写字母，将其从候选单词中删除并重试。为了演示此过程的工作原理，当向候选人显示“纽约市警察局”时，系统找不到匹配项，然后比较第一个单词“ new”和最后一个单词“ station”，并发现“ station”出现在Wikipedia中小写字母比大写字母更常见。然后，它下放电台以生成新的候选短语“纽约市警察”。这也没有匹配，因此将“ new”与“ police”进行比较，警察被撤职，产生了最终匹配的“ New York City”的新候选人。

该系统还使用单独的算法编译文本中的人名列表，并删除也是人名的位置候选。这样可以避免提及“ Paris Hilton”与巴黎相匹配，但允许“ Paris Hilton hotel”与之匹配。最后，使用手动编译的“黑名单”来删除误报率较高的位置。例如，虽然提到“上帝”的确可能是指埃塞俄比亚的城市，或“地狱”是指美国的城市，但这些名称的绝大多数出现都与地球上的位置无关，因此手动黑名单允许删除这些常见的误报。同样，有些国家/地区也以著名人物或组织的名字命名，例如Castro，古巴或Duma。

地名词典

为了识别给定的候选人是否可能是地球上的某个位置，需要一个数据库来记录地球上所有已知地点的名称及其大概位置的列表，称为“地名词典”。许多国家/地区特定的地名词典提供了其土地区域的高分辨率覆盖。澳大利亚政府提供了一个在澳大利亚的322,000个地点的数据库，而加拿大政府则提供了一个类似的350,000个地点的数据库。全球地名词典将世界上所有可用的地理信息汇总到一个数据库中。地名盖蒂词库在地球上大约900,000个不同的地方包含超过100万个条目，其中包括至少可以追溯到罗马时代的历史名称。亚历山大数字图书馆项目地名词典在其数据库中有590万个名称。GeoNames.org网站将尽可能多的国家地名词典汇集在一起，以提供约800万个地名的数据库。

消除和确认候选人

地理编码的最后阶段是对早期阶段的候选者进行消歧和确认。语言非常模棱两可；根据上下文的不同，单词的含义可能完全不同。另一方面，地理坐标是明确的：一组坐标和必要的投影信息唯一地定义了空间中的单个位置。因此，该阶段必须利用语言环境将文本引用歧义化为明确的地理表示形式。例如，对“城市”的引用可以指委内瑞拉的一个城市，也可以指美国的11个城市之一。但是，如果本文的其余部分提到芝加哥，斯普林菲尔德和迪凯特，则可能是指美国伊利诺伊州的厄巴纳。

中文触发词字典构建调研

CAMEO中文化字典调研

这一块主要分为两个部分： 寻找官方字典和目前中文词典的寻找。

官方字典

CAMEO中存在一定的中文信息存储，目前最核心的处理方式还是通过翻译，并没有核心的中文词典库的发布。

或许是由于中东的形式，最近官方推出了阿拉伯语库，可能在未来中文库也会推出的。

---

arabic_dictionaries

Arabic language actor and verb dictionaries for CAMEO-style event data

Jupyter Notebook · MIT · 1 · ★ 2 · 1 · 0 · Updated Jan 30, 2019

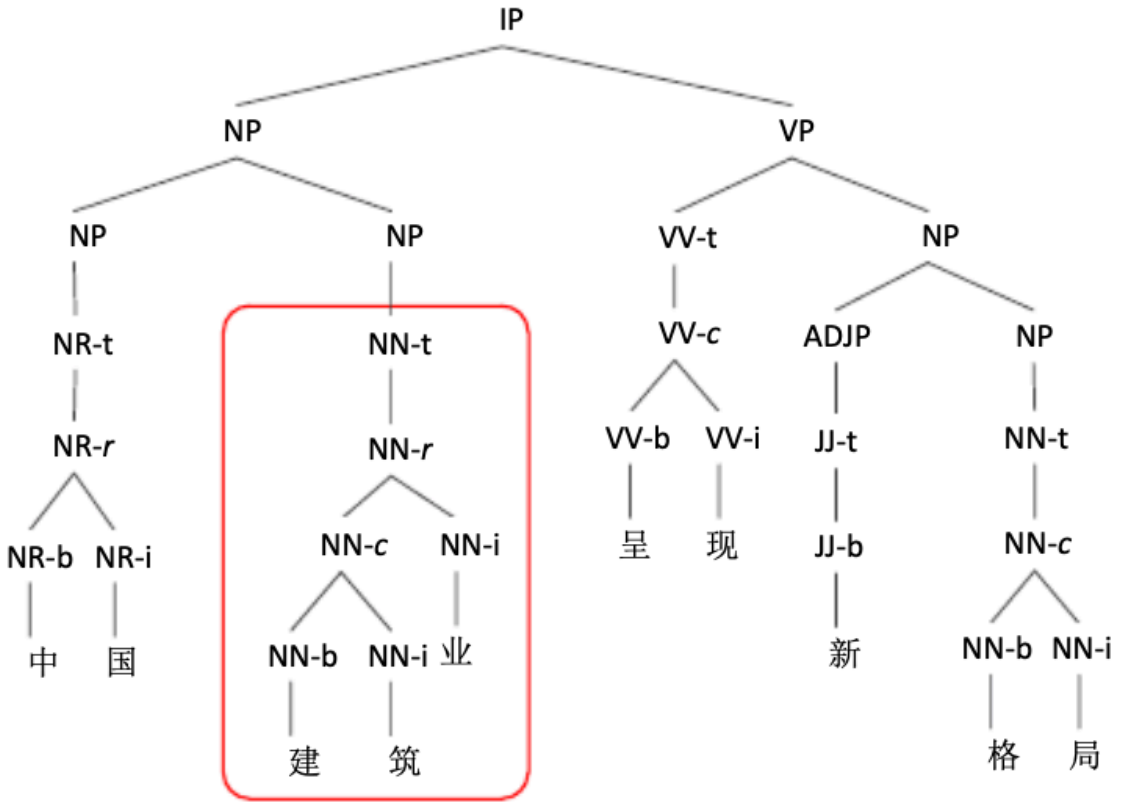
目前中文词典

在*Chinese Parsing Exploiting Characters*论文中，Zhang ，Wang等定义了一套面向中文词内部层次结构解析的

标准。该标准主要包含两部分：

1. 从语义组合的角度对中文词内部结构基本语义单元类型进行了界定。 具体而言， 定义的语义单元类型（节点类型）有两种：原子子词 (atom subwords) 和组合子词(composition subwords)， 标签分别为“ a ”和“ c ”。 原子子词的词义不能由其构成单元的语义直接组合而成， 例如， “蜻蜓”。 与之相反， 组合子词的词义则可以由其构成单元的语义组合而成， 例如，“副总统”。
2. 从词义偏向关系的角度定义了构成单元语义之间的四种关系：偏左（语义单元的中心部分在左边）、偏右（语义单元的中心部分在右边）、并列（两部分是并列关系）和无偏向关系， 其标签分别为“ l”， “ r”， “ p”， “ n”。

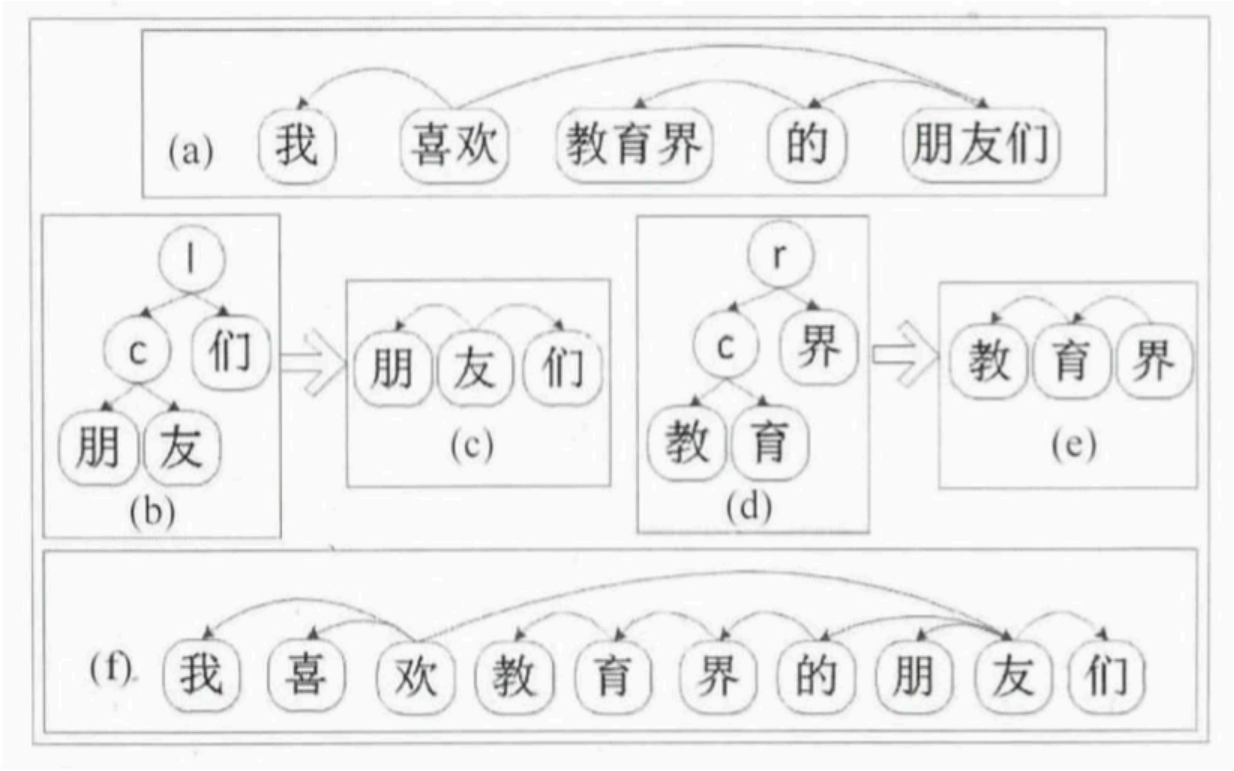
> 论文地址如下：https://www.aclweb.org/anthology/P13-1013.pdf

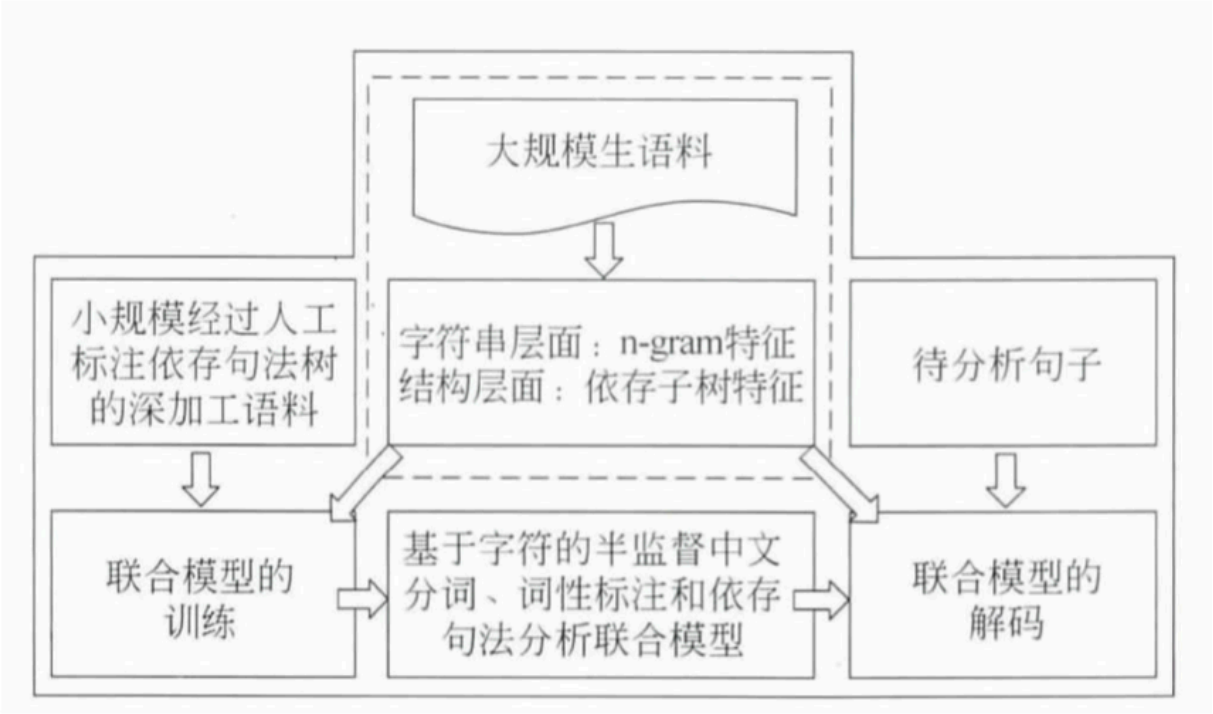


传统的中文自然语言处理任务或者先进行分词， 然后直接以词为基本处理单元， 或者直接以字符为基本处理单元。 上述两种处理方式有着各自的优缺点。 一方面， 以词为单元保留了自然语言的语义信息， 一个词就是一个基本的语义单元， 能够很大程度上降低对句子理解上的歧义， 但是存在数据稀疏问题。 例如， “访问者”， 即使是词表之内的词， 由于其在语料中出现的频率很低， 模型仍然难以准确学到它的语义信息；另一方面， 以字符为单元虽然能够解决数据稀疏问题， 但是字符相比于词歧义更大， 同时句子输入序列过长， 因此无法充分利用中文原子词的语义信息。 例如：“蜻蜓”， 它的语义由字符序列作为整体构成， 而无法由字符语义组合而成。 可见， 如何探索适合中文自然语言处理的基本语义单元具有重要的研究意义。

基于字符的中文分词、词性标注和依存句法分析联合模型论文在以上的基础上做了一系列的改进，提出一种半监督的中文分词、词性标注和依存句法分析联合模型， 将Zhang 标注的词语内部结构转化为依存结构，将传统的基于词语的依存句法树扩展成了基于字符的依存句法树，在此基础上采用增量转移策略实现了真正意义上的基于字符的中文分词、词性标注和依存句法分析联合模型并参考中文分词的序列标注思想，将中文分词的转移策略拓展为4种动作：Shift_S，Shift_M，Shift_B和Shift_E。

> 论文地址如下：http://www.cnki.com.cn/Article/CJFDTotal-MESS201406001.htm





厦门大学的团队借鉴现有相关工作， 提出一种新的中文词内部层次结构定义标准， 该标准首先定义了基本语义单元， 并在此基础上定义了以这些单元为基础的词内部结构， 结构中包含了节点类型和节点内部关系。 进一步， 本文提出中文词内部层次结构的标注规范， 并人工标注了带有内部层次结构的53 918 个中文词料库。

> 论文地址如下：http://kns.cnki.net/kcms/detail/35.1070.N.20200112.0954.002.html

目前并没有找到现成的中文词典库，我已经向厦门大学团队发送了邮件，询问库是否能公开，目前还没有回复。

中文触发词字典构建调研

近年来，在信息抽取领域，事件触发词的识别方法主要有三种：基于统计的方法，基于规则的方法和机器学习方法。

基于统计的方法是指人工统计出句子或文本中的所有触发词，建立一个较完整的触发词字典，通过此字典来判断其他 词语是否为触发词。该方法简单易行，技术上要求 不高，但它是一种典型的经验性方法，且要求训练语料规模足够大且足够经典，但事实上，由于非遍历性为首统计语料的限制，此方法并不能保证统计结果和测试结果的正确性，并且统计过程费时费力。

基于规则的方法则是事先定义一些规则去寻找触发词，在一定条件下该方法能有效地提高触发词的识别效率，减少 工作量，但它是一个偏理论性的方法，只有在理想的情况下定义出涵盖所有语言特征的规则，该方法才能保证有效，而且规则的定义过程耗费大量的人力，如果规则定义得不够好，也可能过滤掉一些本身可以充当触发词的词，导致识别效果较低。 中文语境和词性千变万化，但由于规则的有限性，这种理想化的任务几乎是不可能完成的。 上述两种方法的性能在很大程度上依赖字典和规则的构建，进而依赖构建者的水平并且会耗费大量的人力和时间。

伴随着机器学习的高速发展，基于机器学习的触发词识别能够基于训练集进行自动学习。 它主要利用特征集训练触发词识别分类器，从而把触发词的识别问题转化为分类问题。 机器学习方法引进了自动化模式，大大节省了人力物力的投入，但是，机器学习需要足够量的特征集训练分类器，即要求训练语料和测试语料必须满足一定的规模才能保证识别结果的精确率，机器学习也是一种统计学习方法，不可能照顾到每个实例。

由此可见，三种识别方法各有利弊。在*中文事件触发词的自动抽取研究* 论文中作者综合三种方法，即先定义一些规则 寻找出训练语料中的事件触发词制成初始触发词表，然后用《同义词林》对初始触发词表进行扩展得到扩展的触发词表，进而利用扩展触发词表和机器学习相结合的方法对事件触发词进行自动识别抽取。

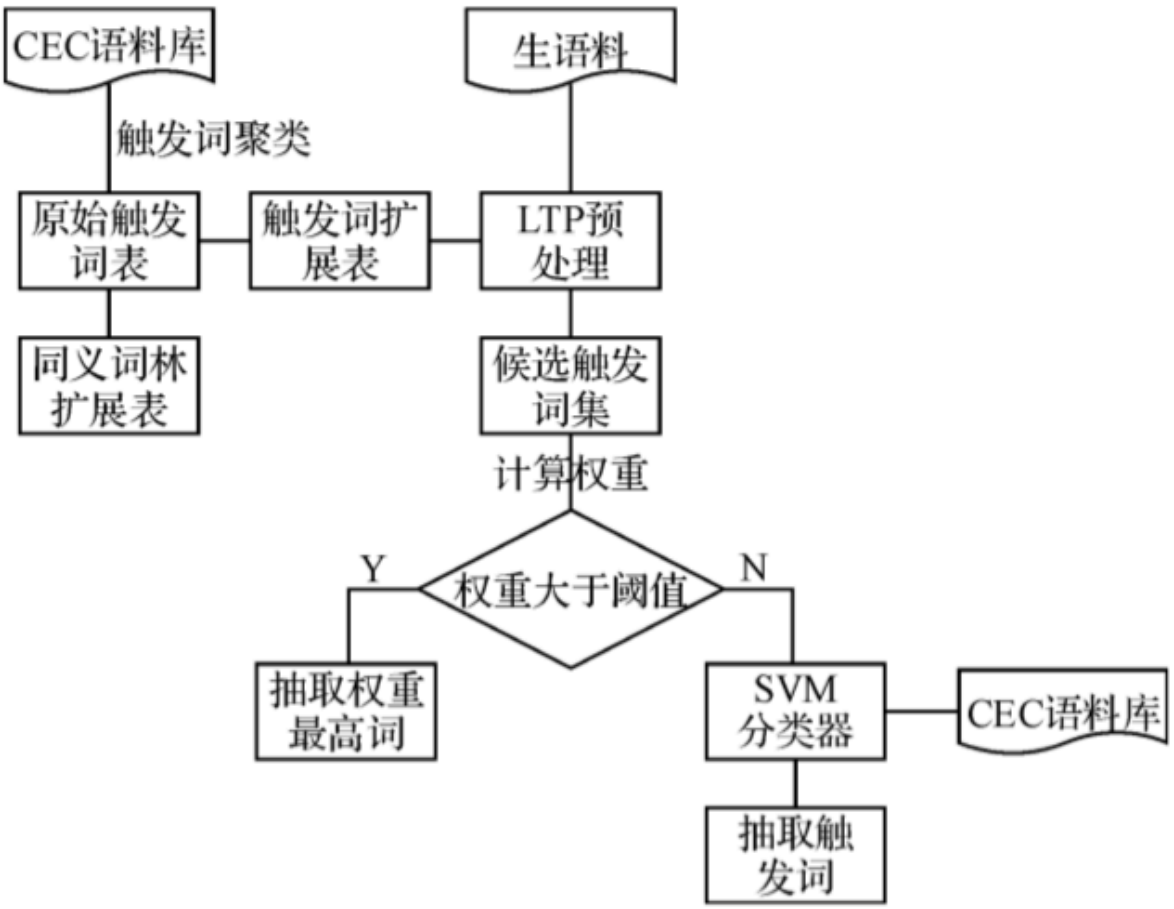
初始表如下：

事件类型	数量	事件触发词	数量
地震	130	地震、震感、余震……	7
交通事故	142	车祸、追尾、撞车……	28
恐怖袭击	102	袭击、爆炸、劫持……	25
食物中毒	118	中毒、呕吐、恶心……	9
火灾	97	火灾、着火、燃烧……	18
伤亡	478	死亡、丧生、受伤……	33
损失	395	倒塌、损坏、烧毁……	98
救援	319	救治、施救、救助……	93
移动	287	赶赴、赶到、送往……	80
合计	2068	391	

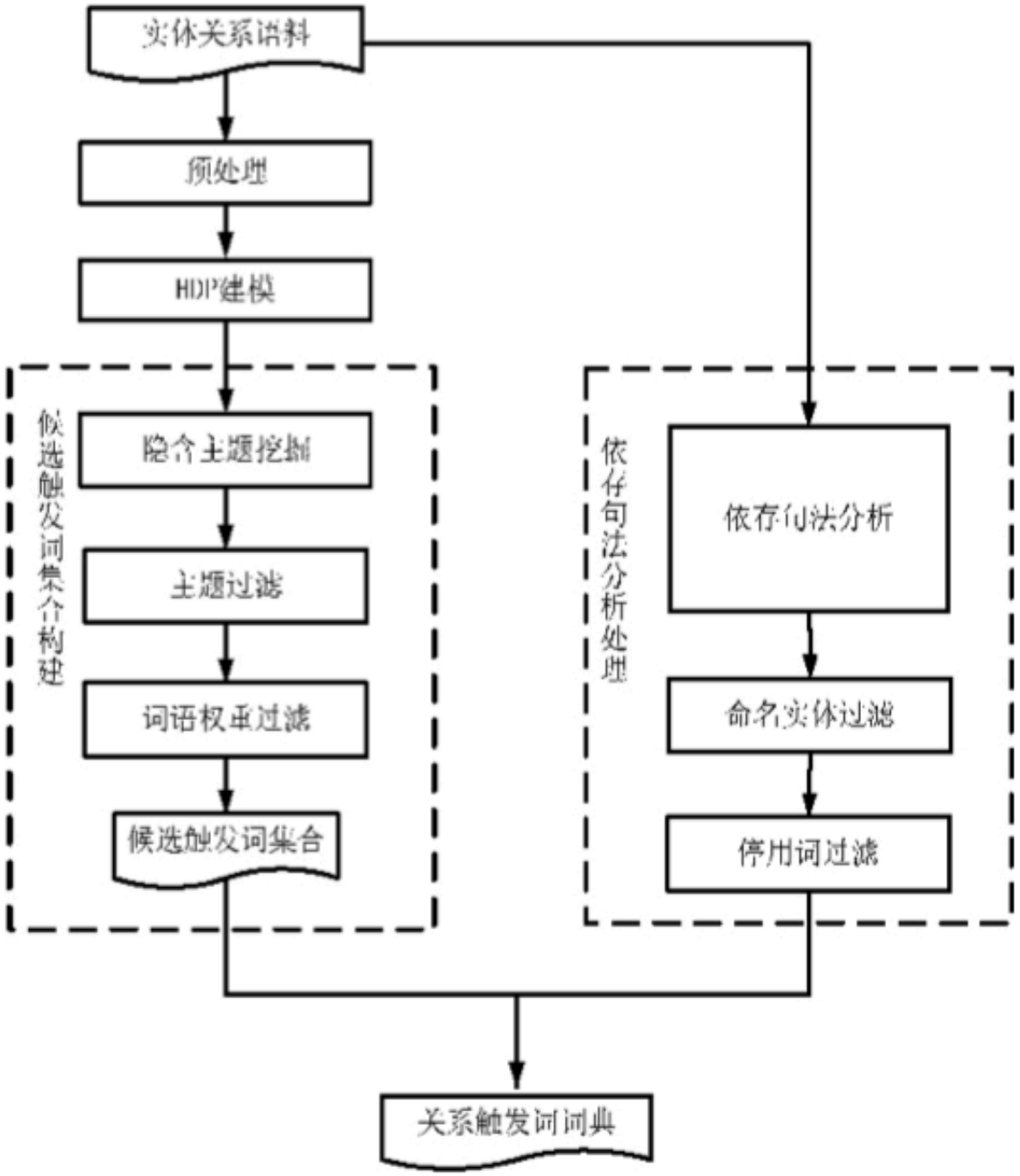
论文中阐述了扩展触发词表和机器学习相结合的触发词抽取方法基于扩展触发词表的事件触发词抽取的召回率较高但准确率却很低， 而基于机器学习的事件 触发词抽取的准确率有明显提高但召回率却低于前一种方法。 如此，可以把两种方法结合起来：

1. 先用基于扩展触发词表的方法构建候选触发词集，计算每个候选触发词的权重；
2. 为score设 定 一 个 阈 值threshold
3. 如果候选触发词中存在score大于threshold的词，则把score最高的词确定为事件触发词；
4. 若不存在，使用SVM机器学习的方法来抽取事件触发词。

如此，对于权重大于阈值的触发词一般为单义词，出现概率较低但对事件的贡献程度大，一旦出现一般就能表征某一类型事件的发生，如果这类词使用机器学习的方法来识别，由于实例比较缺乏， 所以不具典型性，容易造成识别错误，导致召回率降低，因此这类词可直接进行查表识别。 而对于小于阈值的那部分触发词，一般都有一次多意的情况，所以如果使用直接查表的方法极易造成事件的多标，导致精确率降低，因此使用机器学习的方法来解决。



基于初始触发词库的有监督扩展学习方法虽然能对实体关系触发词库进行有效扩展，但仍无法摆脱其对人工构建的初始触发词库的依赖。在另一篇*无监督实体关系触发词词典自动构建* 中提出了一种基于HDP 和依存句法分析的实体关系触发词词典自动构建方法。



以上两种方法构建的触发词词典具有一定的准确率，但仍有待提高。 一方面，这是因为获得的关系句子实例集中存在较多噪声； 另一方面，算法在主题过滤和概率权重过滤过程中涉及的阈值因子对过滤操作具有较大影响。

下周预留任务

HDP研究。