属性抽取实验（一）

作者：纽约的自行车

日期：2021 年 3 月 20 日

1 算法介绍

本文结合使用触发词（trigger）、语义角色标注（Semantic Role Labeling，srl）、依存句法分析（dependency parse, dep）从景点简介中抽取景点属性。使用哈工大 ltp 工具[1]解析文本，srl 和 dep 解析详见 ltp 官网介绍[2]。本文使用的方法可作为属性抽取 baseline 方法。

描述实体属性时，通常用到特定意义的词表示属性值和实体之间的关系，这种词称为 trigger。trigger 在句子中充当了类似桥梁的作用，将实体和属性值联系在一起。例如，句子 ‘‘白云山位于广州市白云区’’ 中的 ‘‘位于’’ 揭示了 ‘‘广州市白云区’’ 是 ‘‘白云山’’ 的**地址属性**，‘‘位于’’ 就是**地址属性**的 trigger。

使用基于 trigger 的方法抽取实体属性大致可分为三个步骤：

1. 建立属性类型和 trigger 的映射；
2. 根据 trigger 构建实体和属性值的联系；
3. 抽取属性。

第一步可人工构建词典完成属性类型和 trigger 的映射。第二步中，[1] 使用 trigger+srl 确定实体的属性块，[2] 使用 trigger+dep 确定实体的属性位置。第三步抽取属性并检查属性值是否符合格式要求。

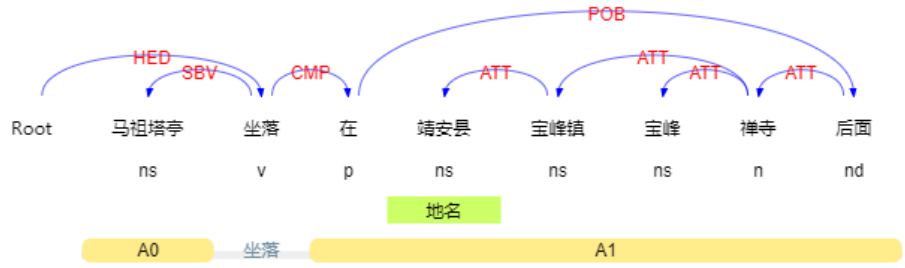

图 1: 使用哈工大 ltp 工具解析句子，包括分词、词性标注、句法分析和语义角色标注

使用 trigger+dep 方法抽取过程复杂，需要构建众多规则，例如图 1中，需要合并所有 “ATT’’ 关系的字符串，并依据 “CMP’’ 得到 ‘‘坐落’’ 和 ‘‘在’’ 的关系，由此将 trigger 替换为 ‘‘在’’，最后根据 “POB’’ 关系提取得到 ‘‘靖安县宝峰镇宝峰禅寺后面’’。使用 trigger+srl 方法抽取更简便，例如图 1中，判断得到 predicate 包含 trigger‘‘坐落’’，由此直接抽取 “A1’’ 片段字符串。但是 ltp

[1]http://ltp.ai/index.html

[2]http://www.ltp-cloud.com/intro

的 srl 标注体系采用动词性 predicate，即 predicate 的词性以动词为主，并且不是所有的动词都是 predicate，所以 srl 的标注能力有限，难以覆盖所有 trigger。因此本文同时使用 srl 和 dep 构建实体和属性值的关系。

2 属性抽取实践

本次实验选取**地址**、**别名**、**占地面积**、**建造时间**四种属性。其中**地址**和**别名**属于通用属性，**占地面积**和**建造时间**属于部分类型景点才具备的属性。ltp 对长句解析效果差，文本只能以逗号和句号分割，以缩短句子长度。

2.1 地址抽取

根据现有语料总结得到部分**地址**属性的 trigger，然后利用《同义词词林》[3]添加部分 trigger。最终使用的 trigger 集合如下：

```
坐落、位于、在、分布、座落、地处、介于、位居
```

通常景点名和景点地址同时出现在一句话中，所以可根据景点名来进一步过滤地址。算法的输入是一段景点介绍文本和景点名称，输出是抽取的属性值。抽取过程：

1. 以句号和逗号分割文本，解析每个句子，包括分词、srl、dep、pos；
2. 保留含有 ‘‘地址属性”trigger 的句子的解析结果；
3. 在 srl 结果中，找到含有触发词的 argument，判断景点名称是否在 “A0’’ 片段中，若存在，提取 “A1’’ 片段的字符作为地址属性值；

在哈工大 ltp 在线演示平台[4]上对句子 ‘‘江西省展览中心位于南昌市八一广场西侧’’ 进行解析，结果如图 2。

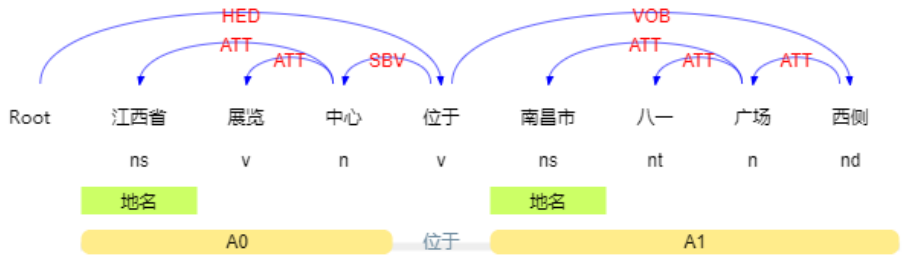

图 2: 地址抽取。返回 trigger‘‘位于’’ 的 “A1’’ 片段。

trigger“‘位于’’ 句子中，景点名 ‘‘江西省展览中心’’ 包含在 “A0’’ 片段中，因此提取 “A1’’ 片段字符串 ‘‘南昌市八一广场西侧’’ 作为地址属性值。

2.2 别名提取

根据现有语料总结得到部分**别名**属性的 trigger，然后利用《同义词词林》添加部分 trigger。最终使用的 trigger 集合如下：

[3]https://github.com/taozhijiang/chinese_correct_wsd/blob/master/HIT-IRLab-同义词词林（扩展版）_full_2005.3.3.txt

[4]http://ltp.ai/demo.html

称、称为、名、俗称、称作、称之为、叫做、叫作、叫、号称、俗名、誉为、喻为、称做

在 srl 的 predicate-argument 体系中，只有动词才能成为 predicate，而部分别名属性的 trigger 是非动词，例如 ‘‘俗名’’。针对非动词的 trigger，本文结合正则表达式提取属性值。

算法输入是一段景点介绍文本，输出是属性值。抽取过程：

1. 以句号和逗号分割文本，解析每个句子，包括分词、srl、dep、pos；
2. 保留含有 ‘‘地址属性”trigger 的句子的解析结果；
3. 使用 trigger + srl 确定 argument，保留 predicate 含有 trigger 的 argument；
4. 使用 trigger + sdp 抽取属性，抽取与 trigger 的依存关系为 “VOB’’ 的字符串，若该字符串等于 ‘‘为’’，则抽取 srl 中 predicate 等于 “‘为’’ 的 “A1’’ 片段，如图 3；
5. 使用正则表达式抽取属性。

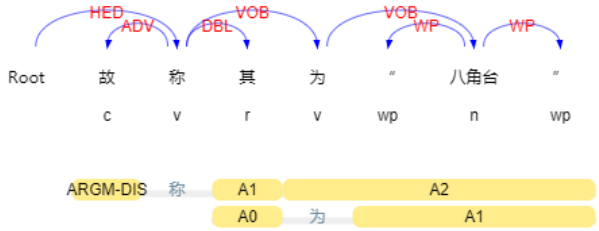

图 3: 别名抽取。与 trigger‘‘称’’ 存在 “VOB’’ 关系的字符串是 ‘‘为’’，抽取 ‘‘为’’ 的 “A1’’ 片段。

2.3 占地面积抽取

根据现有语料总结得到部分**占地面积**属性的 trigger 的如下：

面积、占地、总面积

在描述该属性时通常会加上量词，例如 32 平方米。本文使用 trigger 和量词同时作为属性值位置的判断条件，量词使用 ltp 的词性标注功能识别。通过对语料的依存句法分析，得到两种描述该属性值的方式。第一种是 trigger 和量词存在 “SBV’’ 关系，如图 4所示。第二种是 trigger 和量词之间通过中间词关联，trigger 和中间词存在 “SBV’’ 关系，中间词和量词存在 “VOB’’ 关系，如图 5所示。确定量词位置后，提取与量词存在直接或者间接 “ATT’’ 或 “RAD’’ 关系的字符，拼接后作为属性值返回。例如图 4中的 ‘‘约 4 公顷’’，图 5中的 “4 公顷’’。

算法输入是一段景点介绍文本，输出是属性值。抽取过程：

1. 以句号和逗号分割文本，解析每个句子，包括分词、srl、dep、pos；
2. 保留含有 ‘‘占地面积’’ 属性 trigger 的句子的解析结果；
3. 获取量词在句子中的位置；
4. 判断描述方法是否符合第一种或者第二种情况，若符合则根据规则抽取属性值，然后返回，否则继续判断下一个句子。

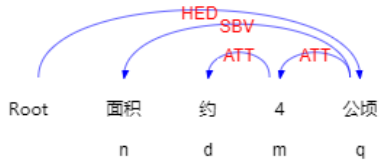

图 4: 占地面积抽取第一种情况。trigger‘‘面积’’ 和量词 ‘‘公顷’’ 存在直接的 “SBV’’ 关系。

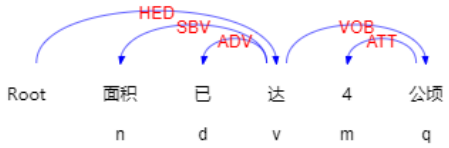

图 5: 占地面积抽取第二种情况。trigger‘‘面积’’ 和量词 ‘‘公顷’’ 通过中间词 ‘‘达’’ 连接，‘‘面积’’ 和 ‘‘达’’ 存在 “SBV’’ 关系，‘‘公顷’’ 和 ‘‘达’’ 存在 “VOB’’ 关系。

2.4 建造时间抽取

根据现有语料总结得到部分**别名**属性的 trigger，然后利用《同义词词林》添加部分 trigger。最终使用的 trigger 集合如下：

始建、成立、产生、生成、建、始、筑、修建、创建、建成、落成、辟、创立、开创、创始、创办、创造、缔造、交工、完工、竣工、创设、创、创导、开立、修成、建筑、建造、修筑、兴修、构筑、修、盖

时间描述方式复杂多样，trigger 和时间的依存关系难以归纳覆盖全部方式。为了应对多种时间表达方式，本文使用三种方法抽取时间属性。第一种方法使用 trigger+ 依存关系；第二种方法使用 trigger+ 规则；第三种方法使用正则表达式。第一种和第二种方法需要对抽取值过滤，只有抽取值中包含词性为 “nt’’ 的词才能被返回。

trigger+ 依存关系方法：通过分析语料的句法分析结果，总结出两种时间和 trigger 的关系。第一种如图 6所示，通过介词将 trigger 和时间关联。在这种情况下，若 trigger 与介词存在 “CMP’’ 关系，则找到与介词存在 “POB’’ 关系的词，然后返回与该词存在直接或间接 “ATT’’ 关系的词作为时间。第二种如图 7所示，trigger 和时间直接关联。在这种情况下，找到与 trigger 存在 “ADV’’ 关系并且词性为 “nt’’ 的词，然后返回与该词存在直接或间接 “ATT’’ 关系的词作为时间。

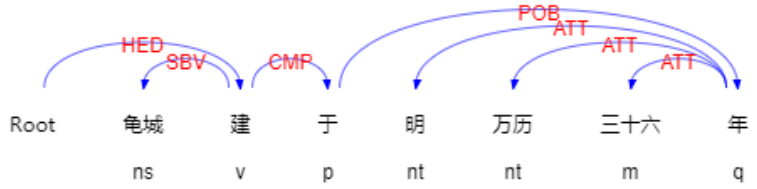

图 6: trigger‘‘建’’ 与 ‘‘年’’ 通过介词 ‘‘于’’ 关联，‘‘于’’ 与 ‘‘明’’、‘‘万历’’、‘‘三十六’’ 存在直接或间接的 “ATT’’ 关系，返回 ‘‘明万历三十六年’’ 作为时间。

trigger+ 规则方法：根据语料整理出两种规则方法。第一种在介词和 trigger 之间寻找词性为 “nt’’ 的词，如图 8所示。第二种提取 ‘‘辟为’’ 前的词性为 “nt’’ 的词，如句子 *白塔山 1958 年辟为公园*，返回 “1958 年’’ 作为时间。

正则表达式方法：最后使用正则表达式，正则表达式如下所示：

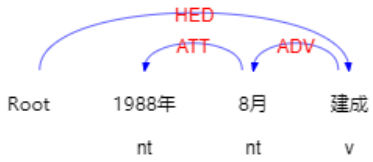

图 7: trigger‘‘建成’’ 与 “8 月’’ 直接关联，“1988 年’’ 与 “8 月’’ 存在直接的 “ATT’’ 关系，返回 “1998 年 8 月’’ 作为时间。

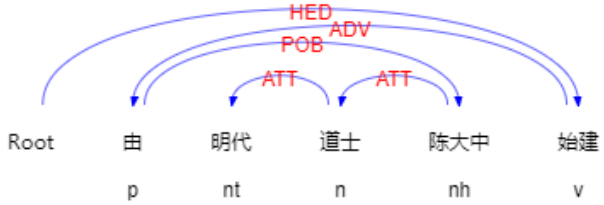

图 8: 返回 trigger‘‘始建’’ 与介词 ‘‘由’’ 之间词性为 “nt’’ 的词作为时间

```
re.compile(r'(建 于 | 年 代 为).*?[, 。 ,\n\s]')
```

算法输入是一段景点介绍文本，输出是属性值。抽取过程：

1. 以句号和逗号分割文本，解析每个句子，包括分词、srl、dep、pos；
2. 保留含有 ‘‘建造时间’’ 属性 trigger 的句子的解析结果；
3. 使用 trigger+ 依存关系方法抽取，判断抽取值，符合要求即可返回抽取值，否则执行下一步；
4. 使用 trigger+ 规则方法抽取，判断抽取值，符合要求即可返回抽取值，否则执行下一步；
5. 使用正则表达式方法抽取，抽取成功则返回值，否则继续判断下一个句。

3 实验

3.1 数据集介绍

从多个旅游网站中共随机选取 152 个景点介绍。整理不规范的描述，例如没有主语的句子。删除无关描述，例如交通路线、门票信息等。人工标注**地址**、**别名**、**占地面积**、**建造时间**四种属性。数据统计情况如表 1所示。

属性类型	地址	别名	占地面积	建造时间
数量	131	19	68	51

表 1: 属性值数量统计表

3.2 实验结果

使用 F1 值作为评价指标。分别计算每类属性的 F1 值，测试结果如表 2所示。代码已上传服务器[5]。

属性类型	P	R	F1
地址	86.75	99.24	92.58
别名	11.92	94.74	21.18
占地面积	43.62	95.59	59.91
建造时间	32.21	94.12	48.00
macro-F1	51.13	95.92	55.42

表 2: 测试结果统计表

地址属性抽取效果最好，这是因为景点简介文本结构单一，大多数仅用 ‘‘位于’’ 即可确定属性值位置。其他属性抽取效果一般，尤其是**别名**，召回率高，但是精确率很低。

3.3 错误样例分析

分析错误样例，总结出主要的四种错误提取原因：

1. 句子中不存在 trigger；
   在描述属性值时，没有 trigger。如下样例所示，‘‘新石器时代’’ 在句子中没有 trigger 相关联。
   例 *1*：新石器时代文化遗址。牛门洞文化遗址位于会宁县头寨子乡牛门洞村，面积约 *16* 平方公里。*1920* 年首先发现彩陶，*1975* 年挖出古代墓穴，出土大量文物。
2. trigger 覆盖不全；
   trigger 由人工从语料中收集，难以覆盖全部情况，如下样例中地址 trigger‘‘处于’’ 未被收录。
   例 *2*：澳门一号广场处于一号湖畔
3. trigger 误触发；
   trigger 误触发现象在**别名**提取中较多，如下样例。‘‘称’’ 是 trigger，导致错误提取 ‘‘齐云庵县北五十里皖公峰下’’ 作为别名。
   例 *3*：但清乾隆 *46* 年 *(1781* 年*)*《潜山县志》则称：“齐云庵县北五十里皖公峰下，全真道人字肱子修烁于此，休邑金商建”。
4. 文本中含有多个同类属性值，错误提取其中与景点无关的属性值。
   当文本中含有多个同类型属性值时，无法确定哪个才是与景点相关。这种现象多出现在**占地面积**和**建造时间**属性中。如下样例中，两次出现 trigger‘‘面积’’，只有在理解句子语义的基础上才能确定 ‘‘约 1164 平方米’’ 是正确属性值。
   例 *4*：原旧构面积约 *700* 平方米；曾在道光十九年（公元 *1839* 年）、民国三年（公元 *1914* 年）、公元 *2004-2005* 年三次修葺。现有规模已经形成了一个面阔 *27.30* 米、通进深 *42.63*

[5]/home/team3user/knowledge graph/attribute extract/

米、占地面积约 *1164* 平方米的四合院式建筑体系。

对于不存在 trigger 的情况，本文无法解决。trigger 覆盖率不全的情况只能人工继续收集语料中的 trigger。trigger 数量多容易误触发导致精确率低，trigger 数量少导致召回率低。一种解决的方法是保持高召回率的同时，增加对属性值过滤操作，过滤明显不具备该类属性值特征的字符串。至于第四种情况，当文中存在多个同类属性值时，只有在理解句子语义的基础上才能判断哪个属性值才是需要提取的，目前算法难以做到语义理解。

4 总结

本文使用的属性抽取方法可作为属性抽取 baseline 方法。这种半监督方法无需人工大量标注数据，但仍需要人工整理各类属性的语言描述特点。对于描述句式单一的属性，本文的方法效果很好。对于描述方法复杂多样的属性效果则很差。

参考文献

[1] PETER EXNER P N. Using Semantic Role Labeling to Extract Events from Wikipedia[J]. 2011.

[2] DIAN YU H J. Unsupervised Person Slot Filling based on Graph Mining[J]. ACL, 2013.