# 属性抽取实验 (二)

作者：纽约的自行车

日期：2021 年 3 月 20 日

## 1 前言

使用文献 [1] 中提出的方法提取属性值，在景点数据上测试，测试结果表明该方法可以较好的完成属性抽取任务。

## 2 训练数据生成

论文中提出的模型属于监督学习方法，因此需要带标签数据，我们借助百度百科自动生成景点的标签数据。生成方法如图 1所示，infobox 中的信息属于结构化数据，属性类型和属性值都可以自动获取，摘要文本属于非结构化数据，将 infobox 中的属性值在摘要文本中标注得到带标签的训练数据。


**图 1:** 数据生成方法

## 3 算法简介

将属性抽取转化为序列标注任务，以商品说明为例，需要提取三种属性 {time,size,flavor}，使用 {B,I,O} 标记方式的标签集为 {B-time,I-time,B-size,I-size,B-flavor,I-flavor,O}。论文中为了提取训练集中未出现的属性类型，省略标签集中的属性类型，只留下 {B,I,O}。在景点属性抽取任务中不存在抽取新属性问题，因此本文在标签集中保留属性类型。

模型结构如图 2所示。Title 表示景点描述文本，Attribute 表示待提取的属性类型名称。

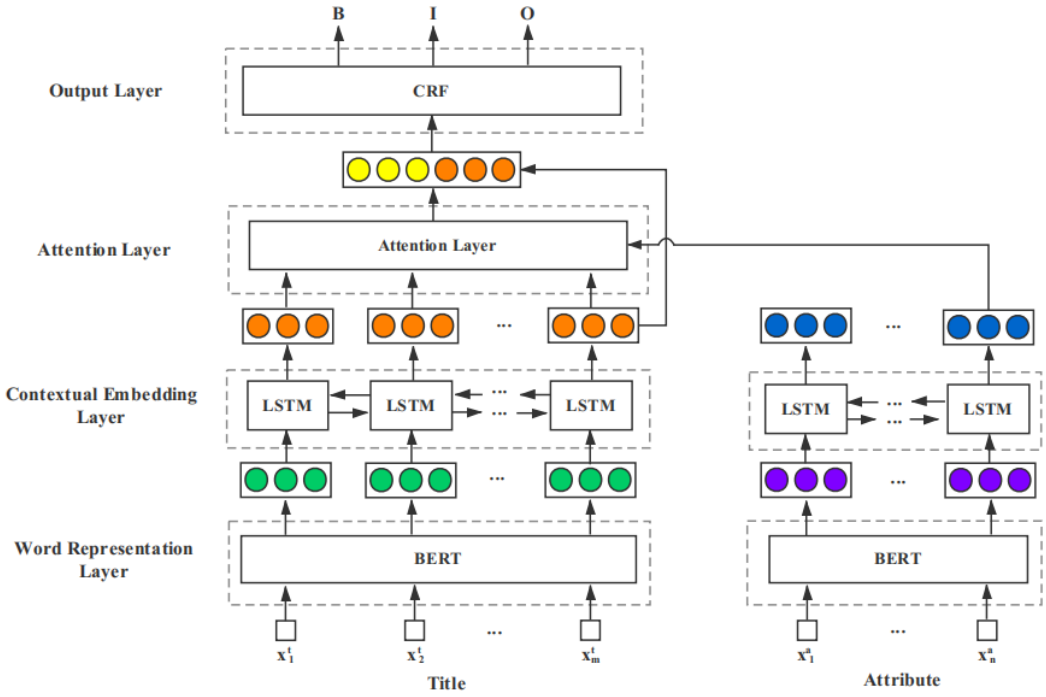


**图 2:** Architecture of the proposed attribute-comprehension open tagging model.

图 2左边对 title 建模，右边对属性建模。两者都用 Bert 预训练词向量，然后输入双向 LSTM。对 title 建模过程中，title 每个位置输出的向量都保留，而在属性建模中，只保留 LSTM 最后时刻的输出向量。模型的重点在注意力层。左侧 LSTM 得到 title 的语义表示，右侧 LSTM 得到属性的语义表示。通过注意力捕获属性和 title 的语义关系，生成蕴含属性理解的 title 表征，最后输入到 CRF 中。注意力计算方法如下：

$$\alpha_i = cosine(h_i^t, h^a)$$

$h_i^t$ 表示每个时刻左侧 LSTM 输出的向量，$h^a$ 表示右侧 LSTM 最后时刻输出的向量。模型的输入数据集格式是 title, attribute, value，attribute 表示一个属性名称，value 是 title 中对应的属性值。

# 4 实验

## 4.1 数据集介绍

本次实验选用的属性类型有：别名、相关人物、建筑结构、发现时间、建造时间、竣工时间、馆藏资源。各属性数量统计如表 1所示。数据集分为训练集和测试集，比例为 7:3。我们使用默认参数训练模型，没有调参过程，因此无需验证集。代码已上传服务器 [1]。

[1]/home/team3user/knowledge graph/attribute extract/BERT-Attribute-Value-Extract/

| 属性类型 | 别名 | 面积 | 建造时间 | 竣工时间 | 相关人物 | 发现时间 | 馆藏资源 | 建筑结构 |
|---|---|---|---|---|---|---|---|---|
| 样本数量 | 2378 | 2375 | 1950 | 462 | 299 | 83 | 66 | 62 |

**表 1:** 数据集中各属性数量统计情况。

### 4.2 测试结果

测试结果如表 2所示。baseline V1.0 表示《属性抽取实验报告》V1.0 版本中使用的方法，主要是结合触发词、语义角色标注、依存句法分析从景点简介中抽取景点属性。

| 属性类型 | baseline V1.0 | | | 本文方法 | | |
|---|---|---|---|---|---|---|
| | P | R | F1 | P | R | F1 |
| 别名 | 11.92 | 94.74 | 21.18 | 74.85 | 70.4 | **72.55** |
| 面积 | 43.62 | 95.59 | 59.91 | 80.27 | 81.58 | **80.92** |
| 建造时间 | 32.21 | 94.12 | 48 | 77.72 | 72.34 | **74.93** |
| 竣工时间 | - | - | - | 82.8 | 56.62 | **67.25** |
| 相关人物 | - | - | - | 58.7 | 62.79 | **60.67** |
| 发现时间 | - | - | - | 76.19 | 69.57 | **72.73** |
| 馆藏资源 | - | - | - | 100 | 5.88 | **11.11** |
| 建筑结构 | - | - | - | 63.64 | 41.18 | **50** |

**表 2:** 测试结果对比。

测试结果表明本文采用的方法显著优于 baseline V1.0 方法。结合表 1和表 2分析，训练样本越多效果越好。此外，属性值结构简单效果也越好。在三个数量很少的属性类型中，**发现时间**的效果明显优于其他两种属性类型，**建筑结构**效果次之，这可能与属性值结构简单有关。**发现时间**的属性值都属于 “xxx 年 xx 月 xx 日’’ 形式，而且大部分只有年份。**建筑结构**的属性值中大都包含 ‘‘结构’’ 二字，与属性类型名相似度高。**馆藏资源**的属性值都为文物名称，且存在多个值，同时数据中还存在大量噪声，例如标注不全和标注错误等样例，这些都导致该属性的召回率极低。与时间有关的三个属性类型中，**建造时间**的样本数量远多于**发现时间**，但两者效果相似，这可能是因为**建造时间**的属性值中含有大量朝代年号等信息，**竣工时间**同样存在此类情况。

综上所述，影响模型效果的因素主要有两大类，样本数量和属性值形式复杂度。样本数量越多效果越好，属性值形式越简单效果越好。

## 参考文献

[1] Huimin Xu, Wenting Wang, Win Mao, Xinyu Jiang, Man Lan. Scaling up open tagging from tens to thousands: Comprehension empowered attribute value extraction from product title[J]. ACL, 2019.