华院数据中文地址魔方大赛

结合地址树推理和正则匹配的中文地址识别

创意说明文档

团队名称：甲基苯丙胺

一、 命题分析

1) 命题复杂性

本次大赛需要对中文地址进行标准化的处理，识别中文命名实体中的 10 级要素。综合考虑中文处理的复杂性、中文地址命名的不规范性、大量地址记录在统计上呈现的规律性，认为本次大赛的命题难度适中，算法实现难度适中，具有重要的现实意义。

2) 问题和困难点

a) 低级数的命名实体被省略。

例：“温岭市太平街道万寿路 206”，省略了省和市，利用先验知识补全之后是“浙江省

台州市温岭市太平街道万寿路 206”。

面对复杂的中文地址记录，算法应该能够对省略的命名实体进行推理。

b) 地址写法的不规范性。

例：“常州武进区人民路 25-10 号”，规范化之后是“江苏省常州市武进区人民路 25-10 号”。

对于本次大赛给定的任务，算法应该能够规范化前几级要素的名字。

c) 街镇乡和路的二义性

“杭州西湖区高技街 49 号”中的“高技街”被识别为路，而“江苏省徐州市铜山区彭政街”中的“彭政街”被识别为街镇乡。

对于这个实体识别的二义性，需要考虑国家行政区划的实际情况。

d) 路号、楼号、单元号、户号的识别

从路号级别开始，几乎无法获得标准的地名数据先验。类似楼号、单元号、户号，随着命名实体级数的增加，可能的地址命名呈指数级增长，也不可能将所有地址作为数据库以备查询。因此，需要对每个要素构建中文实体识别算法，以解决实际问题。

e) 备注的识别。

虽然示例中的备注规律性不强，但可以发现，“(*)”的内容作为备注出现，未识别到的文字也以备注形式出现。

如何写备注，以尽可能的保留有用信息，成为一个难点。

二、 解题思路

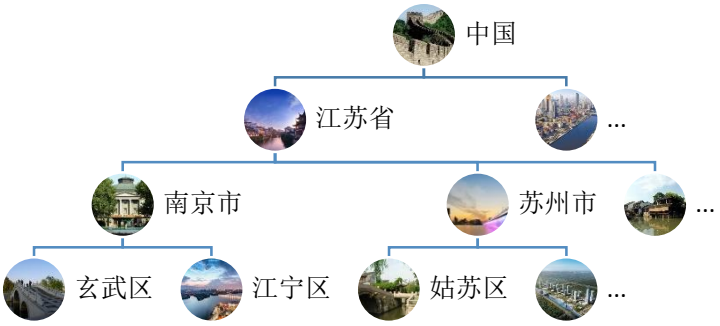
1) 数据清洗

对于少量地址数据，存在“其他”“其他地区”等干扰词汇。在算法执行的第一步，需要去除此类干扰词汇。

2) 充分利用国家**行政区划**作为**先验**知识

通过国家统计局网站公开的数据，能够获得县及以上的名称和行政区划代码。5 级行政区划数据通过互联网获得。

得到数据之后，将其建立为 1 颗中文地址树，结构如下图所示。



3) 采用**级别递增正向最大匹配**算法识别已知地名。

首先对所有已知的地名建立前缀 hash 索引。

对于给定的一条地址记录，例如“南京玄武区玄武湖湖中心”，若直接使用正向最大匹配算法，匹配结果为(南京->南京市，玄武区->玄武区，玄武->玄武区)，显然玄武区出现了 2 次，通过把玄武湖这个实体错误的切割开了。

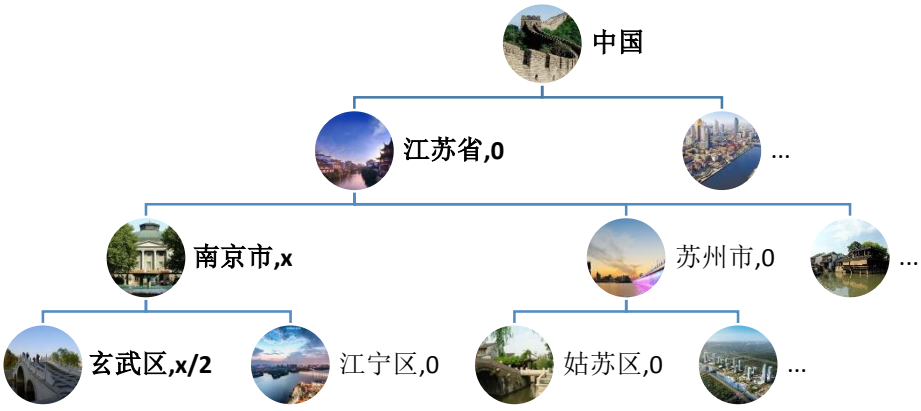
这个问题可以通过使实体的级数严格递增解决。首先识别出上(南京市,2)，之后识别出(玄武区,3)。随后识别的(玄武区,3)级别不增，说明是误识别，忽略。因此，最终识别结果为(南京->南京市，玄武区->玄武区)。

具体实现可以是 Hashmap 或者 Trie 树，实现方式对识别精度没有影响，只存在性能方面的差异。

4) 利用中文**地址树**进行**推理**

延续 3)的例子。算法目前只识别出市和区。若要得到省级信息，就必须通过中文地址树进行推理。

具体计算过程如下：对第 i 个识别出的要素，在地址树上加上分数 $X/2^i$。之后枚举地址树上有分数的节点，计算其到树根的分数和，取分数最大节点到树根的路径作为识别结果。计算过程如下图



因此得到最终结果为：江苏省->南京市->玄武区。

通过此方法，可以有效避免诸如“山东省北京市(辖区)”等识别错误。其识别结果是鲁棒的。根据现有数据，此方案最多能识别 5 级地名。

5) 基于规则的路号、楼号、单元号、户号识别

这部分识别任务没有标准数据支持。故使用基于规则的分割、识别策略。例如路号，一般为“[数字]+号”等情况的组合。通过在数据中寻找各种可能的模式，实现此类要素的识别，是一种切实可行的方案。

另一种做法是采用条件随机场或隐马儿可夫模型做机器学习，由训练集得到优化模型，再将其用于命名实体识别的任务。考虑到训练集容量小，中文汉字个数多，对算法而言相同的命名实体存在很大的语义鸿沟，实际效果不够理想。

最终的算法系统拟采用基于规则的方法实现。

6) 备注的识别

通过分析数据可知，备注常常是一些()内的内容，以及 9 级地址均未能识别的字符串。因此，通过分析这两种情况，可以实现备注的识别。

三、 外部数据引用

国家统计局三级行政区划数据：
http://www.stats.gov.cn/tjsj/tjbz/xzqhdm/201504/t20150415_712722.html

五级行政区划数据：
http://pan.baidu.com/s/1hqxD6vU

四、 参考文献

1. 尹存燕, 黄书剑, 戴新宇,等. 中英命名实体识别及对齐中的中文分词优化[J]. 电子学报, 2015(08):1481-1487.
2. 俞鸿魁, 张华平, 刘群,等. 基于层叠隐马尔可夫模型的中文命名实体识别[J]. 通信学报, 2006, 27(02):87-94.
3. 向晓雯, 史晓东, 曾华琳. 一个统计与规则相结合的中文命名实体识别系统[J]. 计算机应用, 2005, 25(10):2404-2406.
4. 向晓雯. 基于条件随机场的中文命名实体识别[D]. 厦门大学, 2006.
5. 赵琳瑛. 基于隐马尔科夫模型的中文命名实体识别研究[D]. 西安电子科技大学, 2008.
6. 科曼. 算法导论[M]. 机械工业出版社, 2006.