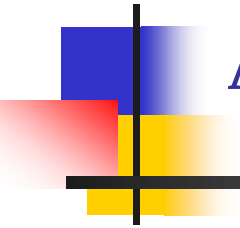


Automatic Recognition of Chinese Location Entity

Reporter: Xuewei Li

12/23/2014

Introduction

- Corpus
  - Previous studies mainly focus on texts which the format is standard;
  - Our research is on the texts of complaint about urban management
- Object
  - Previous studies mainly focus on place names which characteristics are clear, easy to recognize, such as "Beijing" and "city of Zhengzhou in Henan province". ;
  - Our object – location is complicated and longer.

So Identifying the location is difficult by using traditional methods.

The texts of complaint about urban management and location entities examples

Text 1	标题：关于马家堡西路角门西地铁站外面的丁字路口的问题 来信内容：1.马家堡西路角门西地铁站外面的丁字路口，南北向人行横道上北面横着一道30米左右的护栏。.....望有关部门早日解决以上问题
Text 2	标题：整治环境 来信内容：海淀区西四环路定慧北桥以东的定慧福里小区南面的停车场简直就是个垃圾站，.....。定慧福里北面及家乐福前面的道路上，.....。如：从定慧寺车站到定慧桥车站路的北面人行道上经常有狗屎。.....北京是首都，因此也会影响到中国在世界上的形象。
Text 3	标题：海淀区黑泉路一个井盖缺失 来信内容：.....地点在黑泉路南段，北向南方向非机动车道，北五环林萃桥北200米左右。

- As can be seen from above table, location is made up of basic place and indicator. Similarly, the divide-and-conquer strategy can decompose complex problem into smaller parts and solve them. So, we borrow the idea of divide-and-conquer strategy to divide location recognition into basic place recognition and indicator lexicon construction.

Location Recognition

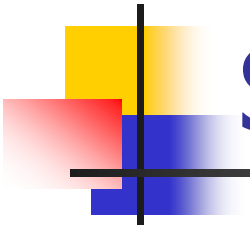
- The Model of Location Recognition
- BasePla Recognition Based on CRF Role Labeling
- Extraction and Expansion of Indicators
- LocEntity Recognition

Definition

- **Definition 1.** Basic place: it is generalized location where event occurs and its length is usually short, denoted BasePla
- **Definition 2.** Indicator: It often appears after the BasePla. Meanwhile, its appearance can make the location where event occurs more exactly, but it appears alone is meaningless, denoted IndicateLoc.
- IndicateLoc=AreaSet ∩ DirectionSet ∩ SpotSet
- **Definition 3.** Location Entity: It is a specific location where event occurs, denoted LocEntity.
- **Definition 4.** BI distance: It is the number of NormalWords that appear continuously between BasePla and IndicateLoc and near the IndicateLoc, denoted BI-Len.

LocEntity = BasePlaSet + NormalWordSet + IndicateLocSet

- 马家堡西路角门西 地铁站 外面 东北角 的 丁字路口 的 问题
  - LocEntity=[马家堡西路角门西地铁站外面东北角的丁字路口]
  - ✓ BasePlaSet={马家堡西路角门西}
  - ✓ IndicateLocSet={地铁站, 外面, 东北角, 丁字路口}
    - SpotSet={地铁站, 丁字路口}
    - AreaSet={外面}
    - DirectionSet ={东北角}
  - ✓ NormalWordSet={的}
- BI_1-Len=0, BI_2-Len=0, BI_3-Len=0, and BI_4-Len=1.



Steps

1. Identify BasePla;
2. Build indicator lexicon semi-automatically;
3. Identify LocEntity using attachment connection algorithm.

BasePla Recognition Based on CRF Role Labeling

- Since BasePla recognition can be converted into sequence annotation and the boundary identification. Similarly, CRF is a kind of conditional probability model for annotation and segmentation ordinal data, which combines the characteristics of the maximum entropy model with hidden markov model, joins the long-distance contextual information, and solves the problem of label bias. So we use CRF model to identify BasePla.
- The basic idea : firstly, we process the corpus with Chinese word segmentation and part-of-speech tagging; secondly, some words are labeled with some roles using restrained role labeling; finally, we select word, part-of-speech and role as features to identify BasePla based on CRF.

BasePla Roles

Role	Description	Example
W	Tail word	红莲北里,大望路
QI	Area indicator	朝阳十里堡地区、戎晖家园周边
FI	Direction indicator	朝阳路北侧、京洲北街南侧人行道
DI	Spot indicator	西红门宜家工地、定福庄路土路
SL	Words before the BasePla	位于方庄东路、家住房山区长阳镇
XR	Words after the BasePla	西大望路交口处、长安街邻近区域
C	Conjunction	帝京路和宝隆路
S	BasePla	海淀/ns五路居
N	Words not related to roles	

Restrained Role Labeling

- POS constraint
  - “马连道/nr 中里”,“马/nr 家堡西路”
  - ”北京信息科技大学/nz 向东200米”
- Tail word and context constraint

$$P = \frac{TF(tail)}{TF(all)}$$

- Conjunctions constraint

tail word > indicator > conjunction > context word

Feature Template of CRF

- we select word, POS and role as features, use B, I, E, O as label, and apply atom feature template and compound feature template to identify BasePla using CRF

atom feature template	W(i), W(i+1), W(i+2), W(i+3), W(i-1), W(i-2), W(i-3), P(i), P(i+1), P(i-1), R(i), R(i+1), R(i+2), R(i-1), R(i-2)
compound feature template	W(i-1)+P(i-1), P(i-1)+P(i), P(i)+P(i+1), R(i-2)+R(i-1)+W(i), W(i)+R(i+1)+R(i+2)

Extraction and Expansion of Indicators

- **Extraction of Indicators**
- Area indicator usually contains “区”(area), “内”(inside) and “外”(outside);direction indicator usually contains “东”(east), “西”(west), “南”(south), “北”(north), “上”(up), and “下”(down); spot indicator is usually noun or noun phrase.
- **Expansion of Indicators**
- Apply HIT-CIR Tongyici Cilin (Extended) to find synonymous and similar words and expand indicators.

LocEntity Recognition

- **Definition 5.** Attachment relationship: For wordA and wordB, wordA is a meaningful word and it can appear alone, however, wordB is meaningless if it appears alone and it must attach itself to wordA. That is to say, wordB is a supplement to wordA and it makes wordA more specific. In brief, wordB depends on wordA, denoted wordB →wordA.
- "朝阳路北侧" where "北侧" is a supplement to "朝阳路" and "北侧" is meaningless when it appears alone, that is, "北侧" attaches itself to "朝阳路" to make the place more specific, denoted that 北侧→朝阳路.

Attachment Connection Algorithm

- Indicator→BasePla

Algorithm 1. Attachment Connection Algorithm

1: Input: every sentence Sen = $W_1W_2W_3…W_n$ in the test corpus TestC, BasePla = $W_i…W_j$ ($1 \le i \le j \le n$), IndicateWSet = {$IndicateW_1$, $IndicateW_2$, … , $IndicateW_n$}
2: BI-Len ← 0, the position of connection pointer ← 0, LocEntity ← BasePla
3: for m ← j+1 to n do
4: for m < n or BI-Len ≤ 4 do
5: if $W_m \in$ IndicateWSet then
6: BI-Len ← 0
7: pointer ← m
8: else BI-Len++
9: endif
10: m++
11: LocEntity ← LocEntity+$W_{j+1}…W_{pointer}$
12: Output: LocEntity

Experiment--Evaluation Criterion

$$P = \frac{NR}{NG} ? 100\%$$

$$R = \frac{NR}{NC} ? 100\%$$

$$F = \frac{2 创 P \quad R}{P + R} ? 100\%$$

Experiment-- Results

Experiment		R	P	F
CRF	BasePla Recognition	85.35%	75.80%	80.29%
	LocEntity Recognition	85.04%	75.52%	80.00%
LocEntity Recognition using our method		88.77%	81.15%	84.79%

Identified LocEntity

1	定慧福里小区南门的停车场
2	朝阳区劲松二区229号楼都城心屿小区西侧停车场
3	北京邮电大学南门对面胡同
4	海淀区车道沟桥，牛顿办公区和嘉豪国际中心的停车场
5	马家堡西路角门西地铁站外面的丁字路口，南北向人行横道上北面
6	西城区百万庄南街3号楼最东面

Thanks!