Automatic Recognition of Chinese Location Entity

Xuewei Li[1], Xueqiang Lv[1], and Kehui Liu[2,3]

[1] Beijing Key Laboratory of Internet Culture and Digital Dissemination Research, Beijing Information Science and Technology University, Beijing, China
li_xuewei163@163.com lxq@bistu.edu.cn
[2] Beijing Institute of Technology, Beijing, China
[3] Beijing Research Center of Urban Systems Engineering, Beijing, China
lkh_2005@126.com

Abstract. Recognition of Chinese location entity is an important part of event extraction. In this paper we propose a novel method to identify Chinese location entity based on the divide-and-conquer strategy. Firstly, we use CRF role labeling to identify the basic place name. Secondly, by using semi-automatic way, we build indicator lexicon. Finally, we propose attachment connection algorithm to connect the basic place name with indicator, then we achieve the identification of location entity. In brief, our method decomposes location entity into basic place name and indicator, which is different from traditional methods. Results of the experiments show that the proposed method has an outstanding effect and the F-value gets to 84.79%.

Keywords: Chinese location entity. Divide-and-conquer strategy. CRF role labeling. Basic place name. Indicator lexicon. Attachment Connection Algorithm.

1 Introduction

Urban management enters into the age of information and people raise issues about urban management through the Internet. Through non-standard writing, texts of complaint format are quite different. Thus, the staff must read verbatim to find important event from exponential growth of texts of complaint about urban management that is shown in table 1, which is inefficient.

By adopting the technique of information extraction, we can extract event automatically by converting the unstructured data into structured data. It not only improves the work efficiency, but also helps the urban management department to master the implementation of policy and find the existing problems in the social management. Besides, automatic recognition of location entity, which is shown as italic and underline in table 1, is an important part of event extraction.

C. Zong et al. (Eds.): NLPCC 2014, CCIS 496, pp. 379–391, 2014.

Table 1. The texts of complaint about urban management and location entities examples

Experimental Corpus and location entities example (italic and underline indicates location entities)	
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 table 1, location entity is made up of basic place name and indicator. Similarly, the divide-and-conquer strategy can decompose complex problem into smaller parts and solve them. In this paper, we borrow the idea of divide-and-conquer strategy to divide location entity recognition into basic place name recognition and indicator lexicon construction. Then attachment connection algorithm was put forward, which is ACA, to connect basic place name with indicator.

The contributions of this paper are twofold. Firstly, the research on location entity recognition in the text of complaint about urban management has never appeared before. Secondly, we propose a model that integrates divide and conquer strategy in location entity recognition, which is quite different from other ones. To the best of our knowledge, this paper is the first one which has introduced indicator in location entity. As shown in the experiment section, our method gains reasonable performance.

The rest of the paper is organized as follows. Some related works are discussed in section 2. Then we will introduce some basic concepts as basic place name, indicator, as well as location entity and present our model in section 3. After that, experiment results and discussions are showed in section 4. Finally, section 5 concludes the whole paper and put forward some work to be done in the future.

2 Related Work

Currently, the domestic related studies on recognition of Chinese place name mainly focus on texts which the format is standard. Cai et al. [1] proposed rule reasoning based method to identify unexpected event place name entity from news corpus and extract place name entities including province, city, county, township and village by analyzing

expressive features of unexpected event place name entity. Li et al. [2] defined it as a binary classification problem and applied a SVM classifier to identify Chinese place names with key words, such as province and city, from People's Daily. Tang et al. [3] employed CRFs-based module for simple location name recognition from People's Daily corpus. Du et al. [4] recognized Chinese place names in news page corpora. Other scholars [5,6,7,8,9] also studied the recognition of simple Chinese place names in standard corpus. Gao et al. [10] analyzed characteristics of the longest location entity and identified it by using the maximum entropy model.

Above researches on Chinese place names recognition are in the standard news corpus, the characteristics of place names in the news is clear, easy to recognize, such as "Beijing" and "city of Zhengzhou in Henan province". Research on the text of complaint about urban management is markedly different from the research above. In addition to diverse format, location entity is complicated and longer. Identifying the location entity is difficult by using traditional methods. In this paper, we divide location entity recognition into basic place name recognition and indicator lexicon construction. Then attachment connection algorithm was put forward, which is ACA, to connect basic place name with indicator and identify location entity.

3 Location Entity Recognition

As discussed before, we use the model of location entity recognition to identify location entity. The definition of the model is as follows.

3.1 The Model of Location Entity Recognition

For the convenience of description, we define following concepts:

Definition 1. Basic place name: it is generalized location where event occurs and its length is usually short, denoted BasePla, the set is BasePlaSet.

Definition 2. Indicator: It often appears after the BasePla. Meanwhile, its appearance can make the location where event occurs more exactly, but that it appears alone is meaningless, denoted IndicateLoc. The set denoted that IndicateLocSet. Suppose IndicateLocSet=$\{a_1,a_2,\ldots,a_m,d_1,d_2,\ldots,d_n,s_1,s_2,\ldots,s_k\}$ then IndicateLoc$\in$ IndicateLocSet. AreaSet=$\{a_1,a_2,\ldots,a_m\}$, DirectionSet=$\{d_1,d_2,\ldots,d_n\}$,SpotSet=$\{s_1,s_2,\ldots,s_k\}$, AreaSet $\cap$ DirectionSet = $\varnothing$, DirectionSet $\cap$ SpotSet = $\varnothing$, SpotSet $\cap$ AreaSet = $\varnothing$.

Where AreaSet is the set of area indicators containing some words that indicate specific range where event occurs; DirectionSet is the set of direction indicators containing some words that indicate specific direction where event occurs; SpotSet is the set of place indicators containing some words that indicate specific spot where event occurs.

Definition 3. Location Entity: It is a specific location where event occurs, denoted LocEntity. Which is defined as below.

LocEntity = BasePlaSet + NormalWordSet + IndicateLocSet

Where NormalWordSet is the set of NormalWord except for BasePla and indicators. Namely, NormalWord $\notin$ BasePlaSet $\cup$ IndicateLocSet. As defined before, $|$NormalWordSet$| \geq 0$ and $|$IndicateLocSet$| > 0$ ($|$A$|$ is the number of elements in set A). To sum up, LocEntity mainly includes the following three characteristics:

1. It is a noun or a noun phrase.
2. It is the longest description of a location.
3. It is associated with a specific event.

Definition 4. BI distance: It is the number of NormalWords that appear continuously between BasePla and IndicateLoc and near the IndicateLoc, denoted BI-Len.

For example, for the sentence:"马家堡西路角门西 地铁站 外面 东北角 的 丁字路口 的 问题"(The problem of the northeast T-junction outside the Majiabu Jiaomen west subway), where LocEntity=[马家堡西路角门西地铁站外面东北角的丁字路口](the northeast T-junction outside the Majiabu Jiaomen west subway), BasePlaSet={马家堡西路角门西(Majiabu Jiaomen West)}, IndicateLocSet={地铁站(Subway), 外面(Outside), 东北角(Northeast corner), 丁字路口(T-junction)}, where SpotSet={地铁站(Subway), 丁字路口(T-junction)}, AreaSet={外面}(Outside), DirectionSet ={东 北 角 }(Northeast corner), NormalWordSet={ 的 }, BI_1-Len=0, BI_2-Len=0, BI_3-Len=0, and BI_4-Len=1.

In the next subsections, we will apply the divide-and-conquer strategy to build our model. The following three steps must be taken:

1. Identifying BasePla based on CRF role labeling;
4. Building indicator lexicon semi-automatically;
5. Identifying LocEntity using attachment connection algorithm.

3.2 BasePla Recognition Based on CRF Role Labeling

Since the 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 of BasePla recognition based on CRF role labeling is as follows: 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 as features to identify BasePla based on CRF.

The Definition of BasePla Roles. Roles table is the basis of the BasePla recognition. The computer can understand the message hidden in the Chinese characteristics according to roles table. Roles table mainly contains following roles:

1. Internal information of BasePla: A part of BasePla, namely tail word of BasePla, marked by W. For example, “红莲北里”(Red lotus North), “海淀区莲花小区”(Haidian District, Lotus area) and “大望路”(Da Wang road) .

2 External information of BasePla: Some words except for BasePla, including context , indicator and conjunctions.

   (a) Context : Some words before and after the BasePla, tagged by SL and XR respectively, such as “至车碾店胡同”(To Cheniandian Lane) and “北京市平谷区供暖”.(Beijing Pinggu District Heating)

   (b) Indicator : It usually appears behind the BasePla. On the other hand, it is divided into three kinds: area indicator, direction indicator and spot indicator, respectively, with QI, FI, DI annotations. such as “马家堡西路角门西地铁站外面东北角的丁字路口”(the northeast T-junction outside the Majiabu Jiaomen west subway)

   (c) Conjunction: It refers to some words connecting two parallel place names, such as “和(and)、与(and), 及(and), 或(or), 或者(or), ‘、’, ……”, marked with C, for example, “帝京路和宝隆路”(Teikyo road and Baolong Road) .

3 Part-of-speech(POS) information of BasePla: The word that POS is ns marked with S. such as “海淀/ns五路居” (Haidian/ns Fifth Avenue Home).

4 Words not related to role: It is not related to the role, marked by N.

To conclude, in this paper we define 9 roles, as shown in table 2.

Restrained Role Labeling. Definition 5. Restrained role labeling: A word is conditional marked with a specific role, rather than any conditions.

Definition 6. Bag of words: the set of unrepeated words that extracted from the texts before and after a word as the specific window, denoted that Bag-w.

Table 2. Roles table

Role	Description	Example
W	Tail word	红莲*北里*(Honglian North),大望*路*(Dawang road)
QI	Area indicator	朝阳十里堡*地区*(Shilipu Chaoyang district),戎晖家园*周边*(surrounding of Rong Hui Home)
FI	Direction indicator	朝阳路*北侧*(the north side of Chaoyang Road),京洲北街*南侧*人行道(On the south side of pavement of Jingzhou North Street)
DI	Spot indicator	西红门宜家*工地*(site of IKEA in the West Red Door),定福庄路*土路*(dirt road of Dingfuzhuang road)
SL	Words before the BasePla	*位于*方庄东路(Located Fangzhuang East),*家住*房山区长阳镇(one who lives in the town of Changyang in the Fang Shan area)
XR	Words after the BasePla	西大望路*交口*处(the intersection of West Dawang Road),长安街*邻近*区域(near Chang'an Avenue)
C	Conjunction	帝京路*和*宝隆路(Teikyo road and Baolong Road)
S	BasePla	*海淀/ns*五路居(Haidian/ns Fifth Avenue Home)
N	Words not related to roles	

By observing the corpus, we find some POS like nr and nz are incorrect and they should be labelled nz, therefore we first constraint the POS role, the rules are as following:

1. POS constraint
   (a) Words with nr POS: nr is the POS of a person name, but person name seldom appears in the text of complaint about urban management. The word that POS is nr is mostly place or component of place. For example, "马连道/nr 中里"(Ma Liandao/nr Zhong), "马/nr 家堡西路" (Ma/nr Jiabao West). So we should convert nr into ns.
   (b) Words with nz POS: nz is the POS of institution names, but sometimes institution names can be indicated BasePla, for instance, the sentence "北京信息科技大学/nz 有多少学生"(How many students are Beijing information science and technology university), where" 北京信息科技大学"(Beijing information

science and technology university)is institution name, but for the sentence"北京信息科技大学/nz 向东200米"(Beijing information science and technology university eastbound 200 meters), where "北京信息科技大学"(Beijing information science and technology university) is BasePla, so we need to convert nz POS into ns by using rule below: For any word which POS is nz, we obtain the Bag-w by setting the window to 4. If Bag-w contains IndicateLoc (the construction of indicators is shown 3.3), we convert nz into ns; or we do nothing.

As described before, role is related to the BasePla and appears near the BasePla, that is to say, role appears in the Bag-w of BasePla. Moreover, the words in the Bag-w of BasePla hold the 65% proportion in all words. If they are all marked with corresponding roles, the feature will not be clear between role words and normal words and the result is bad. We need to choose useful words according to following constrains.

2 Tail word and context constraint

We apply the formula 1 to statistic the probability of a word which is the tail word and get tail words table.

$$P = \frac{TF(tail)}{TF(all)} \tag{1}$$

Where TF(tail) is the count of a word which is the tail word, TF(all) is the total count of a word. The statistics of probability of a word that is the word before or after BasePla is the same as formula 1.

3 Conjunctions constraint

Since not all conjunctions connect two BasePla, we follow the regulation below to get useful conjunction: for all the conjunction, we set the window to 4 and get the set of POS, denoted PosSet. If PosSet contains the POS of ns, the conjunction is useful, and marked with C, or it is marked with N.

Since a word usually plays more than 2 roles, which will result in the difficulty in identifying BasePla. So we define that a word only plays the most important role which it can and sort the different roles according to the importance: tail word > indicator > conjunction > context word.

Feature Template of CRF. As mentioned before, CRF can connect the long-distance contextual information and combine various information, related or unrelated, therefore, we select word, POS and role as features, use B, I, E, O as label, where B is the begining of BasePla, I is the middle of BasePla, that is, between the begin and end, E is the tail of BasePla, O is the word that is not related to BasePla, and apply atom feature template and compound feature template, as shown in table 3, to identify BasePla using CRF.

Table 3. Feature template of 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)

Where W is word, P is POS and R is role. W(i) is current word, W(i+1) is the first word after the current word and W(i-1) is the first word before the current word. R and P is the same to W.

3.3 Extraction and Expansion of Indicators

Indicator describes specific location where event occurs, we category functionally three types: area indicator, direction indicator and spot indicator. In this paper we extract indicators using semi-automatic method and expand indicator using Synonymy Thesaurus.

Extraction of Indicators. Indicator usually appears after BasePla and is limited to 3 types. In this paper, we extract five words after every BasePla to build Bag-w of indicators. However, not all words in Bag-w of indicators are indicators. After that, we would select some words containing some special characters as indicator. As mentioned before, different from NormalWord, indicator usually contains some special characters, for instance, 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. In this paper, we obtain right indicators using features above and proofreading manually.

Expansion of Indicators. Since the number of indicators extracted from limited corpus is small, we need to expand indicators using external resources. In this paper we apply HIT-CIR Tongyici Cilin (Extended) to find synonymous and similar words and expand indicators.

Synonymy Thesaurus was written by Mei[11] in 1983, then HIT-CIR finished HIT-CIR Tongyici Cilin (Extended) based on which by introducing other external resources. The format of HIT IR-Lab Tongyici Cilin (Extended) is shown in table 4.

Table 4. The format of HIT IR-Lab Tongyici Cilin (Extended)

Codes	Similar words
Cb01D01@	时空(time)
Cb02A02=	东(East) 东边(East) 东方(East) 东面(East) 东头(East) 正东(East) 左(Left)
Bc02C17#	右上方(the upper right hand corner) 右上角(the upper right hand corner) 右下方(the lower right hand corner) 左上方(the upper left hand corner) 左上角(the upper left hand corner) 左下方(the lower left hand corner) 左下角(the lower left hand corner)

The method of expansion of indicators is as follows: For every character or word in the indicators, if HIT IR-Lab Tongyici Cilin (Extended) contains, we add its synonymous or similar words to indicators; or we do nothing.

As mentioned before, LocEntity consists of BasePla and indicator. Meanwhile, we have obtained BasePla and indicator as described before. In the next subsections, we will identify LocEntity by connecting BasePla with indicator.

3.4 LocEntity Recognition

Definition 7. 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. For example, "朝阳路北侧"(Chaoyang road north), where "北侧(north)" is a supplement to "朝阳路(Chaoyang road)" and "北侧(north)" is meaningless when it appears alone, that is, "北侧(north)" attaches itself to "朝阳路(Chaoyang road)" to make the place more specific, denoted that 北侧(north)→朝阳路(Chaoyang road).

Indicator is a supplement to BasePla in the corpus of complaint about urban management. On the other hand, indicator attached itself to BasePla, denoted indicator→BasePla. So we propose Attachment Connection Algorithm, namely ACA, to connect BasePla with indicator.

ACA is as follows:

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

Where the reason why BI-Len is less than or equal to 4 is the number of continuous NormalWords between BasePla and indicator is limited to 4 by analyzing LocEntity.

Long LocEntity may contain punctuation and it will be divided into two short LocEntity by punctuation when identifying LocEntity, for example "崇文区忠实里东区（东环居苑）小区门口路两侧"(the sides of the door of Chongwen District in the area of Dongshili east(East loop Home)) will be divided into "崇文区忠实里东区"(Chongwen District in the area of Dongshili east) and "东环居苑"(East loop Home).

In order to recall long LocEntity described above, we combine short LocEntity in accordance with the following rules:

1. There only exists a punctuation between two short LocEntity.
2 Short LocEntity co-exists in a sentence.

4 Experimental Results and Analysis

4.1 Preparation for Experiments

The experimental corpus is 1500 texts of complaint which are crawled from government web site. At the same time, two annotations are made: first label BasePla, second label LocEntity, and final two annotations are proofread by professional staff.

We select randomly 1000 texts of complaint as training data and the rest 500 texts as test data. Then, using NLPIR[12], that is ICTCLAS2013, to process Chinese corpus with Chinese word segmentation and part-of-speech tagging with PKU second standard. After that, we use role labeling to label the corpus. After repeated experiments, we set the value of tail constraint and context constraint to 0.4 and 0.5 respectively.

4.2 Evaluation Criterion

In this paper we use precision rate, recall rate and F-value to evaluate experimental result. (For brief, we use the P, R and F instead hereafter respectively)

$$P = \frac{NR}{NG} \times 100\% \tag{2}$$

$$R = \frac{NR}{NC} \times 100\% \tag{3}$$

$$F = \frac{2 \times P \times R}{P + R} \times 100\% \tag{4}$$

Where NR is the number of right LocEntity identified by our method, NG is the number of LocEntity identified by our method, NC is the number of LocEntity contained in the corpus.

4.3 Results and Analysis

CRF is a fusion of the characteristics of maximum entropy model and hidden markov model[13]. It combines contextual information with the advantages of maximum entropy model. So we choose CRF-based LocEntity recognition as contrast test, whose features and template are the same as CRF-based BasePla recognition.

The experimental results of BasePla and LocEntity are shown in table 5.

Table 5. Experimental results of BasePla and LocEntity

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%

As seen in the above table, the result indicates that our method is far better than CRF. Moreover, it gains up to 84.79% in F-value and 4.79 % higher than CRF-based. CRF-based LocEntity in F-value is 0.29% lower than CRF-based BasePla. The reason are as follows:

1. The feature of BasePla is clearer than LocEntity.
2. It is easy to identify by CRF for the entities which features appears in training data. For example “宣武区鸭子桥南里小区” (area of Duck Bridge South in Xuanwu District), “朝阳区都城心屿小区西侧停车场” (parking lot in the west of Chaoyang District Ducheng Xinyu area), “马家堡西路角门西地铁站外面的丁字路口”(T-junction west Majiabu Jiaomen outside the subway station). It is difficult to identify by CRF for the entities which features don’t appear in training data, for instance, “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”(T-junction outside the Majiabu Jiaomen West subway, in the north of North-south pedestrian crossing) and “西城区百万庄南街3号楼最东面”(On the extreme east 3rd floor of South Street in Xicheng District Baiwanzhuang).

3 By introducing ACA, our model can identify LocEntity, like “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” (T-junction outside the subway station in the Majiabu Jiaomen West, in the north of North-south pedestrian crossing) and “西城区百万庄南街3号3楼最东面” (On the extreme east 3rd floor of South Street in Xicheng District Baiwanzhuang). This improves recall rate.

Table 6 shows some LocEntity extracted by our method.

Table 6. Identified LocEntity

Number	Identified LocEntity
1	定慧福里小区南门的停车场(Parking lot of South Gate in Dinghuifuli Distinct)
2	朝阳区劲松二区229号楼都城心屿小区西侧停车场(parking lot in the west of Chaoyang District Ducheng Xinyu area)
3	北京邮电大学南门对面胡同(the alley across the south gate of Beijing University of Posts and Telecommunications)
4	海淀区车道沟桥，牛顿办公区和嘉豪国际中心的停车场(Parking lot of Chedaogou Bridge, Newton office and Jiahao International Center)
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(T-junction outside the subway station in the Majiabu Jiaomen West, in the north of North-south pedestrian crossing)
6	西城区百万庄南街3号楼最东面(On the extreme east 3rd floor of South Street in Xicheng District Baiwanzhuang)

As can be seen from table 6, we can find that

1. Our approach can identify LocEntity with punctuation, such as “海淀区车道沟桥，牛顿办公区和嘉豪国际中心的停车场” (Parking lot of Chedaogou Bridge, Newton office and Jiahao International Center) and “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” (T-junction outside Majiabu Jiaomen West subway, in the north of North-south pedestrian crossing).
3 Our approach can identify such LocEntity that CRF can’t, such as “西城区百万庄南街3号楼最东面” (On the extreme east 3rd floor of South Street in Xicheng District Baiwanzhuang) and “海淀区车道沟桥，牛顿办公区和嘉豪国际中心的停车场” (Parking lot of Chedaogou Bridge, Newton office and Jiahao International Center) .

In conclusion, our model solves the difficulty of identifying LocEntity to certain degree and injects new ideas for LocEntity recognition.

However, our model also extracts incorrect LocEntity such as “丰台区蒲黄榆二里这个小区”(this area of Huangyuerli in Fengtai Distinct) and ”朝阳区广顺北大街大西洋新城南门应该修建过街天桥”(The Atlantic Metro South Gate should be built overpass in Beijing chaoyang district wide BeiDaJie). That is because the people who complaint the problem of urban management using the word in the different ways and sometimes they use the modifiers before indicator of LocEntity, for example, the pronouns “这个”(this) and verb “修建”(build).

5 Conclusions and Future Work

In this paper, we propose a novel method to identify LocEntity by introducing indicators. Unlike traditional methods, our method follows the divide-and-conquer strategy: we divide LocEntity recognition into BasePla recognition and indicator lexicon construction. First, we use the CRF to identify BasePla, and then we build indicator lexicon semi-automatically. Finally, we propose the ACA to connect BasePla with indicator and obtain LocEntity.

Experiments on the corpus of complaint about urban management indicate that our method can not only ensure higher accuracy but also improve the recall rate. Moreover, the proposed method injects new ideas for LocEntity recognition. However, since our method depends on the precision of BasePla based on CRF and the comprehensive construction of indicators, we plan to expand training data and indicators so that the effect of recognition will be better in the future. Furthermore, the method is corpus independent and can be extended to other corpus once we have the training data in the target corpus.

Acknowledgement. This work is supported by National Natural Science Foundation of China under Grants No. 61271304, Beijing Natural Science Foundation of Class B Key Project under Grants No. KZ201311232037, and Innovative engineering of scientific research institutes in Beijing under Grants No. PXM2013_178215_000004.

References

Cai, H.L., Liu, L., Li, H.: Rule Reasoning-based Occurring Place Recognition for Unexpected Event. Journal of the China Society For Scientific and Technical Information 30(2), 219–224 (2011)

Li, L.S., Huang, D.G., Chen, C.R., et al.: Research on method of automatic recognition of Chinese place names based on support vector machines. Minimicro Systems-Shenyang 26(8), 1416 (2005)

Tang, X.R., Chen, X.H., Xu, C., et al.: Discourse-Based Chinese Location Name Recognition. Journal of Chinese Information Processing 24(2), 24–32 (2010)

Du, P., Liu, Y.: Recognition of Chinese place names based on ontology. Xibei Shifan Daxue Xuebao/ Journal of Northwest Normal University (Natural Science) 47(6), 87–93 (2011)

Li, N., Zhang, Q.: Chinese place name identification with Chinese characters features. Computer Engineering and Applications 45(28), 230–232 (2009)

Li, L.S., Dang, Y.Z., Liao, W.P., et al.: Recognition of Chinese location names based on CRF and rules. Journal of Dalian University of Technology 52(2), 285–289 (2012)

Li, L.S., Huang, D.G., Chen, C.R., et al.: Identifying Chinese place names based on Support Vector Machines and rules. Journal of Chinese Information Processing 20(5), 51–57 (2006)

Huang, D.G., Yue, G.L., Yang, Y.: Identification of Chinese place names based on statistics. Journal of Chinese Information Processing 17(2), 46–52 (2003)

Qian, J., Zhang, Y.J., Zhang, T.: Research on Chinese Person Name and Location Name Recognition Based on Maximum Entropy Model. Journal of Chinese Computer Systems 27(9), 1761–1765 (2006)

Gao, Y., Zhang, W., Zhang, Y., et al.: Application of Maximum Entropy Model in the LLE Identification. Journal of Guangdong University of Petrochemical Technology 4, 014 (2012)

Mei, J.J., Zhu, Y.M., Gao, Y.Q., et al.: Synonymy Thesaurus. Shanghai Lexicographical Publishing House, Shanghai (1993)

NLPIR Chinese word segmentation system, `http://ictclas.nlpir.org/downloads`

Xu, B.: Oral standardization processing based on Conditional Random Fields model. Nanjing University of Science and Technology (2009)