International Journal of Knowledge and Language Processing
Volume 3, Number 4, **2012**

www.ijklp.org
KLP International ©2012 ISSN 2191-2734
pp. 58-**67**

Research of a SE-Aided Location Name Recognition Post-processing Module

Xu Chao[1] and Tang Xuri[2]
[1]School of Chinese Language and Literature
Nanjing Normal University
No.122, Ninghai Road, Nanjing, China
xuchao.nnu@gmail.com

[2] School of Foreign Languages
Huazhong University of Science and Technology
No.1037, Luoyu Road, Wuhan, China
xrtang@126.com

Received September 2012; revised October 2012

ABSTRACT. *The paper presents the design of a search engine aided location name recognition post-processing module. Based on the location name candidates obtained from a prototype location name recognition system, the search engine is used to help filter these candidates to get higher performance. The Methodology is based on the relevance among location names; first a set of location names is obtained from the text in which the location name candidates lie, and each of these candidates is submitted to search engine, then we observe the appearance of the location names set's elements in result pages. Our research shows that without using other linguistic resources, the module achieves 1% improvement in both precision and recall compared with the prototype system.*
Keywords: Location name recognition; Search engine; Relevance measurement

1. **Introduction.** Location name recognition is one of the vital tasks of unregistered words recognition and named entities recognition. Almost every aspects of Chinese information processing, such as word segmentation, Part-of-Speech tagging, machine translation, information retrieval, are highly connected to location name recognition. Also, location name recognition plays a significant role in the geographic information system construction, lexicography and encyclopedic knowledge acquisition, etc. [1]. So it is always one of the most popular topics in the field of Chinese information processing.

1.1. **The State-of-the–art.** Main stream of the location name recognition's studies tend to be based on statistical model and assisted by rules. Almost all of the natural language statistical models have been used in location name recognition. Considering the shortage of the training corpus (the size and coverage of the corpus with location name taggers), data sparseness, restraint on observation window, etc., identified location names generally need

later optimization, done by regular methods or addition of expert knowledge as artificial rule base, context information database, etc.

. Table 1 shows the main technical methods and experimental results of recent studies on location name recognition systems.

TABLE 1. Recent Studies on Location Name Recognition Systems.

	precision	recall	F measure	main technical methods
Jing Qian 2006[2]	91.30%	93.13%	92.19%	Based on maximum entropy model, assisted by dynamic word list and rules
Lishuang Li 2006[3]	89.57%	93.52%	91.50%	Based on support vector machine, assisted by regular template
Hongkui Yu 2006[4]	82.83%	89.73%	86.14%	Based on a cascade hidden Markov model
Degen Huang 2006[5]	92.23%	83.88%	87.86%	Rule based
Nuo Li 2009[6]	87%	80%	83%	Based on maximum entropy model, assisted by rule template
Haipeng Liu 2009[7]	97. 58%	95.73%	96.65%	Based on Conditional Random Fields, assisted by knowledge base
Xiaofei Qian 2009[8]	88.16%	87.32%	87.74%	Combined ways of common location name matching, fragment analysis and combination expansion
Xuri Tang 2010[9]	87.83%	92.28%	89.76%	Based on Conditional Random Fields, assisted by a discourse-based module for the relationship identification
Jiupeng Ju 2011[10]	92.86%	90.91%	91.87%	Based on Conditional Random Fields, assisted by expert knowledge

Note: data in table 1 has no horizontal comparison value due to different training corpus, testing corpus and evaluating methods.

According to what mentioned above, research methodology that based on statistical model and assisted by rule methods have come to the maturity stage. All the statistical models, doing well in other natural language processing field, have had enough experiments in location name recognition and made optimization as far as possible. Yet is optimizing model and improving knowledge base the only way of improving the performance of location name recognition? How much it can be improved indeed? How to break through in this field?

1.2. **How to Breakthrough—Some Samples of Wrong Recognition.**First of all, let us watch some wrong recognition samples, selected from recognition results of the system in [9].

•Type I errors /False negatives:

Tengtou village (滕 头 村) of Fenghua in Jiangsu province achieved the title of the 500 elite in the global ecological environmental protection.

Chunhua (淳 化) is the national famous county of apple base.

What can make Dong Yulin (东 榆 林) more famous is……

•Type II errors/False positives:

Pu Fulu (铺富路)for poor areas

Every Da zhong city (大中城市) should be responsible for overall balance

Jia Jinqiao (架金桥) for the cross-strait enterprises.

Focusing on Qinggou (清沟) and preventing water logging

Da xing (大兴) ethos of investigation and research.

At first glance, there are two reasons caused these wrong recognitions: one is the crossover of location names and common nouns, and the other is some characters are rarely used as part of location names. However, the inner technological causes are shortage of training corpus and data sparseness. Usually, the solution to this kind of error is using rules, although few rules can be really used in each recognition process due to limited amount and poor coverage.

1.3. **Assumptions of Breakthrough.** The wrongly recognized location names mentioned above can be easily identified by manual work. This is because have the language sense of understanding the whole sentence and human beings have world knowledge—a rule base of tremendous scale.

The simple thought is to make the location name recognition system also having a world knowledge base. It is unrealistic to unlimitedly expand the scale of corpus and coverage of the common knowledge base, but search engine is capable of retrieving information from the Internet without limitation. Therefore, our simple thought is that dynamically retrieving related knowledge using search engine.

2. Design Objective and Main Thought.

2.1. **Design Objective.** Based on the location name candidates obtained from a prototype location name recognition system, the search engine is used to help filtering these candidates to improve the performance of recognition.

2.2. **Theoretical Foundation.**

Assumption 1: If there is a location name "Loc_i" in the text "S", it indicates Loc_i has a high relevance with other location names in "S". Written as:

If $Loc_i \in S$ then

Loc_i<-> $\{Loc_{1S} \ldots Loc_{nS}\}$

In it, $Loc_i=c_1,c_2 \ldots c_n$ and $c_1\ c_2 \ldots c_n$ is a string.

For example, Nanjing(南 京) as a location name is highly relevant to other location names in the same text such as Jiangsu(江苏), Gulou(鼓楼), Xin Jiekou(新街口).

Assumption 2: When "$c_1,c_2 \ldots c_n$" occurs in another text "D", if it shows high relevance

with other location names in “D”, it indicates “$c_1,c_2…c_n$” is a location name in “D”. Written as:

If $c_1,c_2…c_n$ <->{$Loc_{1D}…Loc_{nD}$} then $c_1,c_2…c_n=Loc_i∈D$

For example, if there exist such location names as Nanjing(南京), Jiangsu(江苏), Xikang Road(西康路), Hankou Road (汉口路) in a text with a string of “Gulou”, it indicates this “Gulou” is also a location name.

Deduction: If the location name set of “S” is somehow similar to that of text “D”, the string “$c_1,c_2…c_n$” made up of the location name “Loc_i” in “S” may also be a location name in “D”. Written as:

If $Loc_i∈S$ and {$Loc_{1S}…Loc_{nS}$}≈{$Loc_{1D}…Loc_{nD}$} then $c_1,c_2…c_n=Loc_i∈D$

For example, there are two texts “S” and “D”. Gulou(鼓楼) as a location name occurs in “S” and there are other location names such as Nanjing(南京), Jiangsu(江苏), Xikang Road(西康路), Hankou Road(汉口路) in T. If the string “Gulou” also occurs in “D”, while “Jiangsu”, “Xikang Road” and “Hankou Road” also exist, then the string “Gulou” may also be a location name in the text “D”.

In fact, these assumptions and deduction are using world knowledge to identify location names. Within an independent text, it’s probably hard to have enough evidence for identifying whether a string is a location name or not, but if the string is observed in a wider context, it will be easily identified by more evidence. This is the role of the text “D”: a reference text to provide a wider context.

2.3. **Obtaining Reference Texts.** As the corpus size can’t be infinite, it is not guaranteed that location names for recognition always appear in the corpus, if only local corpus is taken as reference. So we considered using search engine that means almost infinite corpora resource on the Internet can be used as a reference text. Namely, location names waiting for recognition are first submitted to search engine, then we take returned result pages as a reference text and compare it with source text to help identify location names.

3. **Algorithm Description.**

- Step1: Extract the location name set “A{$Loc_{1s},Loc_{2s},…,Loc_{ns}$}” from the source text “S” which already processed by a prototype location name recognition module;
- Step2: Extract a candidate location name string “$Loc_i=c_1c_2…c_j$” from the text “S”;
- Step3: Submit “$c_1c_2…c_j$” to search engine and retrieve the returned result page text “T”;
- Step4: Look up the elements of set “A” in text “T”;
- Step5: If the proportion of the found elements in A is larger than the threshold value “H”, then the candidate string “$Loc_i=c_1c_2…c_j$” is identified as a location name in “s”;
- Step6: If A is not empty, then turn to step2, otherwise exit.

For example:
The text “S” as:

19980501-03-003-006 从 以上 情况 来 看 ， 这 次 元首 会议 虽然 没有 像 去年 10月 基希 讷乌 会议 那样 毫无 成果 、 不欢而散 ， 但 成果 十分 有限 ， 未能 使 元首 们 满意 而 归 。 尤其 感到 遗憾 的 是 哈萨克斯坦/ns 总统 纳扎尔巴耶夫 。
……
19980501-03-003-009 叶利钦 虽 对 他 有 看法 ， 也 只好 同意 。 格鲁吉亚/ns 总统 谢瓦尔德纳泽 认为 ， 这 一 步骤 是 非同寻常 的 、明智 的 、必要 的 。俄 电视台 评论 说 ，看来 独联体 元首 们 对 “ 会 赚钱 的 人 ” 寄托 着 比 官员 们 更 大 的 期望 。 （ 本报 莫斯科/ns 4月 30日 电 ）
19980501 - 03 - 003 - 009 但 有 一点 各国 元首 的 看法 是 共同 的 ，那 就 是 经济 合作 应该 成为 独联体 继续 存在 的 基础 。 俄 原 负责 独联体 事务 的 代理 副 总理 雷布金 不久前 到 独联体 各国 进行 游说 时 曾 提出 ，现在 独联体 各国 经济 中 ，私人 资本 的 作用 都 越来越 大 。 各国 应当 为 企业家 创造 有利 条件 ， 必须 消除 资本 和 商品 流通道/ns 路 上 的 障碍 。

A location name set “{哈萨克斯坦(Hasakesitan/Kazakhstan), 格鲁吉亚(GeluJiya/Georgia), 莫斯科(Mosike/Moscow), 流通道(Liutong Road or Stream Pipe)}” can be obtained from “S” . Now submit the string “流通道” to search engine, and then from the result page formed the text “D”:

文家坝堰塞湖泄流通道已经打通- CCTV 搜索 CCTV 视频播放. 节目信息. 文家坝堰塞湖泄流通道已经打通 2008-05-26. 文家坝堰塞湖水位不断上涨，现在泄流通道已经打通。 中央电视台新闻频道 “关注汶川地震”特别节目 ...
嘉陵江被堵河道清开泄流通道-河道-陇南本报兰州 5 月 15 日讯(记者宋振峰)记者今天从省水利厅获悉：截至 14 日下午 4 时，徽县嘉陵镇下游嘉陵江受地震破坏被堵塞的河道已清开约 3 米宽、1 米深的泄流通道，水流得以下 ...
清江：曾经的长江东流通道--彭鲁在距今 6 亿年以前，湖北大部分地区是一片汪洋大海，其中湖北西部是古地中海向东突出的一个海湾，湖北南半部则浸泡在海底，只有湖北北部极少的一部分凸出在海面。
灯泡混流式水轮机过流通道几何尺寸的优化设计...从最佳的水力性能和水流运动规律出发,对灯泡混流式这种新型水轮机的过流通道进行分析研究,优化出了灯泡混流式水轮机过流通道的几何形状及尺寸系列,为这种新型水轮机的 ...
间隙式节流通道磁流变液减振器实验研究...利用新型智能材料磁流变液,设计了一种混合工作模式的磁流变液减振器.该减振器结构上采用间隙式节流通道,外加磁场方向与磁流变液的流动方向垂直.在 MTS 实验机上对该减振 ...

Look up the elements of the set “{哈萨克斯坦(Hasakesitan/Kazakhstan), 格鲁吉亚(GeluJiya/Georgia), 莫斯科(Mosike/Moscow) }” in text “D” and the result is 0, so it is believed that “流通道” is not a location name in text “S”.

4. **Preliminary Experiments.** To be clear about problems the location name recognition post-processing module would encounter, some preliminary experiments were done.

4.1. **The Construction of Preliminary Experiment Platform.** The experiment was done on basis of two systems: ICTCLAS(Institute of Computing Technology, Chinese Lexical Analysis System)[11] and the CRFs location name recognition system of Xuri Tang in [9]. Test corpus is the People Daily tagged corpus of January and May 1998.

4.2. **Corpus Analysis.** With the restraint of network speed, the average download speed of

web pages is about 5s. If the top 10 links pages which search engine returns are used as a reference text, each candidate location name string needs to wait for 50s. According to one month corpus of the People's Daily, there are 20000 to 30000 location name tokens. It will cost 300 hours to analyze all of them. This is not acceptable.

Fortunately, after analyzing the corpus carefully, we found a method to cut down time complexity.

4.2.1. **Experiment 1: ICTCLAS and the People Daily corpus of January 1998.** First of all, ICTCLAS is used to do word segmentation and Part-of-Speech tagging, comparing with the corpus proofread by manual work as model answer. Experiment results show: precision 98.75%, recall 98.40%.

TABLE 2. the Results of Experiment 1

Total errors of location name tokens	346
Total errors of location name types	282
Errors of location name tokens with occurrence frequency <=2 (denoted by TKE2)	261
Errors of location name types with occurrence frequency <=2 (denoted by TPE2)	225
Proportion of TKE2	75.0%
Proportion of TPE2	79.8%

It is clear that the proportion of TPE2 is close to 80%. Is it feasible to focus on those location names with occurrence frequency <=2? Since ICTCLAS' training corpus is just the People's Daily corpus, and all the location names are on the word list, so whether this analysis result will appear in other lexical analysis system and corpus is not guaranteed.

4.2.2. **Experiment 2: CRFs-based location name recognition system and The People Daily corpus of January 1998.** Obtain location names from The People Daily corpus of January 1998 using a CRFs-based location name recognition system of Xuri Tang in [9], comparing with the corpus proofread by manual work as model answer. Experiment results show as below.

Precision: 0.963902, recall: 0.982323;

Number of location name tokens recognized by CRFs system: 28425;
Number of location name types recognized by CRFs system: 3603;
Number of location name tokens in model answer: 27890;
Number of true positive tokens: 27399.

According to table 3, if obtained location names are ranked by occurrence frequency, the location names with occurrence frequency 1 to 2 are only 10.99% of the total location name tokens, which leads to 69.6% of errors. These less common location names extremely threats the whole system's precision.

TABLE 3. the Results of Experiment 2

Total errors of location name tokens	1026
Total errors of location name types	771
location name tokens with occurrence frequency <=2	3125
Errors of location name tokens with occurrence frequency <=2 (denoted by TKE2)	714
Errors of location name types with occurrence frequency <=2 (denoted by TPE2)	652
Proportion of TKE2	69.6%
Proportion of TPE2	84.6%

Further investigation shows that in these 652 wrongly recognized words with occurrence frequency <=2, there are 634 word types (82.2% of the total errors) with all the occurrences wrongly recognized. That's to say, whenever the location name appears, it is always false positives without other recognition results.

So it is concluded that location names with occurrence <=2 has the characteristics of low frequency, high error rate and simple error pattern. If these location names are submitted to search engine for further processing, it is available to obtain better recognition performance with less time complexity.

5. **Experiment and Analysis of a SE-Aided Location Name Recognition Post-processing Module.**

5.1. **Experimental Platform.** Hardware environment: Intel i3-2100 CPU, 3G memory; network environment: CERNET (The China Education and Research Network); software environment: Windows XP 32bit and Visual C++ 2003.

Experimental Parameter: prototype system is the CRFs-based location name recognition system of Xuri Tang in [9]; test corpus is The People Daily corpus of May 1998, with Baidu as search engine.

5.2. **Experimental Methods and Process.** The experimental process is as follows:

- First, analyze the text recognized by prototype location name recognition system and extract location name strings with occurrence frequency 1 to 2.
- For each location name string form a context set of location names "A" according to discourse information.
- Submit this location name string to search engine and look up elements of set "A" in the search engine's result page.
- If proportion of occurred elements is higher than a predefined threshold (practical value

is 50%), this string is identified as a location name.

Two experiments were done: one used the first page of search engine's result; the other used the search engine's top 10 result links.

5.3. **Base Line and Top Line.** The base line of the experiment is the recognition result of prototype system. The precision is 93.72% while recall is 95.08%.

The top line of the experiment is the location name tokens wrongly recognized with occurrence frequency<=2 are all correctly recognized. On this condition, precision is 97.36% while recall is 98.70%.

5.4. **Experimental Results.**

Experiment A:only using the first page of search engine's result). Our test corpus has 3321 location name tokens with occurrence frequency<=2. Then 2442 of them are correctly recognized, accounting for 73.53% and other 879 location name tokens are wrongly recognized, accounting for 26.47%. So after post-processing, precision is 94.73% while recall is 96.11%, both with 1% improvement.

Experiment B: using the linking pages of search engine's top 10 results. The precision and recall has no significant difference with that only using the first page. So we didn't do any further experiments of B.

5.5. **Experimental Results Analysis.**

- True correction cases:

  True correction cases are those false positives corrected to true negatives.

  With the help of manual analysis, positive correction cases can be classified into 3 categories: other named entities, strings having location name like structure or characters often used in location name, and other tagging errors of prototype system.

  (1)Other named entities.

  Organization names: 宝丰(Baofeng), 方正(Founder); other proper names: 阿昌(Achang), 才旦(Caidan).

  (2)Strings having location name like structure or characters often used in location name. This category accounts for more than 70% of true correction cases.

  保宁津(BaoNingJin), 布市(BuShi), 处江湖(ChuJiang lake)

  第五共和国(The fifth republic), 对口县(DuiKou country), 法律关(Legal Guan), 临门(LinMen), 防撬门(FangQiaoMen)

  (3)Other tagging errors of prototype system

  桃花飞(TaoHuaFei), 罗盘计(LuoPanJi), 买牙膏(Buy toothpaste), 蒙蒙晕(MengMengYun)

- False correction cases:

  False correction cases are those false negatives or false positives are still not correctly recognized.

  With the help of manual analysis, false correction cases can be classified into 4 categories:

(1)The string closely related to location names in a context
This category accounts for 80% of false correction cases. Here are some kinds of examples.
A. Names: 阿尔代尔(Aldair), 安达露西亚(Anda Lucia), 福田纠夫(Takeo Fukuda)
This kind of names has close relationship with the location names in the context. When submitting it to the search engine, the returned results will also contain the same location names. For example, the context location name set of "Anda Lucia" is { Spain,Barcelona,Europe } and the search engine's result page also contains these location names, so "Anda Lucia" is still considered as a location name. But obviously it is a false positive error.
B. The string overlapped with names or location names: 阿德尼瓦说(Aduh Newar Shuo), 怀柔城(Huairou town), 俄原(Eyuan)
C. Other named entities with strong local uniqueness: 苏宁(Suning), 意大利里拉(Italian Lira), 巴勒斯坦人(Palestinian).
(2)Complex location names: 北京四环(Beijing's Fourth Ring Road), 北大西门(the western gate of Peking University),番禺丽江(Li river of Panyu district), 曲阜孔庙(Confucius Temple in Qufu). These are false negatives still not correctly recognized.
(3)Model answers needing discussing: 紫禁城(the Forbidden City), 北极圈(the Arctic Circle), 十里长街(ten miles long street). These proper names whether are location names or not should be discussed.
(4)Word segmentation inconsistencies: 兰州市/ns or 兰州/ns 市/n (Lanzhou City), 浙西/ns or 浙/ns西/f (West Zhejiang).

6. **Conclusions.** The SE-Aided location name recognition post-processing module in the present research has achieves 1% improvement in both precision and recall compared with the prototype system.

Considering the original precision and recall of the prototype system are quite high, the evaluation of the research results is positive. And this post-processing module needs little linguistic resource. We didn't use any language model of high complexity or expert knowledge base.

The future work: First, apply search engine to the establishment of geographical elements database and location name reference disambiguation; second, consider extracting the main part of the page which the search engine's result linking to, and ignore unrelated parts; what's more, deletion of duplicated web pages should be implemented.

REFERENCES

[1] K. Y. Liu, *Chinese text automatic word segmentation and labeling*,1st edition, THE COMMERCIAL PRESS, Peking, 2000.

[2] J. Qian, Y. J .Zhang, and T. Zhang, Research on Chinese name and location name recognition methods based on the maximum entropy, *Journal of Chinese Computer Systems*, vol.27, no.9, pp.1761-1765, 2006.

[3] L. S. Li, D. G. Huang, and C. R. Chen, Automatic identification of Chinese location names with combination SVM and rule , *Journal of Chinese Information processing*, vol.20, no.5,pp.51-57,2006.
[4] H. K. Yu, H. P. Zhang and Q. Liu, Chinese named entities recognition based on cascade a hidden markov model, *Journal of Communication*, vol.27, no.2, pp.87-94, 2006.
[5] D. G. Huang and Y. H. Sun, Automatic identification of Chinese location names, *Computer Engineering*, vol. 32, no.3, pp.220-222, 2006.
[6] N. Li and Q. Zhang, Chinese location name recognition processing analyzed by characters for location names, *Computer Engineering and Applications*, vol.45, no.28, pp.230-232, 2009.
[7] H. P. Liu and X. J. Wang, Named entities recognition of short messages based on conditional random fields and knowledge base, *Journal of Guangxi Normal University*, vol.27, no.1, pp.177-180, 2009.
[8] X. F. Qian and M. Hou, Chinese basic location name recognition, *Application of Language and Characters*, vol.8. no.3, pp.129-135, 2009.
[9] X. R. Tang, X. H. Chen, and C. Xu, Research on Chinese location name recognition based on discourse, *Journal of Chinese Information Processing*, vol.24, no.2, pp.24-32, 2010.
[10] J. P. Ju, W. W. Zhang, and J. J. Ning, Geographical space named entities recognition with combination CRF and rule. *Computer Engineering*, vol.37, pp. 210~215, 2011.
[11] Http://www.ictclas.org/, Retrieved August 21, 2012.