EXTRACTION AND VISUALIZATION OF GEOGRAPHICAL NAMES IN TEXT

Xueying Zhang
zhangsnowy@163.com

Guonian Lv
Zhiren Xie
Yizhong Sun

210046 Key Laboratory of Virtual Geographical Environment (MOE)
Nanjing Normal University, P. R. China

Abstract

Geographical names are the most popular location expressions used to represent geographical references in unstructured text. This paper aims to explore methods for extraction of geographical names in Chinese text and their visualization in Geographical information systems. Extraction of geographical names is a subtask of information extraction that seeks to locate and classify geographical names in text into predefined feature type categories. Approaches based gazetteers can achieve satisfactory performance in English, Spanish and Scottish text. But they are not suitable to Chinese text because of its linguistic characteristics and the lack of a standard gazetteer. Statistical models such as hidden markov models (HMMs) and maximum entropy models (MEMs) and maximum entropy Markov models (MEMMs) are discussed in many documents for word segmentation and extraction of geographic names. Conditional random fields (CRFs) are a probabilistic framework for labeling and segmenting structured data, such as sequences, trees and lattices. The primary advantage of CRFs is their conditional nature, resulting in the relaxation of the independence assumptions required by HMMs, MEMs and MEMMS in order to ensure tractable inference. Our proposed model based CRFs and the algorithm make it possible to effectively and efficiently extract both simple and complicate geographical names in Chinese text without using gazetteers and word segmentation techniques. Naturally one location may have various geographical names and several locations may share one identical geographical name. A set of candidate references or potential locations for an extracted geographical name are provided according to the geographical information system. Then geographical names in text are visualized with both highlight colors in the original context and a spatial overview. In conclusion, this paper discusses a CRFs based statistical model and algorithms for extraction geographical names in Chinese text, in order to bridge the gap of location representations between text and geographical information systems. It is meaningful for intelligent location-base service, geospatial information query and supplement of spatial data attributes.

Key words: Chinese Text, Geographical Name Recognition, Conditional Random Field, Cognitive Salience

1 Introduction

Named entities are usually defined as the real things or instances in the world that are themselves natural and notable class members of subject concepts in natural language processing. They (often multi-world expressions) refer to atomic, specific objects that belong to reference types such as persons, organizations, locations, events, products, time intervals, etc. Before, the terms ‘location name’, ‘geographical name’, ‘place name’, ‘toponoym’ , ‘geographical entity’, ‘spatial named entity’ or ‘geographical named entity’ were often used, sometimes interchangeably in a confusing way. In this paper, we prefer geographical names rather than others. For example, in the sentence “There is a swimming pool at the south of Nanjing University”, the “swimming pool”, the “swimming pool” is a geographical name and refers to a geographical entity but represented with an nominal place name, and anther geographical entity is named with an organization name and a expression of spatial relation.

Extraction of geographical names from unstructured text can be regarded as a sub-task of named entity recognition (NER) in natural language processing. A majority of the previous methods are based on standard large gazetteers, which are proven to be very valuable for this task (Mikheev, 1999). The GIPSY for automatic geo-referencing of text with an algorithm, and a thesaurus incorporating synonymy relations and domain-specific databases are used to detect feature types in the face of linguistic variability and referent sizes heuristically. The TIPSTER looks up candidate referents in the DARPA gazetteer, in which an optional rank number for an entry indicates how salient the referent represented by the entry is (Li 2003). MetaCarta Text Search used an supervised learning algorithm to induce contexts that are positive or negative indicators of terms being geographical names, and to estimate confidence in these indicative contexts. SPRIRT project reports that geographical name recognition with a combination of gazetteer and sop-lists exceeded better performance (Clough 2005). These methods have been applied in geographical information retrieval and mapping services, especially news text and assigning spatial metadata to web page religions.

Unlike English, there is no blank to mark word boundaries in Chinese text. Chinese word segmentation has achieved great progress in the past ten years. However, the results are not satisfactory when unknown words exist in the texts. Moreover, it is unavailable to a standard gazetteer covering most of Chinese geographical names. Therefore, the above-mentioned methods can not be efficient for this task carried on Chinese text (Le, 2006). However, the previous research still focused on syntax rules and word segmentation. A max-margin markov networks was illustrated for identifying unknown geographical names of Chinese text (Li, 2008). The combination of lexical reliability and contextual reliability is proven to be helpful (Huang, 2006; Tan, 2001).

Yu proposed a cascaded hidden Markov model aimed to incorporate person name, location name, and organization name recognition into an integrated theoretical frame. Simple named entity was recognized by lower HMM model after rough segmentation and complex named entity such as person name, location name and organization name was recognized by higher HMM model using role tagging. However, the performance of this model is highly correlated to the word segmentation system (Yu, 2006).

Conditional random fields (CRFs) are a probabilistic framework for labeling and segmenting structured data, such as sequences, trees and lattices. The primary advantage of CRFs over HMMs is their conditional nature, resulting in the relaxation of the independence assumptions required by HMM in order to ensure tractable inference. CRF outperform HMMs on a number of real world tasks in many fields, including bioinformatics, computational linguistics and speech recognition. Recently, CRFs has been widely used in NER and part-of speech (POS) tagging with good performance (Chen, 2006; McCallum, 2003; Lu, 2007). In this paper, we explore a cascaded CRF based model, which aims to extract simple and composed geographical names in Chinese text without using gazetteers and word segmentation techniques. Then geographical names are visualized with both highlight colors in the original context and a spatial overview, for the purpose of bridging the gap between text and space.

2. Basic theory of CRF

A conditional random field (CRF) is a type of discriminative probabilistic model most often used for the labeling or parsing of sequential data, such as natural language processing (e.g. word segmentation and part-of speech) or biological sequences. A CRF is an undirected graphical model in which each vertex represents a random variable whose distribution is to be inferred, and each edge represents a dependency between two random variables. Formally, an undirected graph is defined as $G=(V,E)$. There is a node $v \in V$ corresponding to each of the random variables representing an element Yv of Y, then (Y, X) is a conditional random field. In theory the structure of graph G may be arbitrary, provided it represents the conditional independencies in the label sequences being modeled. The graphical structure of a conditional random field may be used to factorize the joint distribution over elements Yv of Y into a normalized product of strictly positive, real-valued potential functions, derived from the notion of conditional positive dependence.

The probability of a particular label sequence y given observation sequence x to be a normalized product of potential functions is defined by

$$p_\theta(y \mid x) \propto \exp\left(\sum_j \lambda_j t_j(y_{i-1}, y_i, x, i) + \sum_k \mu_k s_k(y_i, x, i)\right) \quad (1)$$

where $t_j(y_{i-1}, y_i, x, i)$ is a transition feature function of the entire observation sequence and the labels at positions i and $i-1$ in the label sequence; $s_k(y_i, x, i)$ is a state feature

function of the label at position *i* and the observation sequence; and λ_j and μ_k are parameters to be estimated from training data (Lafferty, 2001).

In order to define feature functions, it is necessary to construct a set of real-valued features *b(x, i)* of the observation, which expresses some characteristic of the empirical distribution of the training data. These data should also hold of the model distribution. For example, if the observation at position I is the word "September", *b(x, i)* equals to 1, otherwise 0.Each feature function takes on the value of one of these real-valued observation features *b(x, i)* if the current state or previous and current states take on particular values. All feature functions are therefore real-valued (Wallach, 2004). An example of transition function is considered by

$$t_j(y_{i-1}, y_i, x, i) = \begin{cases} \text{b}(x, i) & \text{if } \text{y}_{i-1}\text{=1N and } y_i\text{=NNP} \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

Then the two functions con be simplified by

$$s(y_i, x, i) = s(y_{i-1}, y_i, x, i) \quad (3)$$

$$F_j(y, x) = \sum_{i=1}^{n} f_j(y_{i-1}, y_i x, i) \quad (4)$$

where each $f_j(y_{i-1},y_i,x,i)$ is either a state function $s(y_{i-1},y_i,x,i)$ or a stransition function $t(y_{i-1},y_i,x,i)$. Lets assume that Z(x) is a normalization factor, the probability of a label sequence y given an observation sequence x will be defined by

$$\text{p}(y \mid x, \lambda) = \frac{1}{Z(x)} \exp(\sum_j \lambda_j F_j(y, x)) \quad (5)$$

3. Extraction of geographical names using CRF

Corresponding to the theory of CRF models and the task of extraction of geographical names, we proposed an approach which follows the procedures described in Figure1.

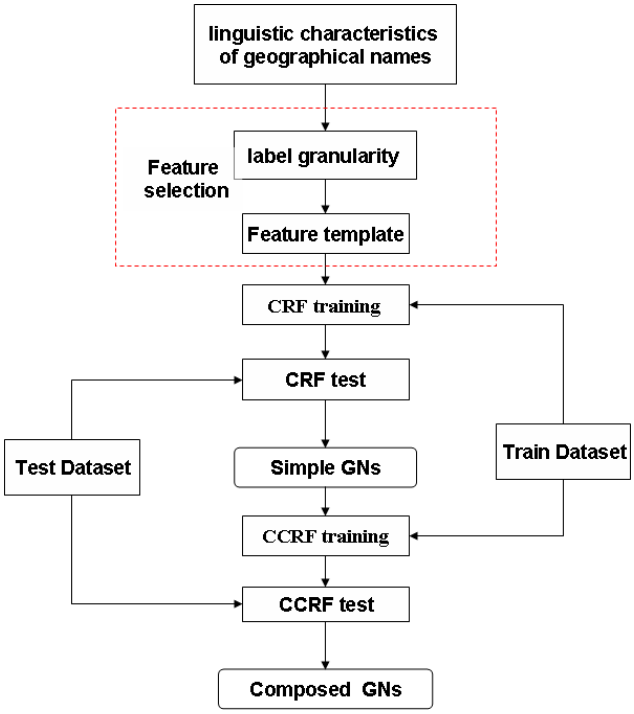

Figure1. The diagram of extraction of geographical names using CRF

3.1 Linguistic of geographical names

China has over 5000 years history and special multi-culture. Most geographical names have been changed temporally and spatially over times. In general, one geographical name may be represented with unlimited number of characters, e.g. a character or 20 characters including a few named entities such as person names, organization names or spatial relations. A simple geographical name means it contains only one named entity, otherwise it is a composed name. In addition, our large-scale statistical investigation shows that 80% Chinese geographical names include generic terms which identify their geographical feature types.

3.2 Feature selection

It is widely accepted that word segmentation cannot improve the performance of named entity recognition. Feature granularity of natural language includes the types: single character (1-gram), word, phrase, concept, N-gram, document, logical block and re-parameterization (Zhang, 2006). 1-gram labeling can solve the data sparsing problem. Undoubtedly, geographical names are always a small part of text. It is reasonable and flexible to select 1-gram as the feature granularity.

In this task, features are extracted from a corpus which has been tagged geographical names. The features are composed of observation values and corresponding labels that

are used to train the estimation parameters of value. In the test module of our approach, input data is used for feature extraction, and put into extraction model. Because CRF models highly depend on features, features selection is an important part of the model. Just as the above mentioned characteristics of Chinese geographical names, we select 4 left and 4 right characters for feature extraction. The feature template using CRF is shown in Table1. Feature selection is an iteration process. The best one could be gained according the balance of training time and experimental performance.

Table 1. Features template using CRF

Feature type	**Relative position**
Front neighbor feature (left)	W_{-4},W_{-3}, W_{-2}, W_{-1}
Back neighbor feature (right)	W_1, W_2, W_3,W_4
Current feature	W_0
Front combined feature	W_{-1} W_0
Back combined feature	W_0 W_1
Transition state	The label state of W_{-1}

3.3 Cascade CRF

Machine learning models usually construct multi-layers with a linear or hierarchical combination. The latter method called cascaded CRF (CCRF) looks at the bottom model as the input of the topper model, which can achieve better combination characteristics. The composed geographical names have much more complicated structures than the simple ones. It is just adaptive to extract composed geographical names with CCRF models. As shown in Figure 2, the lower models are used to extract simple geographical names, and then the outputs are transferred to the upper models.

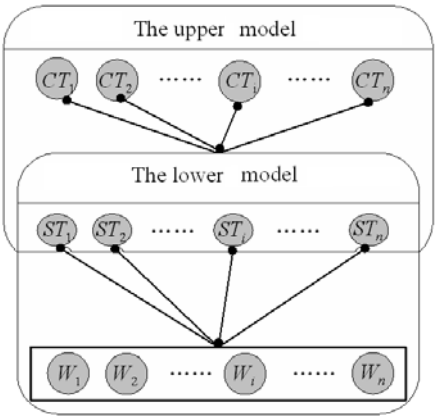

Figure 2. Cascaded CRF for extraction of geographical names

The following example will further describe the mechanism.

Input text: 位于黑龙江省哈尔滨市的哈尔滨市儿童公园为孩子们准备了特殊的贺岁

礼物。(The Harbin Children Park in the city of Harbin in Heilongjiang Province prepared special new year gifts for children.)

Simple extraction: 位于 /ST黑龙江省 (Heilongjiang province) /ST /ST哈尔滨市 (The city of Harbin) /ST 的 哈尔滨市(The city of Harbin) /ST 儿童公园 (Children Park) 为孩子们准备 了特殊的贺岁礼物 。

Composed extraction: 位于 /CT黑龙江省哈尔滨市 (The city of Harbin in Heilongjiang province)/CT 的 /CT哈尔滨市儿童公园 (Harbin Children Park) /CT 为孩子们 准备 了 特殊的贺岁礼物 。

4. Experimental evaluation

We implement two experiments by means of CRF++ downloaded from the website of http://crfpp.sourceforge.net/. One dataset is the People's Daily Corpus (in short PER) with a one million word of Mandarin Chinese, released by the Institute of Computational Linguistics, Peking University. This corpus contains 6 month's data from People's Daily (Jan.-June 1998). The other is MSRA corpus provided by the SIGHAN bakeoff organizers. There are about 1.5 million characters in the training corpus and 223 thousand characters in the testing corpus. It is noted that these corpora are tagged location names which are a little different from geographical names defined in this paper. They are manually modified before they are input to our experimental system. Classical measures such as precision, recall and F1 are selected to evaluate experimental performance.

Table 2. Experimental results with the CCRF models

Train	Test	Precision	Recall	F1	Number of extracted geographical names
PER (Jan.-May)	PER (Jan.)	94.01	94.91	94.46	26185
PER (Jan.-May)	PER (June)	94.30	94.35	94.33	30126
PER (Jan.-May)	MSRA	73.40	73.10	73.25	2674
MSRA	MSRA	93.23	87.78	90.43	3211
MSRA	PER (Jan.)	73.61	67.84	70.61	18718
MSRA	PER (June)	71.90	69.68	70.77	22249

Table 2 indicates that the proposed model can effectively extract geographical names

from Chinese text. The interesting thing is that better performances are always achieved in the same corpus. The main reason is that train and test datasets have consistent tagging specification. Supervised machine learning models highly depends on the scale and quality of train datasets. In real world applications, it is reasonable that about 70% geographical names can be extracted. So we believe that combination of machine learning methods and gazetteers may be a prospective way.

5. Visualization

Naturally one location may have various geographical names and several locations may share one identical geographical name. A set of candidate references or potential locations for an extracted geographical name are provided. Disambiguation of geographical names relates a given geographical name to their appropriate latitude and longitude based a gazetteer. It is assumed that there is always a high degree of spatial correlation in geographical references that are in textual proximity. Then a cognitive salience model is defined by

$$S(r_k, T_j, R_j) = C(r_k) \times S(T_j) \times \sum_{r_i \in R_j} \left[F(r_i, r_k) \times D(r_i, r_k) \right] \quad (6)$$

Where $C(\mathrm{r_k})$ is the reference category salience of r_k; $S(T_j)$ is the discourse distance salience determined by linguistic unit category; $F(r_i, r_k)$ is the frequency salience determined by co-occurrence of r_i and r_k; $D(r_i, r_k)$ is the distance salience determined by geographic distance between r_i and r_k.

Finally we developed a prototype system which can online visualize geographical names with both highlight colors in the original context and a spatial overview (see Figure 3).

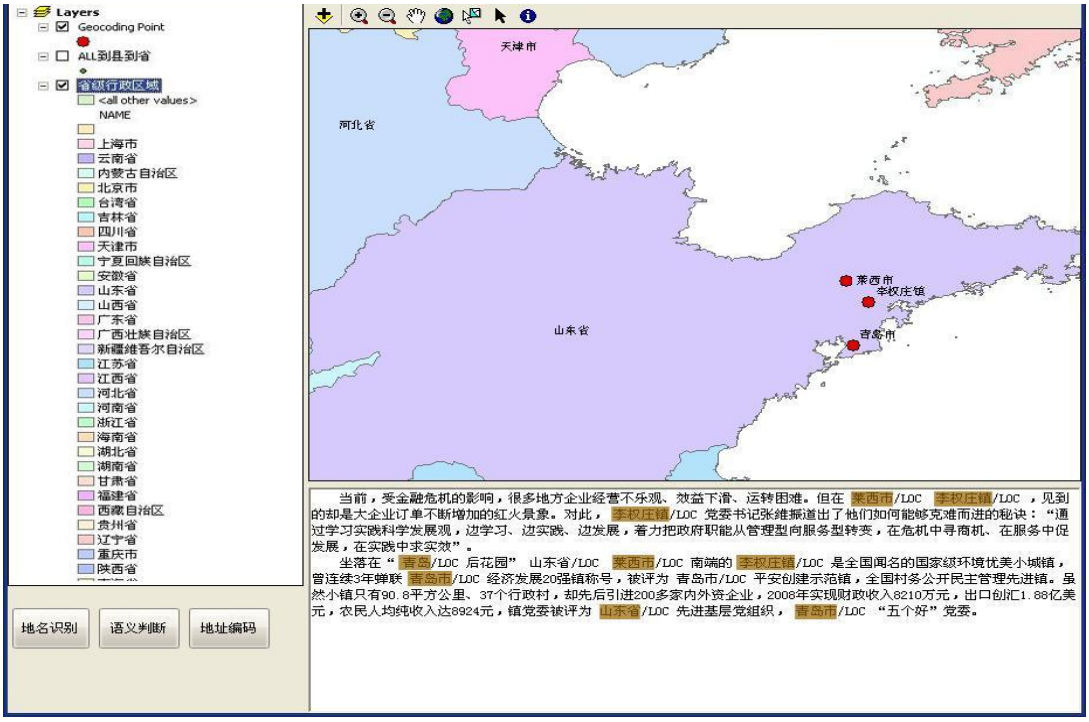

Figure 3. The interface of the prototype system

6. Conclusion

We have presented a CRF based model to extract geographical names in Chinese text. At the same time, the disambiguation and visualization of geographical names are explored. The purpose of this work is to bridge the gap of location representations between text and geographical space. It is meaningful for intelligent location-base service, geospatial information query and supplement of spatial data attributes. Our further efforts will pay attention to identifying and formalization geospatial relations as geospatial input data of geographical information systems.

References

[1] Mikheev, A., Moens, M., Grover C.(1999). Named entity recognition without gazetteers. In: Proceedings of the Ninth Conference of the European Chapter of the Association for Computational Linguistics (EACL), Association for Computational Linguistics, 1–8.

[2] Li, H., Srihari, R.K., Niu, C., Li W. (2003). InfoXtract location normalization: a hybrid approach to geographic references in information extraction. In: Proceedings of Analysis of Geographic References, Association for Computiational Linguistics, 39-44.

[3] Clough P.(2005). Extracting metadata for spatially-aware information retrieval on the Inter-net. In: Proceedings of the ACM Workshop on Geo-graphic Information Retrieval. ACM Press, 25–30.
[4] Le, X. Q. (2006). Research on intelligent web search engine of unstructured spatial information. Dissertation, Institute of Remote Sensing Applications, CAS, Beijing.
[5] Li, L., Ding, Z., Huang, D.G. (2008). Recognizing location names from Chinese texts based on max-margin markov network. In: Proceedings of International Conference on Natural Language Processing and Knowledge Engineering, 1-7.
[6] Huang, D.G., Sun, Y. (2006). Automatic recognition of Chinese place names. Computer Engineering, 32(3):219-222.
[7]Tan, H., Zheng, J. Liu, K. (2001). Research on method of automatic recognition of Chinese place names based on transformation. Journal of Software, 12(11):1608-1612.
[8] Yu, H., Zhang, H., Liu, Q., Lv, X, Shi, S. (2006). Chinese named entity identification using cascaded hidden Markov model. Journal on Communications, 27 (2): 88-95.
[9] Chen, W.L., Zhang, Y.J., Isahara. H. (2006) Chinese named entity recognition with conditional random fields. In: Proceedings of Fifth SIGHAN Workshop on Chinese Language Processing, 118-121.
[10] McCallum. A., Li, W. (2003). Early results for named entity recognition with conditional random fields feature induction and web-enhanced lexicons. In: Proceedings of the 7th Conference on Natural Language Learning, 188-191.
[11] Lu, P., Yang, Y.P., Gao, Y.B., Ren, H. (2007). Hierarchical conditional random fields (HCRF) for Chinese named entity tagging. In: Proceedings of the 3rd International Conference on Natural Computation, 24-28.
[12] Lafferty, J., McCallum, A., Pereira, F.(2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of the 18th International Conf. on Machine Learning, 282–289.
[13] Wallach, H.M. (2004). Conditional random fields: An introduction. Technical Report MS-CIS-04-21, University of Pennsylvania.
[14]Zhang, X.(2006). Rough set theory based automatic text categorization and the handling of semantic heterogeneity. German Social Science Center.