International Journal of Knowledge and Language Processing
Volume **2**, Number 3, July **2011**

www.ijklp.org
KLP International ©2011 ISSN 2191-2734
pp. **36-47**

Structural Analysis and Computation of Chinese Toponyms

Dongmei Zhou, Maohua Qian, Min Hua, Dan Liu, Xuri Tang
Wuhan Textile University
No.1 Textile Road, Wuhan, China
{ ttzhou402@163.com； sarahq@163.com; xrtang@126.com}

Received June 2011; revised July 2011

ABSTRACT. *This paper argues for a new analytical unit, namely Toponym Constituent, for structural analysis of Chinese toponyms. A Toponym Constituent is a determined construction composed of distinctive morpheme, orientation morpheme, descriptive morpheme, category morpheme and part morpheme in the given order, among which only descriptive morpheme is of open class morphemes. Using the analytical unit, the paper conducts a statistical investigation and reveals the major structural patterns used in Chinese toponyms. A toponym recognition model based on Toponym Constituent analysis and Conditional Random Fields is implemented, which achieves precision ratios of 98.02% and 92.93% in close and open test respectively, showing that Toponym Constituent can be used to describe the internal state transition of Chinese toponyms effectively.*

Keywords: Toponym Constituent; Toponym; Structure Analysis; Toponym Recognition

1. **Introduction.** Named entity recognition, in which toponym recognition is an important part, has been one of the major topics in SIGHAN Backoff in recent years. Experiences have also shown that the incorporation of linguistic knowledge with statistic methods may yield promising results in such tasks. The rapid development of Chinese information processing calls for the studies of the internal structure of Chinese toponyms and their external context. The knowledge obtained along the process may be of great use to improve the performance of Toponym Recognition systems. However, as far as we know, in linguistics, the structural analysis based on large-scale Chinese toponym database structure of the formal description of toponym research in linguistics is still rare, possibly because toponym is only a very small and inconspicuous part-of-speech category in linguistic studies. In grammatical studies, a toponym or geographical name is regarded as a word of location with geographic identities, whose grammatical distribution is consistent with the class of nouns class. For instance, a toponym cannot be modified by "not", but can be the complement of “is not”, which shows its identity as a noun[1]. In semantics, toponyms fall into the category of proper names, which is a special type of natural language and is used to refer specifically to a particular region in the objective world. Even in the field of Chinese Information Processing, automatic toponym recognition is just one of the three major tasks in automatic Named Entity Recognition [2]. Linguistic studies associated with toponyms

are focused on three aspects: name of streets, words/morphemes of orientation, and the relationship between toponyms and history and culture. [3] conducts a thorough study on the structures of 55 street names and obtains the structural pattern using word classes such as numbers, verbs and adjectives etc. [4-6] expounds on the variation of street names in Beijing, their associations with history and people, and on the digital forms of toponyms. In regard of words of orientation, [1] classifies them into simple forms and compound forms, and points out that this category of words possesses the characteristics of function words. [5] makes a detailed analysis of the usage of 14 words of orientation (东(east), 南(south), 西(west), 北(north), 前(front), 后(back), 左(left), 右(right), 上(up), 中(middle), 下(low), 内(interior), 里(inside), and 外(outside)) in toponyms of Beijing, and shows the systemic usage of these words in the toponyms. [7] describes the words of orientation in terms of orientation marks, and divides them into quasi-orientation-marks and typical-orientation-marks. Quasi-orientation-marks include "旁(besides)", "变(side)", "头(head)", "部(part)", and "心(heart)" etc., thus expanding the scope of words of orientation. The close link between toponyms and national history and culture has been discussed in the above literatures,and is given a systematic study in [8], which regarded toponyms as "the fruits of mankind's understanding and the accumulation of the ways of human thoughts and psychological characteristics", a very special cultural phenomenon. Both the choice of Chinese characters in toponyms and their structures reflect the features of national culture and cognition process, and carry some historical deposit [9].

There are, however, two limitations on the current analysis of toponyms in linguistics. One is the size of the study. The study scope of toponyms in existing literatures is mainly concentrated on some big cities, and some street names of these cities, but little is done on other cities and streets, or other administrative levels such as county, township, town, village name, and various geographical elements such as rivers, mountains, hills, and streams etc. Yet all of the above mentioned geographic elements can be found in a conspicuous number in everyday language use. The other limitation is the lack of formalized structure analysis. To perform computation on toponyms, it is necessary to analyze the structures of toponyms so as to reveal the internal structural pattern of them. But existing studies are mainly focused on the historical changes and cultural knowledge of toponyms, without a universal and formalized treatment on them.

In the field of natural language processing, the recognition of toponyms relies heavily on machine learning models, and the analyses of the internal structure of toponyms are generally not sufficient. For example, [10] builds a toponym recognition system with characters used in toponyms, and the combination probability between these characters. [11] builds a rule database for toponym recognition, which is based on feature words in toponyms such as "县(county)" and "省(province)" etc. [12] makes use of support vector machine for toponym tagging and uses four states of internal structures: initial, middle, end and suffix. It is easy to see that the above mentioned studies put more emphasis on toponym recognition model, but their use of the knowledge of toponym structure is not sufficient. Feature words of toponyms can be obtained easily, and therefore become the most widely used structure knowledge of toponyms. But the rules of internal toponym structures, the position of toponym feature characters (or general names) in toponym structures and the combination relationship between toponym feature characters and other parts are less studied and thus not often applied. Because of the lack of structural analysis,

machine learning models often use the state sequence tag set based directly on a simple relationship, such as four-label-tagging or six-label-tagging, to mark the internal relationship. The lagging behind of toponym structure studies and the lack of toponym structure knowledge are the major barriers in improving the accuracy of automatic toponym recognition.

Given the above understanding, this paper makes a more comprehensive and in-depth analysis of the internal structure of Chinese toponyms and builds a system which recognizes toponyms by computing toponymhood of strings with Conditional Random Fields. Following the principle of recursiveness, this paper firstly proposes to use "toponym constituent" as the basic unit, and then conducts structural analyses on toponyms found in the corpus of *People's Daily*[1] to build a toponym structure database. A statistical analyses is then done to investigate the distribution of different morphemes, namely distinctive morpheme, orientation morpheme, category morpheme, part morpheme and descriptive morpheme in toponyms, together with the cultural and cognitive characteristics of these elements' involvement in the process of toponym formation. These analyses help to obtain a more comprehensive and in-depth understanding of the internal structure of toponyms. Meanwhile, to study the effectiveness of "toponym constituent", this paper constructs a tag set for the internal state of toponym structures and then use the tag set to build a toponym recognition model system based on Conditional Random Fields, which achieves a precision of 98.02% and 92.93% in close and open test respectively.

2. **Toponym constitute.** Generally, Chinese words can be classified into simple words and compound words according to the complexity of their internal structures [14]. As for toponyms, three categories[2] can be obtained by their complexity (Table 1): Simple Toponyms made of one or more morphemes; Compound Toponyms made of Simple Toponyms and other subsidiary structures; and Shortened Forms.

TABLE 1. Toponym Classification

Category names	Examples
Simple Toponym	北京(Beijing), 高界(Gaojie), 卢湾(Luwan)
Compound Toponym	北京市(City of Beijing), 玄武区(District of Xuanwu), 南大街(Southern Street), 中华人民共和国(People's Republic of China)
Shortened Toponym	京(Jin), 沪(Hu) , 辽(Liao), 湘(Xiang)

Simple words and compound words are often used in analyzing Chinese word structures. They are both found inside all the three categories. However, these analytical units are fairly abstract when applied for particular categories such as toponyms and may not provide enough information for such purpose. For example, the delimitation between Simple Toponyms and Compound Toponyms is blurred. Toponyms such as "南大街" and

[1] The corpus covers half a year's daily newspaper (from January to June of 1998) of *People's Daily*, which is segmented and tagged with POS by Peking University.

[2] There's another category in toponym recognition, namely Complex Compound Toponym, such as "北京市海淀区(Haidian District, City of Beijing)", "南京市鼓楼区(Gulou District, City of Nanjing)". Complex Compound Toponyms generally refers to one location, not two.

“广州东(East of Guangzhou)” contains morphemes of location and morphemes of orientation, and should be classified as Compound Toponyms. But these words do not differ greatly from Simple Toponyms such as “北京” and “广州(Guangzhou)” in terms of syntactic distribution. Thus the distinction do not help much in toponym recognition. In addition, those analytical units can not reveal the internal relationship between constituents. Another example is “北京市” and “东市口”, which are both regarded as Compound Toponyms, but the relationship between “市(city)” and “北京” is different from the relationship between “口(exit)” and “东市(East Market)”.

Analysis on toponyms shows that the internal structure of toponyms abides by the principle of recursiveness[15], in which one structure is recursively used to build a larger unit. In this paper, this structure is named as Toponym Constituent, which is defined as below:

Toponym Constituent (TC) is the basic analytical unit of toponyms, which consists of five optional elements in sequential order: Distinction Morpheme (Dst), Orientation Morpheme(Ori), Descriptive Morpheme(Mod), Category Morpheme(Cat) and Part Morpheme(Part).

As the elements in TC are optional, it is thus not necessary for all the five elements to be present together to make a TC. The feature of sequentiality requires the five elements to appear in a TC in the given order. A toponym may consist of one or more Toponym Constituents. This is adhere to the principle of recursiveness in language, as complex structural units are often composed by recurrence of simple structural units[15], which makes it possible for the mechanism to adapt to the ever-changing situation of toponyms.

The term “morpheme” used in TC is identical with what is defined in [16]: a morpheme is the smallest combination of meaning and form. However, TC differs from Simple Word or Compound Word in two aspects. Firstly, A TC can be a Simple Word such as “北京”, or a Compound Word such as “北三环(Northen Sanhuan)”. The second aspect is that the structure of a TC is deterministic in that the morphemes are always arranged in a sequential order, and the elements inside are optional, allowing for the absence of one or more elements. But for a specific TC, its components are deterministic with a fixed order and relation between them.

By way of TC, a universal analytical approach can be achieved for both Simple Toponym and Compound Toponym, thus avoiding the problem of blurred delimitation between them. Table 2 illustrates the approach. In the table, No. 2 and No. 3 are of one TC which can be used independently in natural language, but No.1 and No. 6 have different structures, where in No.1 has only one TC and No.6 has two TCs of which No.1 forms a part. As is seen, the use of TC provides a universal approach to describe both Simple Toponyms and Compound Toponyms.

TC also makes it possible to describe the internal relationship among constituents inside a toponym, thus make it easy to identify those toponyms with the same structural patterns or those with different structural patterns. For example, the morpheme “市” in “北京市” and “壮族自治区(Autonomous Region)” are identical in distribution and structural function, thus tagged as “Cat”. While the morpheme “北(North)” in “北京” and “西(west)” in “广西壮族自治区” are both morpheme of orientation, but the two occur in different positions of order and perform different structural functions. The TC approach reveals this

difference.

TABLE 2. Samples of toponym constituent analysis

Index	Toponyms	Structural analysis
1	北京(Beijing)	L-Struct [Ori<北> + Cat<京>]
2	广州东(East of Guangzhou)	L-Struct[Mod<广> + Cat<州> + Ori<东>]
3	南大街(South Street)	L-Struct[Ori<南> + Mod<大> + Cat<街>]
4	昆仑山(Kunlun Mountains)	L-Struct[Mod<昆仑> Cat<山>]
5	西山(Western Hills)	L-Struct[Ori<西> Cat<山>]
6	北京市(The City of Beijing)	L-Struct[Ori< 北 > + Cat< 京 >] L-Struct[Cat<市>]
7	广西壮族自治区(Guangxi Zhuang Autonomous Region)	L-Struct [Mod< 广 > Ori< 西 >] L-Struct[Mod<壮族> Cat<自治区>]
8	南横街东口(East Exit of South Side Street)	L-Struct[Ori< 南 > Mod< 横 > Cat< 街 >] L-Struct[Ori<东> Part<口>]

TC based approach also provides a mechanism to identify different forms of toponyms with the same reference of geographic entity. In Chinese, one geographic entity may be referred to by different toponyms. For example, the toponyms “西双版纳傣族自治州”, “西双版纳”, “西双版纳自治州”, and “西双版纳州” can be used to refer to the same geographic entity. This is hard to explain using the concepts of Simple Toponym and Compound Toponym, but can be easily explained with TC-based approach. All the other forms of representation can be seen as the result of “drop-off” of elements of the toponym “西双版纳傣族自治州”. Table 3 gives a brief description of such phenomenon. A number of rules can be deduced from observation of this table. One possible rule is that if a toponym consists of more than one TC, the first TC is to remain in the dropping off, while the others may be completely or partially dropped and that in the partial dropping off, the “Cat” morpheme remains while the “Mod” morphemes can be dropped.

TABLE 3. Dropping Off of TCs

西双版纳傣族自治州	L-Struct [Mod<西双版纳>] L-Struct [Mod<傣族> Cat<自治州>]
西双版纳	L-Struct [Mod<西双版纳>]
西双版纳自治州	L-Struct [Mod<西双版纳>] L-Struct [Cat<自治州>]
西双版纳州	L-Struct [Mod<西双版纳>] L-Struct [Cat<州>]

3. **Morphemic analysis of geographic name components.** [16] holds that morpheme is the basic unit of word formation. However, toponym is a particular type of linguistic unit. Its external syntactic functions and semantic features would inevitably require that the morphemes inside it behave differently from morphemes in other word classes. These characteristics are mainly shown in three aspects: (1) the number of morpheme types in toponyms are limited, which includes only distinctive morpheme, orientation morpheme, descriptive morpheme, category morpheme and part morpheme; (2) except for descriptive morpheme, the other types of morphemes are closed and have members which can be listed in a exhaustive manner; (3) descriptive morphemes have distinctive national cultural characteristics, which may lead to a systemic distribution in terms of the characters used in these morphemes.

3.1. **Closed class morpheme**. In Toponym Constituents, distinctive morphemes, orientation morphemes, descriptive morphemes, category morphemes and part morphemes are of closed classes which can be listed out exhaustively in Chinese.

3.1.1 **Distinctive morpheme.** Distinctive morphemes are used to make finer distinction of geographic entities. Example 1 illustrates the phenomena. Theoretically, all the distinctive adjectives can serve as distinctive morphemes in toponyms, but the statistics of the toponym structure database show that there are only 13 distinctive morphemes, including "大(big)", "小(small)", "新(new)", "老(old)", "古(early)", "白(white)" and "青(blue)" etc.

Example 1

小西山(Small West Hill)	大西山(Big West Hill)
老城区(Old Urban Areas)	新城区(New Urban Areas)
民主德国(Democratic Germany)	联邦德国(Federal Germany)
白尼罗河(White Nile)	青尼罗河(Blue Nile)

3.1.2. **Orientation morpheme and part morpheme.** Orientation morpheme and part morpheme share some similarities to a certain extent. As a matter of fact, the orientation morpheme here mentioned is the same as "the typical orientation mark" in [7], including mainly "东(east)", "南(south)", "西(west)", "北(north)", "前(front)", "后(back)", "左(left)", "右(right)", "上(up)", "下(down)", "中(center)", and "里(internal)" etc., while "the semi-orientation mark" in [7] are almost the same as part morphemes, such as "口(mouth)", "头(head)", "尾(tail)" and "底(bottom)" etc.. Orientation morpheme and part morpheme present some differences in signifying spatial orientation. Generally, orientation morphemes can mean both external and internal spatial orientation of an object, while part morphemes can only signify the inner spatial orientation or places, as can be seen in Example 2.

Example 2

岔口(Mouth of Fork)	城关(Portal of City)	桥头(Head of Bridge)
杉树脚(Foot of Firs)	岔南(South of Fork)	城南(South of City)
桥西(West of Bridge)	港北(North of Port)	

The second factor distinguishing orientation morpheme from part morpheme is different positions of the two morphemes in toponyms. Orientation morphemes can be found before the descriptive morphemes as in "北大仓(North Dacang)" and "北大街(North

Big Avenue)", or after descriptive morphemes as in "百下区(Lower part of Bai District)" and "察南(Southern part of Cha)". Yet part morphemes can only occur after descriptive morphemes, not before them.

Analysis based on spatial cognition shows that the orientation morphemes before and those after descriptive morphemes indicate different cognition results on space. The former takes the language user as the frame of reference, and project the toponym as a "point" in the space, while the latter takes the place the toponym signifies as the frame of reference, and projects it as an "area". The statistics in toponym structure database show that 37 orientation words in total appear before descriptive morphemes, including the single syllable morphemes such as "东(east)", "南(south)", "西(west)" and "北(north)" etc., and the double-syllable morphemes such as "中心(center)", "(东北(northeast)", "北部(northern)" etc.. There are 1235 toponyms, about 7% of the toponyms in the database contains orientation morphemes before description morphemes. In the database, 1080 toponyms, about 6% of the database, contains orientation morphemes used after the description morpheme, in which 24 orientation morphemes are used. It should be noted that 38 toponyms contain both the former and the back morphemes in the database.

23 part morphemes in total appear in toponyms, such as "口(mouth)", "头(head), "边(side)", "滨(river-side)", "脚(foot)", "尾(tail)" and "嘴(mouth)" etc., and 336 toponyms contain part morphemes, accounting for 2% of the total.

3.1.3. **Category morphemes.**Category morpheme is also called "General name" in Chinese linguistic studies, which is generally used to mark category of the toponym. Category morpheme often presents the language users' view on the outstanding geographic features of the place signified by the toponym. Because places differ from one another in geographic features, there exists a fairly large number of category morphemes. In the toponym database, there are 351 category morphemes in total. And the category morpheme can be further divided into 3 sub-types: administrative regions, natural factors and artifacts, illustrated in Example 3.

Example 3

Administrative divisions: City, Country, District, Street, Prefecture (州),Province, Village, Town, Jiang(疆),Tun(屯),Du, County, Township, League(盟) etc.

Natural factors： Bay, Mountain, Zhen(圳), River, Sea, Lake, Di(地),Gorge, Pu(浦),Chuan(川), Lin(林),Island, Ling(岭),Ling(陵) etc.

Artifacts: Li(里), Road, Fang(坊), Gate, Nei(内), Tan(坛), Cheng(城), Dao(道), Yuan(园), Ting(亭), An(庵), Gang(港), Lou(楼), Jin(津), Pu(铺) etc.

Nearly 80% of toponyms involve one or more category morphemes, which suggests that category morphemes are quite essential in toponym identification and are often called "toponym feature word" in many toponym recognition system and serve as an important feature in distinguishing toponym from other word classes. Both [11] and [12] make use of this type of morpheme in toponym recognition

3.2. **Open Class Morphemes.**

3.2.1. **Descriptive Morpheme.** Descriptive morpheme refers to the constituent which provides a description of the place not included in orientation morpheme, distinctive morpheme, part morpheme and category morpheme, such as the "安定(safe and steady)" in "安定门(Anding Gate)", "临高(Before height)", "麟游(Linyou)", "凌源(Lin Yuan)" and

“马达加斯加(Madagascar)”. The descriptive morphemes can be classified into four categories in terms of the historical relation process between symbol and the signified:

(1) Transliteration. The morpheme comes into existence via transliteration, namely, choosing the closest corresponding pronunciation in the target language for translation. In Chinese toponyms, not only the foreign places, but also the places of minority nationality regions, are transliterated. What’s more, many of them have been widely spread and accepted before the regulation of transliteration comes into being.

(2) Ancient Chinese Characters, which is generated in the long process during which the characters were adopted, such as “泗(si) and “淮(huai)”. Using these Chinese characters alone can make it clear the signified places.

(3) Feature description, which is an important method in toponym formation. The features used in the description vary from topographical features to plants and animals, produce, historic events or figures. Structurally, various types of syntactic patterns such as Subject Predicate Structure, Predicate Objective Structure, Predicate Complement Structure and Modified Nominal Expressions are used for such purpose.

(4) Wishes and Inclination. Historically, toponym can be a manifestation of the language worship, instead of a simple describing method. In other words, this can be seen as an exaggerated employment of “language action” which expresses best wishes of language users. The has led to the fact that a number of commendatory morphemes are used this category, such as “安定(safe and steady)” and “东安(safe in the east)” in syntactic patterns of Head Modifier Structure, Subject Predicate Structure and Predicate-Objective Structures etc.

4. **Structural Patterns of Toponyms.** Using TC as the basic analytical unit, the paper conducts a thorough analysis on a large number of Chinese toponyms and builds a toponym structure database of approximately 20,000 Chinese toponyms. Investigation on the database reveals the following regularities in the structural patterns of toponyms:

(1) The majority of the toponyms are constituted by one or two elements. In the database, the number of toponyms made up by one TC is 12,939, taking up 69% of the total.

(2) There is none of the five types of morphemes in TC which should be found in all toponyms. Descriptive morpheme is the only type of morpheme that can constitute a toponym alone. For instance, the toponyms “天目(eye of sky)”, “偃师(Yanshi)”, and “福建(Fujian)” include only descriptive morphemes. However, in some other toponyms, as are in type 6 in Table 4, there is no descriptive morpheme at all. Category morphemes may not appear in all types of toponyms, as are in type 2, type 4 and type 5 in the Table. That is contradictory to the common sense that almost toponyms contain one ore more category morphemes. As a matter of fact, this assumption ignores about 20% toponyms in the database.

(3) The internal structure of some toponyms are rather complicated, at most four TC can be found in one toponym in the sequential order, as are seen in Example 5.

Example 5.
井冈山市(City of Jingan Hill)：
L-Struct[Cat<井>] L-Struct[Cat<冈>] L-Struct[Cat<山>] L-Struct[Cat<市>]

凤阳府城镇(Town of Fenyan Fu Cheng)：
L-Struct[Mod< 凤 > Ori< 阳 >] L-Struct[Cat< 府 >] L-Struct[Cat< 城 >] L-Struct[Cat<镇>]

沟河庄乡(Xiang of Ditch River Village)：
L-Struct[Cat<沟>] L-Struct[Cat<河>] L-Struct[Cat<庄>] L-Struct[Cat<乡>]

In example 5, most of the repeated TCs are category morphemes. This can be attributed to the process of formation of these toponyms. On one hand, toponyms have close relationship with geographic features and thus often takes on category morphemes of geographic category morphemes. On the other hand, toponyms are also subjugated to the historical change of government and administration, which will attach those administrative category morphemes to the original toponyms.

(4) The internal structural patterns of toponyms are relatively complicated. Table 4 gives the statistics of structural patterns. It can be seen that type 1, type 2 and type3 takes up to 77.8% of the total, while the rest is about 22%. In addition, these types are of more or less the same distribution. The complexity of the internal structures of toponyms is one of the major setbacks for toponym recognition.

5. **Toponym recognition.** The structural analysis based on TC provides the basic framework for knowledge representation of toponyms, which can be used to build statistical models for toponym recognition. Toponym recognition can be viewed as a task of annotation of hidden states, for which TC provides a tag set for all the internal hidden states. For example, “北京市(City of Beijing)” can be described as “北/LOC 京/LCC 市/LCC”, which indicates the hidden status transition sequence from orientation morpheme to category morpheme and then to category morpheme again. The state tag set provides a sound linguistic rationale for the toponym recognition system based on statistic machine learning model, and thus forms reliable conditions for high precision automatic toponym tagging system. In this paper, the model of Conditional Random Fields[3] is chosen as machine learning model to build a system which is designed to judge whether a given string is a toponym or not without taking the context into consideration. For example, given two strings, such as “渌口(Lukou)” and “一个村(a village)”, it should be able to decide “渌口” is a toponym while “一个村” is not. In the experiment, the wordlist used for training are the words used in the first five months(Janguary-May) of the *People’s Daily* corpus, while the wordlist used for test is constructed from the last month of the corpus. During the training, tag set for toponyms is obtained by toponym structural analysis (see Table 5). Other non-toponyms are analyzed with a six-name tag set[4] (B,C,D,I,E,S)to mark the location of characters in the word. The experimental results are given in table 6.

[3]The experiment is done with CRF++(0.51) by Taku Kudo, which is available at http://crfpp.sourceforge.net/.
[4]E.g. 如/B 果/E 没/B 有/E 工/B 人/C 阶/D 级/E 的/S 支/B 持/E

TABLE 4. Structural Patterns with Frequency High than 200

NO.	Structure Type	Freq	Rate	Examples
1	L-Struct[Mod<> Cat<>]	7724	41.5%	狗叫屯(Gouijao Village),古巴共和国(The Republic of Cuba),古交市(Gujiao City), 三惠桥 (Sanhui Bridge)
2	L-Struct[Mod<>]	3510	18.8%	天目(Tianmu), 偃师(Yanshi),福建(Fujian)
3	L-Struct[Mod<> Cat<>] L-Struct[Cat<>]	3258	17.5%	福井县(Fujing County), 九台市(Jiutai City),酒泉市(Jiuquan City), 喀拉山山口(Kala Mountain Mouth)
4	L-Struct[Mod<> Ori<>] L-Struct[Cat<>]	392	2.1%	阿北乡(Abei Village), 巴东县(Badong County),南县(Banan County),白下县(Baixia County)
5	L-Struct[Mod<> Ori<>]	348	1.8%	安西(Anxi),白面下(Baimian Xi),保北(Baobei)
6	L-Struct[Ori<> Cat<>] L-Struct[Cat<>]	342	1.8%	北碚区(Beibei District), 北道区(Beidao District ）, 北甸子乡(Beidianzi Village）
7	L-Struct[Ori<> Mod<> Cat<>]	327	1.8%	北冰洋(Arctic Ocean ）,北大仓(Beidacang）,北大街(Bei Avenue)
8	L-Struct[Ori<> Cat<>]	286	1.5%	北部湾 (Southern Bay), 东庄(Dongzhuang), 南海 (South China Sea)
9	L-Struct[Cat<>] L-Struct[Cat<>]	218	1.1%	厂甸(Chang Dian),池河(Chi River), 关镇 (Guan Village), 界岭 (Jie Mountain)
10	L-Struct[Mod<>] L-Struct[Mod<> Cat<>]	202	1.1%	楚雄彝族自治州(Chuxiong Yi nationalitiy autonomous prefecture), 椿树二期(Chunshu second-stage), 广安大街(Guangan Avenue)

TABLE 5 Simple toponym tag set and examples based on toponym structure

Tag set	explanation	examples
LDC	Distinctive morpheme	白尼罗河(White Nile River)：白/B-LDC 尼/I-LMC 罗/I-LMC 河/E-LCC
LOC	Orientation morpheme	龙阳(LongYang)：龙/B-LMC 阳/E-LOC
LMC	Descriptive	龙舟池(LongZhou Pond)：龙/B-LMC 舟/I-LMC 池

	morpheme	/E-LCC
LCC	Category morpheme	龙株岗(LONGZhu Mound)：龙/B-LMC 株/I-LMC 岗/E-LCC
LPC	Part morpheme	渌口(LuKou)：渌/B-LMC 口/E-LPC

TABLE 6 Toponym recognition experimental result based on Conditional Random Fields

Tag type	**Testing corpus**	**precision(%)**	**Recall rate(%)**	**F(%)**
Closed test	PU corpus January	98.02	89.68	93.66
Open test	PU corpus June	92.93	80.85	86.47

From the experimental results, it can be seen that the system achieves a fairly high accuracy in recognizing toponym.

6. **Conclusions.** This paper analyzes nearly 20,000 toponyms based on a new analytical unit —Toponym Constituent for internal structure of toponyms and constructs a large scale Toponym Structure Database. The statistical investigation of the database provides an overall understanding of the internal structure of Chinese toponyms, and the general principles governing the formation of toponyms. The toponym recognition system based on CRFs is able to achieve precision ratios of 98.02% and 92.93% in close and open tests respectively, without using contextual information. The experiment testifies that Toponym Constituent can be used to describe effectively the internal state transition of Chinese toponyms. We also believes that the obtained internal structural state of indicated by the toponym tag set can also be used in other machine learning modes such as Maximum Entropy and Hidden Markov, and provide linguistic knowledge resources and framework to achieve high accuracy of toponym recognition. This is also our plan for further researches.

REFERENCES

[1] WenLian, *Location, time and direction*, Shanghai Education Press,1957.

[2] L. Hirschman and N. Chinchor, "Muc-7 named entity task definition.," in *Proceedings of the 7th Message Understanding Conference (MUC-7)*, Fairfax, Virginia, 1997.

[3] Ma Qingzhu, Street Names and its ways of formation, *Linguistic Study Symposium*, Vol. 6,Tianjing Education Press, pp. 153～177,1991.

[4] Zhang Qingchang, Some Linguistic Problems on the Street Names Evolution in Beijing since Ming and Qing dynasty, *Selected Papers of Modern Chinese Vocabulary*, Zhou Jian, Ed, The Commecial Press, 1985, pp. 62-67.

[5] Zhang Qingchang, Three Problems in Beijing Street Names, *Chinese Language*, vol. 1996, pp. 428-432, 1996.

[6] Zhang Qingchang, 14 Orientation Words in Beijing Street Names, *Chinese Language*, vol. 1996, pp. 10-15, 1996.

[7] Xing Fuyi, Li Xiangnong and Chu Zexiang, Time, Orientation and Place, *Introduction to Grammar Study*, Ma Qingzhu et al. Ed. The Commercial Press, 1999, pp. 472-483.

[8] Deng Huirong, Probing the Thinking Mode and Social Psychology of the Hans in the View of Chinese Toponyms, Academic Exchange, vol. 2003, pp. 138-141, 2003.
[9] Wu Zhirong, Discussion on Toponym Wording, *Cartography*, vol. 2006, pp. 42-43, 2006.
[10] Liu Kaiying, Automatic Word-separation and Tagging of Chinese Texts, The Commercial Press, 2000
[11] Huang Degen, Automatic Recognition of Chinese Toponyms, Computer Engineering, vol. 32, pp. 220-222, 2006.
[12] Li Lishuang, Huang Degen, Chen Chunrong and Yang Yuansheng, Identification of Location Names from Chinese Texts Based on Support Vector Machine, Journal of Dalian University of Technology, vol. 47, pp. 433-438, 2007.
[13] Yu Hongkui, Zhang Huaping, Liu Qun, Lv Xueqiang and Shi Shuicai, An approach based on Chinese named entity identification using cascaded hidden Markov model, *Telecommunication Journa*, vol. 27, pp. 87-94, 2006.
[14] Zhu Dexi, *Lecture Notes on Grammar Biejing*, The Commercial Press, 1982.
[15] Qian Guanlian, *Theory of Language Holography*, The Commercial Press, 2002.
[16] Liu Shuxin, *Chinese Descriptive Lexicology*, The Commercial Press, 2005.