Research on a Model of Extracting Persons' Information Based on Statistic Method and Conceptual Knowledge*

XiangFeng Wei[a], Ning Jia[a,b], Quan Zhang[a], HanFen Zang[c]

[a]Institute of Acoustics, Chinese Academy of Sciences, China
[b]Graduate University of Chinese Academy of Sciences, China
[c]Captical University of Economics and Business, China
{wxf, zhq}@mail.ioa.ac.cn, gnin_aij@sina.com, zanghf@163.com

Abstract. In order to extract some important information of a person from text, an extracting model was proposed. The person's name is recognized based on the maximal entropy statistic model and the training corpus. The sentences surrounding the person's name are analyzed according to the conceptual knowledge base. The three main elements of events, domain, situation and background, are also extracted from the sentences to construct the structure of events about the person.

Keywords: Person's Name Recognition; Hierarchical Network of Concepts; Main Elements of Events; Semantic Parsing; Information Extraction

1. Introduction

In news reports, person is one of the most important elements. On the internet politicians, famous enterprisers, celebrities and hot persons are always the focus what people want to know. It would be very useful and valuable to compile and extract the information about a person with the assistance of using automatic information extraction technology. ACE (Automatic Content Extraction) program is organized by NIST (National Institute of Standards and Technology, U.S.A). It is aimed to develop automatic content extraction technology to support automatic processing of human language in text form, including the detection and recognition of entity, value, time, relation and event. In the description of events, ACE defined the allowable roles of an event, such as person, place, position, etc. Persou, which is studied by Liu and etc. (2006), is a web search engine for searching persons' information. It can distinguish automatically or semi-automatically the persons who own the same name, and produce their resume and activities. A fame evaluation system studied by Zan and etc. (2003) can calculate the correlative degree between the basic information such as name, specialty, affiliation, character words of a person and the information of a web page. If the count of the strong correlative web pages is greater than a threshold, the relative person will be a famous person. In above studies, they all adopted statistic method. The advantage of using statistic models is high efficient performance

* This work is supported by the National Basic Research Program of China (973 Program, the contract No. 2004CB318104) and the Fund for Excellence by the Director of the Institute of Acoustics, Chinese Academy of Sciences (the contract No. GS13SJJ04).

to process large-scale data, but the precision is restricted by training set and it is not suitable to process sparse individual data, although the data is very important to the user.

This paper proposed an model of information extraction based on statistic algorithm and conceptual semantic knowledge to extract persons' information from the text. The primary information of a person includes name, gender, age, nationality, department in organization, career, title, and so on. The name of a person was recognized by using maximal entropy statistic model. The text around the name was analyzed based on conceptual model. According to some associated key words about the primary information, the information was extracted from the context. If there are some events in the text, the main elements of events will be extracted based on the conceptual structures of a sentences and the conceptual knowledge base, which was stored in computer in advance.

2. The model of extracting persons' information

Figure 1 shows the whole process of extracting a person's name, his or her primary information and events about a person. In this process, recognizing persons' names is one of the key sub-processes. In order to recognize a person's name, there must be some tagged texts which indicate the right persons' names. In the tagged texts, the contexts around the names present some statistic characters. Therefore, recognizing a person's name was transformed into classifying a word belongs to a name or not. This paper focused on Chinese names. A Chinese name is composed of two parts. The first part is family name and the second is given name. Most family names are limited in about 100 Chinese characters. These characters are used to activate persons' names. The Chinese character (one or two) behind the family will be considered as candidates of a person's name. A classifier can divide all candidates into name or non-name according to the statistic model based on the tagged texts.

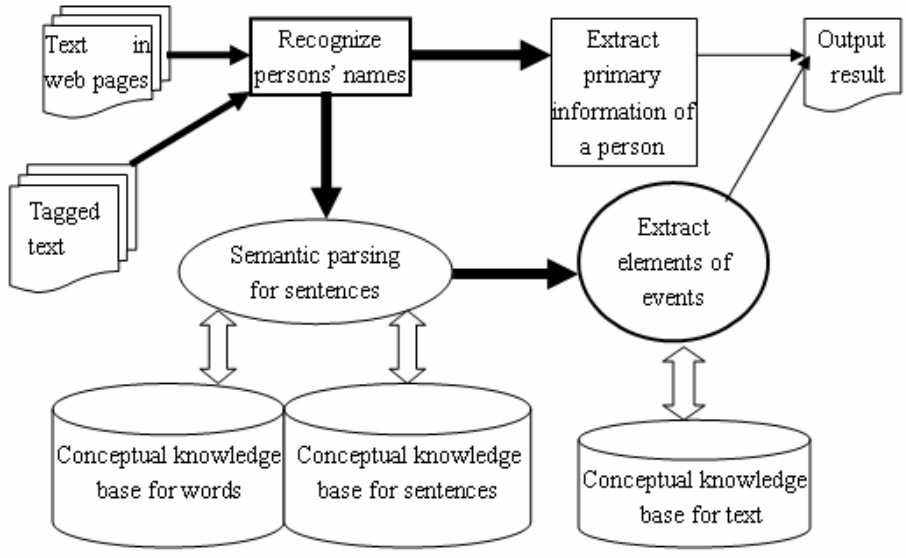

Figure 1: The model of extracting persons' information.

Once a person's name is recognized, the surrounding sentences will be analyzed to extract the primary information and the events about the person. To extract the primary information of a person, there are some activating words to point out the information. For example, 'he' pointed to a person's name indicates that the person's gender is male, and 'president' points out the person's title. To extract the information of events about the person, we must analyze the related sentences at first. All sentences are mapped into the conceptual structures, and the main elements of events are extracted from text based on conceptual knowledge bases. There are

three kinds of conceptual knowledge bases for words, sentences, and text respectively, as showed in figure 1.

3. Recognition of Chinese names

The maximal entropy model is a well adaptive, flexible, and high-efficient statistic model. It can synthetically process all kinds of related and irrelated characters from data. The key point of the model is fitting the known data and equably distributing the unknown data. To apply the maximal entropy model, the restricted condition is described as an character function f(a,b) {0,1} with alternative value. For example, a function to judge whether a noun is name is described as formula (1).

$$f(a,b)=\begin{cases}1 & (a=\text{Pr}esident)\cap(b==Noun),\\ 0 & otherwise.\end{cases} \quad (1)$$

For a character function, its expectation relative to the experiential probability is calculated as formula (2).

$$E_{\tilde{p}} f_i = \sum_{a,b} \tilde{p}(a,b) f_i(a,b). \quad (2)$$

Its expectation relative to the model is calculated as formula (3).

$$E_p f_i = \sum_{a,b} \tilde{p}(b) p(a \mid b) f_i(a,b). \quad (3)$$

In the training data set, the two expectation are assumed to be equal as formula (4).

$$E_p f_i = E_{\tilde{p}} f_i. \quad (4)$$

If there is more than one restricted condition, a classifying problem will be changed into working out the optimized answer as formula (5).

$$P = \{p \mid E_p f_i = E_{\tilde{p}} f, i = 1,2,4 \ \ ,k\},$$
$$p^* = \arg\max_{p \in P} H(p). \quad (5)$$

The optimized answer is showed as formula (6) by using Lagrangian arithmetic.

$$p^*(a \mid b) = \frac{1}{\pi(b)} \exp\left(\sum_{i=1}^{k} \lambda_i f_i(a,b)\right),$$
$$\pi(b) = \sum_{a} \exp\left(\sum_{i=1}^{k} \lambda_i f_i(a,b)\right). \quad (6)$$

λ_i is the parameter of the statistic model. It can be worked out by learning from the training data set. If λ_i is known, the probability distributing function can also be worked out.

The character function is very important for maximal entropy model. It is also an advantage of the maximal model because the flexible configuration of character functions can make full use of all kinds of information to improve the performance. To recognize Chinese names in text, a

Chinese character used as surname or given name, a word (contains one or two Chinese characters) used as given name, and Chinese characters used as non-name are all useful information to construct character functions in the maximal entropy model. When a Chinese character in surname character set does not act as a person's name, it often acts as a word by combining with the Chinese character before or behind it. So a word, which is composed of two or more Chinese characters but not act name, should appear frequently in the training text set. It is useful information to construct character functions. There are sixteen character functions in our maximal entropy model to recognize Chinese names.

The first step of recognizing a person's is selecting a rough name candidate set according to Chinese surname character set and the useful context. Secondly, non-name and name was divided by the maximal entropy model. The third, the probability of one Chinese Character as given name and two Chinese characters as given name are both calculated. If the maximal probability is greater than a threshold δ, the name candidate is considered as a person's name. The training set is the text in People's Daily (a newspaper of China), January 1-20, 1998. The rest text in the newspaper January, 1998 was selected as test set. There are total 1,932,693 Chinese characters and 8,670 person's names in the training set. In test set there are total 1,056,961 Chinese characters and 6909 person's names. With different δ in the maximal entropy model, there were different result data as in Table 1.

Table 1: The result of recognizing persons' names with different δ

	$\delta = 0.3$	$\delta = 0.4$	$\delta = 0.5$	$\delta = 0.6$
Accurate names	6392	6341	6241	5984
Inaccurate names	2903	1588	1492	765
Σ	9295	7929	7733	6749
Precision rate (%)	68.77	79.97	80.71	88.66
Recall rate (%)	92.52	91.78	90.33	86.61

In table 1, the item 'Σ' means the total number of the recognized persons' names by the system. The item 'Recall rate' in table 1 is the percent of accurate names in the actual persons' names (the total number is 6909), which were tagged by human being.

4. Parsing sentences with conceptual model

It is hypothesized that there is only one linguistic conceptual space in human brain, which can be mapped from more than 6,000 languages in the world. This linguistic conceptual space is divided into four layers: conceptual primitives, semantic categories of sentences (SCs), contextual elements and contexts, by Huang (2004). They can be mapped from words, sentences, sentence group and article respectively. In our practice, the linguistic conceptual space is based on a redesigned symbolic system for classifying concepts. Most concepts are distinguished by two kinds of concepts: concrete and abstract. Some concepts are between them. The concrete concepts are material, such as people and the objects in the nature. The abstract concepts are not visible and touchable. The abstract concepts are described from five properties: dynamic(v), static(g), attribute(u), value(z), result(r).

Action-Effect chain is the groundwork of the symbolic system of concepts. The most essential fundamental relationship between two concepts is action. The action brings on an effect, and the effect creates a new action, the new action brings on a new effect. It is a timeless repeating process. It is called Action-Effect chain, by Huang (1998). Based on the Action-Effect chain, a hierarchical symbolic network of primitive concepts is established. The symbols are associated each other through conceptual relations. Any word can be explained by the primitive concepts or their compounding. For example, the mapped conceptual symbol from the word '*think*' is 'v80', and the mapped conceptual symbol from the word '*idea*' is 'r80'. The close relation

between the two words is uncovered by the common symbol '80', as their relation in the fixed phrase '*think an idea*'. So the conceptual relationship between two words can be expressed perspicuously by the symbols.

Starting with the Action-Effect chain, some primitive concepts are attached to the primitive SCs. The most essential fundamental SCs are action, process, transfer, effect, relation, state and judgment. They are the primitives of the semantic category of a sentence (SC). The compound SC is mixed by two primitive SCs. Under a conceptual semantic parsing model, all sentences can be mapped into the SCs. For example, the sentence '*They* ***have also continued****, in the practice of historical activities and by making comparisons,* ***to seek, reveal and develop*** *the truth that guides their advance.*' is a complex sentence with more verbs. The head meaning of the sentence lies on the verbs '***seek, reveal and develop***'. According to the head verbs, the sentence is mapped into the compound SC mixed by action SC and effect SC. A SC is constructed of the main semantic chunks whose number and properties are transcendental in the linguistic conceptual space. The main semantic chunks are divided into two kinds of semantic chunks. One is eigen semantic chunk (EK); the other is generalized object semantic chunk (GBK). EK is corresponded with the characteristic parts of a sentence, like the head verbs '***seek, reveal and develop***' in the example sentence. GBK are corresponded with the objects what the sentence describes, as the subject or object in a sentence, like '*they*' or '*the truth*' in the example sentence.

In order to get the right SC of a sentence, the approach of 'Hypothesis Testing' is adopted. When parsing a sentence, the parser firstly hypothesized the EK according to some special concepts. The second, the SC of a sentence is hypothesized according to the EK. The third, the parser tests the hypothesized SC and its structure according to the words in the sentence and the conceptual knowledge base in the computer. If all semantic chunks in the sentence are corresponded with the transcendental concepts in the conceptual knowledge base, the SC is confirmed. Otherwise, the SC is rejected.

5. Extracting the main elements of an event

After parsing the semantic conceptual structure of a sentence, it is possible to parse the semantic conceptual structure of text. Because text is made up of sentences, parsing the text must be constructed on the results of parsing sentences. Besides the name, the primary information of a person includes gender, age, title, career, etc. Sometimes they appear in the sentences near the person's name. For example, the male role is often surrounded with 'he', 'his', 'boy', 'man' and so on. After recognizing the person's name, the sentences (usually from 3 to 5 sentences) near the name were selected to find out the associated words like 'he'. Then the pattern or structure extracted from the sentences will be matched to the pattern which is abstracted from the training set according name and associated words. If the patterns are matched, the information attached the pattern will be extracted.

When extracting the main elements of an even, it is more difficult because the relationships between the main elements are not a simple co-occurrence model. For an event, the main elements include time, location, people, process, state etc. The main elements are distributed in the text. They can be well organized into a story by human brain. To extract the main elements of an event and their relationships in text by computer, we consider that the essential main elements of an event include: domain, situation, and background.

Domain is the most basic information of the context in sentence group. It describes any activity about human being. The categories of domain is classified according to the first kind of extended concept primitive and the second kind of extended concept primitive in the conceptual symbolic system. Domain also includes instinctive activities, disaster, and state. Situation is the dynamic description of an event. It indicates the semantic relationships between the participants of an event. Background describes the subjective and objective conditions of an event. Background is divided into two. One is the background of the event; the other is the

background of the narrator. The background of the event mainly includes the source of text, language type, date, time, location etc. The background of the narrator mainly includes the age, nationality, standpoint etc.

In order to extract the main elements in text, it is necessary to cut the text into sentences and then analyze the semantic chunks of the sentences (details in section 4). Domain information is implicated in the conceptual concepts of some words. It is convenient to extract the domain information from the semantic chunks which is made up of words. If more than one semantic chunk contain domain information, which domain should be extracted? There is a principle to tackle this problem. Because the station and significance of semantic chunks in a sentence is different, domain information in different semantic chunks possesses different priority. The priority is revealed as the following: Eg>El>C>B or A[1]. Different domains also possess their own priorities. If there are more than one domain to be selected in different sentences, the highest priority domain will be selected. After affirming domain, a SC expression with domain can also be confirmed according to the conceptual knowledge base. Based on the SC expression with domain, the framework units of situation can be constructed. The framework unit of situation must be described as the following: EK name[EK conceptual symbol] | GBKm name[GBK conceptual symbol], the value of m is from 1 to 3. If the same EK or GBK appears in different sentences, the framework units of situation must be combined according to the sequence of appearance. In general, the background is extracted from the supplemental semantic chunks, such as time supplemental semantic chunks, space supplemental semantic chunks, and so on.

By processing the text sentence by sentence, all sentences are mapped into conceptual symbols and the structure of a semantic category. Domain, situation and background are extracted from the conceptual symbols and semantic chunks. To affirm the domain of text, there is also a 'Hypothesis Testing' algorithm. Hypothesizing the domain of text, and then the domain will be tested by the conceptual knowledge base for text. If the conceptual symbols of the sentences are matched to the knowledge base, the domain will be affirmed. Once domain is affirmed, it is easy to construct situation and background according conceptual knowledge base.

6. Conclusion

As an automatic or semi-automatic technology of information processing, information extraction especially web information extraction will be applied into wider fields in our lives. Information extraction is aimed to retrieve the entities and their relationships, the content of special templates and events in text by establishing pattern, domain knowledge, tagged corpus and statistic model. The accuracy of the recognition of named entities is beyond 70%, and the accuracy of the recognition of events seems stuck at 50%-60% (see URL: http://www.cs. cmu.edu/～ref/mlim/chapter3.html). In this paper, an approach of recognizing Chinese name based on the maximal entropy was proposed and the performance is at about 80% precision and 90% recall. In order to extract the primary information of a person and the events from text, the surrounding sentences near the person's name are analyzed into semantic conceptual structures based on a conceptual symbolic system and the conceptual knowledge base. With the 'Hypothesis Testing' algorithm, which is used in analyzing sentences and texts, the main elements of an event are extracted into a semantic framework, which contains domain, situation, and background. We have manually annotated a Chinese corpus, which contains about 350 thousand Chinese characters, for extracting name, gender, title, and events of a person. We are studying how to describe the information and the events in Web Ontology Language (OWL) and improve the performance of extracting persons' information with combining statistic training model and conceptual knowledge.

[1] Eg is a kind of EK, which lies on the top layer of a sentence. El is a kind of EK, which lies on the clause or phrase in a sentence. C, B and A are GBKs in a sentence, C means the Content, B means the oBject, and A means the Actor.

References

Huang ZengYang. 1998. *The Theory of Hierarchical Network of Concept*, Tsinghua University Press, Beijing, China, 6-9.

Huang ZengYang. 2004. *The Basic Theorem and Mathematic and Physical Expression in Lingual Concept Space*, Ocean Press, Beijing, China,. 9-10.

Liu Yue, HongBo Xu and XueQi Chen. 2006. The Progress of Researches on Web Mining and Search. *Proceedings of the Academic Conference for 25th Anniversary of Chinese Information Processing Society of China,* pp.18-33.

URL: *http://www.cs.cmu.edu/ ~ref/mlim/chapter3.html*

Zan HongYing, YuMei Su, Bin Sun and ShiWen Yu. 2003. The WebPages Relevance Research Based on the Shallow Parsing. *Proceedings of the 2003 National Joint Symposium on Computational Linguistics of China*, pp. 501-506.