Person Name Identification in Chinese Documents Using Finite State Automata

Bing Shen Zhongfei (Mark) Zhang*
Computer Science Department
Binghamton University
Binghamton, NY 13902, USA
{bshen, zzhang}@binghamton.edu

Chunfa Yuan
Computer Science Department
Tsinghua University
Beijing, China
cfyuan@tsinghua.edu.cn

Abstract

This research is about automatic identification and extraction of person names in Chinese text documents. Solutions to this problem have immediate and extensive applications in many areas especially in Web Intelligent Agents related applications such as Web search engines, Web data mining, and automatic Web information analysis. We have noted that while finite state automata (FSA) based techniques have been extensively used in NLP and IE in English, they have not yet been extensively used in processing Chinese text, and in particular, to our knowledge, no work has been reported in using FSA in person name identification and extraction. Motivated by this need, we have proposed a person name identification method based on FSA, called NICF. Evaluations show that NICF works very well in terms of identification recall and accuracy, as well as the processing speed, and thus holds a great promise for future applications.

1. Introduction

With the rapid growth in the number of electronic resources published and distributed in Chinese over the Internet, automatic mining Chinese data from the Web or retrieving information across the Internet becomes even more important than ever. Driven by this need, many efforts have been reported in Chinese Natural Language Processing (NLP) and Chinese Information Extraction (IE). Similar to English NLP, it is well known that the current NLP research is unable to provide robust text understanding capabilities [1], resulting in significant focus on developing effective and robust IE techniques and technologies [2]. IE concerns with automatic extraction of key entities in a text document, such as the four W's: who, what, when, and where. An important topic of IE is to automatically identify and extract person names from the text. This project addresses this issue, i.e., person name identification in Chinese text documents.

While there are reported efforts in the literature on this topic, to our knowledge, no one has used finite state automata (FSA) techniques for person name identification in Chinese text documents. Consequently, in this project, we have proposed a specific method of person **N**ame **I**dentification in **C**hinese text documents using **F**SA, called NICF. Clearly, NICF method has immediate and extensive applications in many areas especially in Web Intelligent Agents related applications such as Web search engines, Web data mining, and automatic Web information analysis.

2. Background and Challenges

The finite state automata (FSA) theory is an essential tool used in many areas in computer science, including pattern matching, database, and compiler technology. Due to the speed and the compactness of FSA representations, it provides efficient and convenient tools to represent the linguistic phenomena.

A finite state automaton is a 5-tuple representation $A = (Q, \Sigma, E, S, F)$, where Q is a set of states, Σ is a finite set of the alphabet, $S \in Q$ is a set of initial states, $F \in Q$ is a set of final states, and E is a set of relationships $Q \times (\Sigma \cup \{\varepsilon\}) \times Q$. An FSA represents a collection of states and transitions. It takes an input sequence of symbols such as alphabetical letters, consumes one symbol at a time, and transits one state to another state until it reaches the end of the input or halts on a state. The whole set of the recognized input symbols accepted by an FSA is called the language of the FSA. The language of an FSA is equivalent to regular expressions. The recognition of an input symbol sequence can be solved in a linear time by an FSA. The efficient performance of an FSA is made possible by the following reason: there is a method to transform each automaton into a uniquely equivalent automaton with the least number of states [5]. In other words, a non-deterministic FSA (NFA – with ε transitions or multiple transitions on the same input symbol) can be made deterministic. Furthermore, a deterministic FSA (DFA) can be minimized. Any regular expression recognized by an NFA can be recognized by an equivalent minimal DFA [5]. Such representations have successful applications in various language processing.

While an FSA has powerful representations for regular languages, it is well known that the linguistic expressiveness of the FSA fails to fully recognize natural languages such as English and Chinese.

Nevertheless, recent research shows that an FSA may still be used to achieve some "simple" aspects of NLP, such as IE [5]. Examples of the applications based on these "aspects" include morphology, parsing, speech recognition, and machine translation. While FSA-based techniques have been extensively used in IE in English [5], little work has been reported in using FSA for IE in Chinese. This project attempts to use FSA for person name identification in Chinese text documents.

The problem of person name identification in Chinese documents is difficult and challenging, and hence is still an open problem. In addition to the challenging difficulties existing in the counterpart problem in English, this problem also exhibits the following more difficulties: (1) Chinese names are typically arbitrary; there are no "rules" or patterns to follow. Consequently, a Chinese name may consist of virtually any Chinese words. As a result, hundreds and thousands new names are generated everyday. (2) Different news agencies and different people may translate the same foreign name to different Chinese names; there is no standard to follow. (3) In a Chinese document, person names do not have "boundary tokens" such as the capitalized initial letters for a person name in an English document. This poses an even greater challenge to the problem.

The existing approaches to person name identification in Chinese documents in the literature fall into the following three categories.

(1) Statistical methods [7]. These methods use a large corpus of tagged documents for statistical training, and the identification problem is reduced to a standard classification and labeling problem after the training is done.

(2) Guessing-from-surnames methods [6]. These are the most commonly used methods in person name identification in Chinese documents. There is a set of commonly used surnames in China. Consequently, when we encounter any of these words, we can presume that they might signify an appearance of a person name. The advantage of these methods is that they are simple and robust as long as the encountered words are truly surnames. The disadvantages, however, are that there are false positives as it is not always true that these words are surnames in the text, and they only work for Chinese names but not for foreign names.

(3) Context-based methods [8]. Methods in this category attempt to take advantage of the contextual information typically existing in a Chinese document to "infer" the occurrence of a person name. Examples of the contextual information include the title of a person, certain verbs, and certain adjectives.

3. NICF Method

Based on the above discussion of previous work in person name identification in Chinese documents, we combine a typical guessing-from-surnames method and a typical context-based method to propose a new approach using FSA for person name identification in Chinese documents. Using FSA is due to its efficiency in handling languages. To limit the number of states in an FSA, this method focuses on one small region of a sentence at a time. The method works as follows. First, we segment and tag a sentence. We then segregate the sentence into small regions. These regions and their associated tag sequences are fed into an FSA. The FSA consumes one tag at a time, makes necessary state transitions, or stops when it reaches the final state. After all tags have been consumed, an input sequence that leads to the final state is accepted, which indicates that we have found a person name contained in the current region. If an input sequence that halts on a state is rejected, this indicates that the current region contains no person name.

Any NLP capability needs a tag set to mark words and special strings in documents for further language analysis. The taggers are special symbols that stand for word forms, such as verb, adjective, and noun. A large annotated corpus is a collection of text data with assigned taggers. Consequently, a tag set consists of all the designed taggers to be used for annotating a text corpus. Since there is no standard tag set in Chinese NLP, many research groups resort to designing their own tag sets to meet their particular needs. In NICF, we design nine tags to mark sentences for the person name identification purpose in Chinese documents.

In order to correctly tag a Chinese document, we must first segment the text into words. We use an existing dictionary-based method [4] to segment Chinese sentences into words and other Chinese characters.

Based on this method [4], we use the longest string match algorithm to find words in a sentence. The longest Chinese word in the dictionary we have has eight Chinese characters. Consequently, the longest-word-matching algorithm can recognize words up to eight characters. For each Chinese character encountered in a sentence, the algorithm examines the next one up to the subsequent seven characters to find the longest possible word by looking up the dictionary hash table. The best-matched string becomes an accepted Chinese word. The word segmentation brings down a sentence into small regions, and the tagging is conducted based on these regions.

Since most of native Chinese person names fall into a simple form – a surname followed by a given name, we collect a total of 545 commonly used Chinese surnames. This collection contains the well-known "*Hundreds of Chinese Surnames* (百家姓)" and many common

minority surnames such as Aixinjueluo (爱新觉罗). The typical Chinese given names consist of one or two characters. Although any Chinese characters can be a given name in principle, there are some Chinese characters typically unlikely to be a person's name due to their meanings as well as the community conventions existing in many places in China. Examples of these characters include 死 (death), 尸 (corpse), 了, and 的. Consequently, these characters may be used to help further break down a region into smaller regions.

Common foreign names are usually transliterated from the pronunciations in the original languages. They can be any lengths, which poses a great challenge in identifying these names. Fortunately, there are some Chinese characters that appear particularly often in person name transliterations such as 尔 (er), 姆 (mu), and 斯 (shi). These characters typically do not carry obvious meanings and are normally not considered as common Chinese words. Other characters may commonly be used in other occasions in addition to in the transliterated foreign names. For example, the character 克 (ke) frequently appears as a part of a transliterated person name, but on the other hand, it also often forms many common Chinese words with other characters such as the word 克制 (self-restraint). In this case, contextual information must be used to disambiguate these situations.

Contextual information provides important clues in determining whether a Chinese character is a part of a person's name or not. The NICF tag set is applied to marking surnames, special Chinese characters, contextual words, and other words. These tags are served as the input to the NICF FSA for final identifications.

In order to use FSA to correctly identify person names, we always focus on a phrase between the tag "w" (Chinese words) and the tag "u" (symbols or special characters) as a region consisting of a potential person name and the contextual clues that help identify the person name. We have manually summarized 23 patterns through analyzing the small region text in various Chinese documents. Based on the 23 patterns, we first construct a non-deterministic FSA to describe all possible patterns. The input symbols either follow the path to be consumed one by one entering into the next state, or are rejected or accepted. Based on the NFA constructed, we convert it to the corresponding deterministic FSA with the minimal number of states shown in Fig. 1 using the standard conversion algorithm [5]. Once a transition leads to the state with label *n*, the information of the contextual clues and the length of the string are used to determine whether it can reach the final state *N* and be accepted, or halts at the state n and is rejected. Consequently, an input string will be determined whether it is recognized as a person name or not.

4. Experimental Results

We have implemented NICF as a prototype system, which is also called NICF. NICF is implemented to be able to handle both the simplified Chinese (GB2312 encoding) and the traditional Chinese (Big5 encoding). NICF parses the text data in either of the encoding schemes to segment each sentence into small regions and to tag the rest characters directly; for each region, NICF parses it using the minimal deterministic FSA shown in Fig. 1. Finally, this FSA indicates whether there is a person name or not, and labels the person name if it is identified.

In order to extensively test and evaluate NICF, we have composed two data corpora. One corpus consists of 100 sentences from a public data source (Penn Chinese Treebank [3]). Since this data set is small, we are able to construct detailed ground truth data manually. Thus, this data set allows for detailed testing and parameter tuning. The second corpus consists of different news articles collected from five major Chinese news Websites over a period of half year. This data set allows for real testing of the person name identification capabilities NICF has in the real Web documents, as this is the ultimate goal for this project.

In the evaluations reported in this section, we define the identification recall *R* and the identification accuracy *C* as follows. Let *L* be the number of detected person names, *M* be the total number of ground truth person names in the input corpus, *P* be the number of correctly detected person names out of *L*. Then, $C=P/L$, $R=P/M$.

We first use the Penn Treebank data set for the feasibility test for NICF. The 100 sentences consist of 5,422 Chinese characters, of which there are a total of 45 person names. Based on the definitions given above, NICF achieves 93.33% recall and 95.45% accuracy.

The web news data set is a 20MB text corpus. It contains 290 Web documents and includes 2053 person names. NICF has achieved 98.79% recall and 92.86% accuracy. Note that the recall for the Web news corpus is higher than that for the Penn Chinese Treebank corpus even though the corpus is much larger than that of the latter. This is due to the fact that in many news articles, the transliterated foreign names are accompanied by the original names in the original languages in parentheses. NICF makes use of this pattern as it is designed ultimately for Web applications.

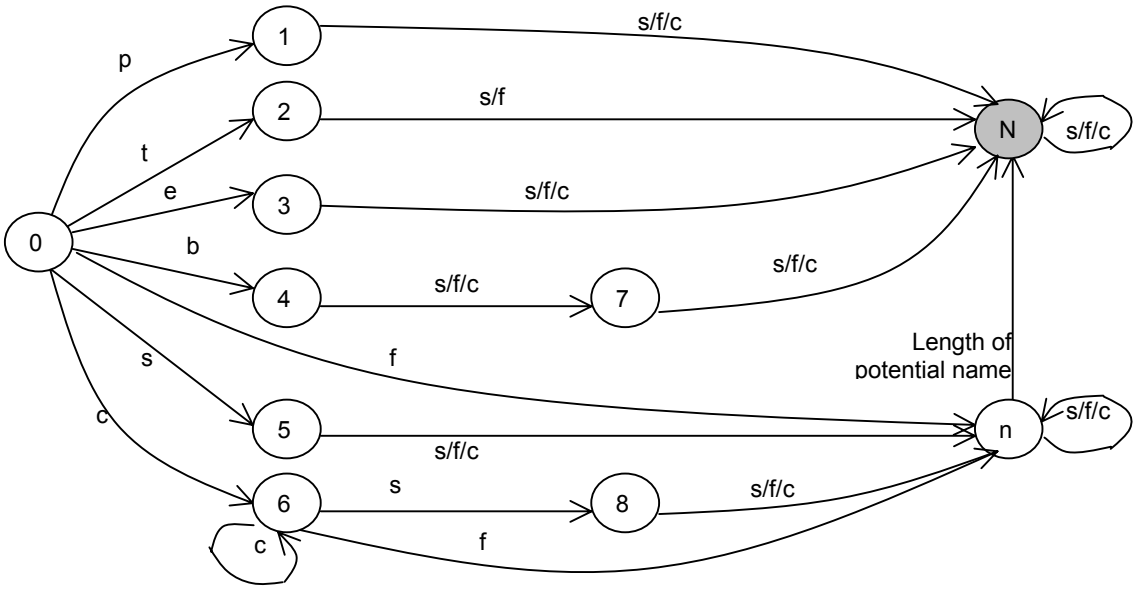

Figure 1: The minimal-state deterministic FSA.

Since the ultimate goal of NICF is for Web applications, the time issue is important. However, we note that of the rich literature of NLP, only a few proposed systems have reported the performance evaluations in terms of processing speeds. In order to see NICF's performance in processing speed, we test it on the Penn Chinese Treebank data corpus. Given the whole data set of the 100 sentences with 5,422 characters and 45 person names in total, NICF takes 960 milliseconds to complete the whole processing, including word segmentation, tagging, and person name identification, under a platform of Pentium 500 CPU with 128MB memory. While the NICF prototype is just for proof of the concept and the code is not optimized yet, this has already shown the great promise NICF holds in future Web applications.

5. Conclusions

This research is about automatic identification and extraction of person names in Chinese text documents. Solutions to this problem have immediate and extensive applications in many areas especially in Web Intelligent Agents related applications such as Web search engines, Web data mining, and automatic Web information analysis. We have noted that while finite state automata (FSA) based techniques have been extensively used in NLP and IE in English, they have not yet been extensively used in processing Chinese text, and in particular, to our knowledge, no work has been reported in using FSA in person name identification and extraction. Motivated by this need, we have proposed a person name identification method based on FSA, called NICF. Taking the advantage of the typically efficient processing speed in FSA, NICF performs fast in processing Chinese text to identify and extract person names, which makes it suitable for future Web related applications. Evaluations also show that NICF works very well in detection recall and accuracy, which further indicates the great promise for future applications.

References

[1] Ralph Grishman, "Information Extraction: Techniques and Challenges", *Fifth Applied Natural Language Processing Conference*, 1997.

[2] MUC-7, *Proc. 7th Machine Understanding Conference,* 1998.

[3] Mary Ellen Okurowski and John Kovarik, "Chinese Little Grove 100 Sentences Treebank". http://umiacs.umd.edu/labs/CLIP/tocampsen.html

[4] Erik Peterson, "A Chinese Named Entity Extraction System", 1996. http://epsilon3.georgetown.edu/~petersee

[5] Michael Sipser, *Introduction to the Theory of Computation*, PWS Publishing Company, 1997.

[6] Jiaheng Zheng and Hongye Tan, "Research of Chinese Person Names Identification Based on Surname", *International Conference on Chinese Computing*, 2001.

[7] Junsheng Zhang, "Chinese Person Name Recognition Based on Corpora", *Chinese Information Journal*, Vol.3 1992.

[8] Yue Zhang and Tiansun Zhnag, "Combination Based for Distinguishing Chinese Name Automatically", *Mini–Micro Systems*, Vol.18 No.10 1997, pp 43–48.

* Responsible for all the correspondences. We thank CLIP at UMIACS for [3] and Erik Peterson for [4] as well as comments from the anonymous reviewers.