PRACTICAL TECHNIQUES OF CHINESE AUTOMATIC WORD SEGMENTATION IN AN APPLIED SYSTEM - CASS

Chun-yu KIT

Department of Applied Linguistics
City Polytechnic of Hong Kong
Fax:7888706 Phone:7888489 E-mail:ALCYKIT@CPHKVX.BITNET

ABSTRACT

It has become obvious in recent years that automatic segmentation of Chinese character string into words is a key issue in Chinese Information Processing. It is regarded as another bottleneck besides Chinese Character Coding to be resolved. Therefore, it is now a common research topic in China (Mainland, Taiwan and Hong Kong) and overseas wherever Chinese computing is undertaken. The main purpose of this paper is to demonstrate general principles of an applied system of Chinese automatic word segmentation through introducing the design and implementation of the CASS system. The overall architecture of the system, segmenting algorithm, dictionary construction, recognition and handling of ambiguous segments of character string will all be discussed.

I. Introduction

Chinese automatic word segmentation, also known as automatic word identification, is a crucial issue to overcome in Chinese Information Processing. It aims at recognizing words, including idioms and terminology, in written Chinese text. Without this essential step, Chinese computing can not proceed to further processing, e.g. parsing, after the Chinese characters have been input into computer, because it is on the word level that the computing of natural language is done.

Since this issue arose, it has been attacked from many necessary sides. At the technical side, there are a number of difficult problems to overcome in implementing segmentation methods and handling ambiguous segments of character string. An ambiguous segment here refers to a part of a character string which can be divided into at least two different groups of words. Techniques for building segmenting system can be classified into two different levels: mechanical segmentation and knowledge processing. The former carries out basic segmentation methods and necessary support, dictionary construction for example. The latter, which employed many practical AI techniques applicable to Chinese computing, checks and corrects mistaken segmentations. Making use of knowledge of language has become more and more important in improving the accuracy of word segmentation.

These two kinds of practical techniques are urgently needed in CIP today. The CASS system is presented in this paper as an illustration of how these techniques may be employed. The abbreviation CASS stands for Chinese Automatic Segmenting System.

II. Objectives and Overall Architecture of CASS

CASS system was implemented as the central part of a large-scale state project in China to build up several Chinese word banks in the domain of economics. Word selection is based, in part, on the usage frequency in a corpus of economic literatures composed of more than ten million Chinese characters. Word segmentation and frequency calculation was assigned to the CASS system.

CASS was carried out using VS FORTRAN on an IBM-4361 (1988, Beijing, China). It provides two execution modes. One, the batch processing, is the usual way to process text data in large amounts. The other is usually used to deal with small texts in interaction with a human user, in order to improve the accuracy. The interactive mode is also used to obtain special experimental data about Chinese word segmentation.

By presetting parameters, the user can also maks choice on use of interaction and knowledge processing. CASS brings out all machanical methods of word segmentation simultaneously. Users can choose any of them. It is probably one of the most distinguishing features of the system and may be regarded as a new progress in word segmentation of Chinese that all segmenting methods (according to ASM(d,a,m) model, see next sections), no matter forward scanning or backward scanning, are carried out with a dictionary arranged in normal sequence.

The dictionary used by CASS is organized within a special structure, known as first-character-index structure. It is stored as a VSAM KKDS file to improve access speed and decrease the memory and disk space used. With its support, mechanical segmentation in CASS reaches an average speed of 200 characters per second in test runs.

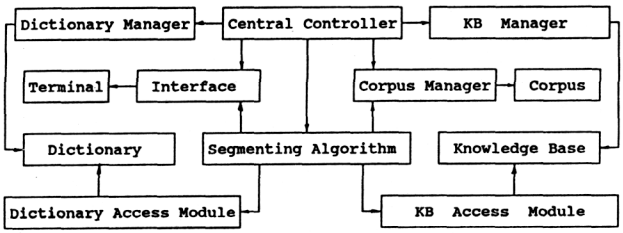

Figure 1

CASS system is composed, shown as Figure 1, of five functional components: 1.The Central Controller that coordinates operations of the whole system; 2.The Segmenting Program, which contains three individual programs: Segmenting Algorithm, Dictionary Access Module and Knowledge Base Access Module; 3. Utility programs, including Dictionary Manager, KB Manager, Corpus Manager and a small interface used in interaction between CASS and its users; 4. Dictionary; 5. Knowledge base for word segmentation.

The processing flow of CASS is shown in figure 2, in which A and B indicate interactive processing and batch processing modes respectively, and 1 2 3 and 4 indicate typical flows for the four different segmenting modes that users can choose one each time. Segmenting method and output options are also selected by users at presetting. If the word frequency calculation program needs

to be activated to update the data in dictionary, a password must be supplied, otherwise the output of segmentation will be displayed on the screen or stored in sequential files, depending on indication of previous settings.

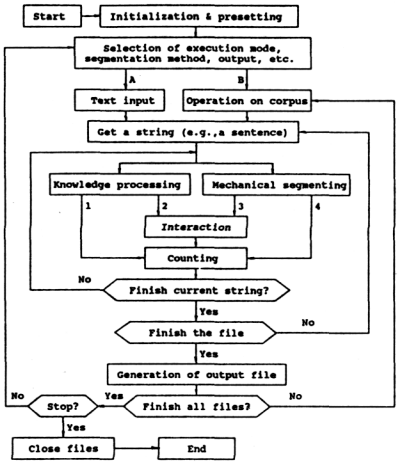

Figure 2

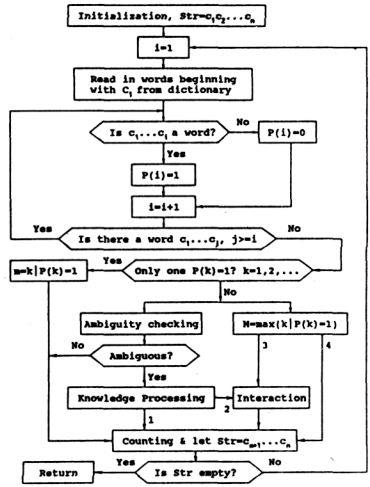

Figure 3

III. Selection & Implementation of Segmentation Methods in CASS

Since the issue of word segmentation arose in CIP, a number of segmentation methods[5] [7] have been proposed. In fact, all methods of Chinese automatic word segmentation are essentially based on character string matching. A method of segmentation is therefore determined, as we found, by three factors: first, the direction of scanning, which may be forward or backward; second, the change of the number of characters in the matched string through each round of matches, i.e., adding or omitting characters one by one to get through the matching process; third, the selection of result from maximum match or minimum match. A round of matches means all matches carried out in identifying a word. While the word is determined, another round of matches begins if the segmentation is still going on. By the way, of maximum match and minimum match, the latter is not really suitable to contemporary Chinese, because almost every character in the written form of this language makes a word or a single-used morpheme.

Consequently, a systematic structure model ASM(d,a,m),[7] [8] in which ASM refers to automatic segmenting methods, has been suggested to classify, and present as well, all methods of Chinese automatic word segmentation. The three parameters are defined as follows:

$d \in D=\{+1,-1\}$, +1 refers to forward scanning,
-1 the backward;
$a \in A=\{+1,-1\}$, +1 means adding charactesr in a round of matches,
-1 omitting;
$m \in M=\{+1,-1\}$, +1 is maximum match,
-1 the minimum.

Therefore, the Maximum Match Method, commonly known as the MM method, which has the properties of forward scanning, character omitting and maximum match, can be represented within this model as MM=ASM(+1,-1,+1). Similarly, the Backward (or Reverse) Maximum Match Method can be described as BMM=ASM(-1,-1,+1). Based on this model, we have brought forward other two new methods ASM(+1,+1,+1) and ASM(-1,+1,+1), which prove to be most beneficial to the whole process of word segmentation.

Among maximum match methods, those with character omission, where characters in the matched string are cut away one by one from one end till a word is found, can not accumulate information about the structure of the output string. So they do not provide data for the follow-up testing and correction of incorrect segmentations. In contrast, methods with character addition, where the number of characters in the matched string is increased one by one, allow information about the structure of the output string to be retained and later used to test and correct mistaken segmentations in knowledge processing.

Today, technical development of Chinese word segmentation focuses more and more on increasing segmentation accuracy, and it depends almost wholly upon knowledge processing. In view of this consideration, we chose ASM(+1,+1,+1) as the kernel method for segmenting algorithm in CASS system. Other methods with maximum match can be carried out using this algorithm. Its flow chart is shown in Figure 3.

IV. Construction and Features of the Dictionary

A machine dictionary for automatic word segmentation is usually organized into the first-character-index structure, which is generally composed of two component parts, the index and the data. The index can be wholely put into memory, but the data area that stores almost all words and regularly-used phrases of a language is so huge that it must be located on a mass storage device, e.g. hard disk, preferably within the format of a direct access file, like a relative record file, to allow a very fast access speed.

Figure 4

Figure 5

The first-character-index structure is shown as Figure 4. Each record in the index consists of a character, i.e. the first-character, and one (or more than one) pointer which indicates the location in the data area of words beginning with that character. As for how many words are put into a record in the data area and how words with the same beginning character but different number of characters, i.e. word length, are arranged, what can be done on these aspects depends absolutely upon the development environment. But two critical factor, i.e. space requirement and access