Segmenting Chinese in Unicode

Thomas EMERSON (湯姆・愛摩森)
Senior Software Engineer
Basis Technology Corporation
tree@basistech.com

Abstract

The automatic segmentation of Chinese text is an ongoing problem in information retrieval (IR) and computational linguistics: "words" in written Chinese are not delimited by spaces so tokenizing (the first phase of many IR tasks) is considerably more difficult than for Western languages.

This paper presents an overview of the segmentation problem, detailing previous research into its solution and introduces Basis Technology's *Chinese Morphological Analyzer* (CMA), a new, general purpose hybrid segmentation system. The CMA is Unicode based, and can handle both Simplified and Traditional Chinese text from a variety of locales, including Mainland China, Taiwan, Hong Kong, and Singapore.

1. Introduction

With few exceptions the first phase of any computational processing of written text is to separate it into its constituent tokens.[1] This is a well defined (and relatively trivial) task for many languages: tokens are separated by white-space or punctuation and generally correspond to the common definition of "word." However, East Asian languages present a particular complication: tokens are not separated by white space in Chinese, Japanese, or Thai. Korean is written with intertoken white space, yet these tokens are rarely individual "words" so further analysis must be performed.

This paper describes a system for segmenting arbitrary free-form Chinese text being developed at Basis Technology Corporation. The segmentation engine is Unicode-based throughout, offering numerous advantages over existing segmenters that are tied to a particular character encoding and locale. The paper is organized as follows: Section 2 defines the terminology and notation used in the remainder of the paper. Section 3 provides an overview of the driving forces behind the development of the segmentation engine, giving various applications of the segmenter. Section 4 describes the architecture and basic operation of the segmentation engine, including the methods used for segmenting and disambiguating text. Section 5 presents the role played by Unicode in the design and implementation of the segmenter, and Section 6 concludes.

This paper makes no assumptions on the reader's knowledge of spoken or written Chinese, or on any particular knowledge of morphology, lexicography, or grammar. Terminology will be defined as needed.

[1]The term 'token' is used instead of 'word' in an attempt to avoid (or at least postpone) the semantic baggage that 'word' carries with it. This is especially necessary when dealing with Chinese, as will be discussed later in the paper.

2. Terminology and Notation

"Chinese" or "Chinese text" as used in this paper refers to modern written Mandarin. This specifically excludes Classical Chinese and written Cantonese, Taiwanese, or other "dialect" of Chinese. When a specific language is meant it will be explicitly stated.

Chinese text is romanized using the *Hanyu Pinyin* system, with diacritical tone marks. For example, 美国人 is romanized as Měiguórén. Unless otherwise specified, Simplified Chinese (SC) forms are used. Written Chinese characters will be referred to interchangeably with either the Chinese word *hanzi* (汉字) or by its English translation "character".

There is no perfect definition of "word" in Chinese: the distinction between 'character' (字 zì) and 'word' (词 cí) was not even made until the early 1950s when language reform began in the People's Republic of China. At least one leading Chinese grammarian has gone so far as to propose the concept of "word", as it is used in Western linguistics, is unnecessary to describe Chinese (Chao 1968). This rather extreme view is no longer in vogue, and much theoretical work is being done to examine the difference between word and phrase in Modern Chinese, see especially Packard (1997) for an excellent collection of papers on current theories. Because of the ambiguity of the term 'word' when discussing Chinese, its use will be avoided as much as possible when discussing the function of the segmenter.

The terms 'dictionary' and 'lexicon' will be used interchangeably in this text, though technically they denote different entities with different purposes and content.

3. Why Segment?

There are many applications where accurately segmented text is a necessity, or at the very least where it is useful to know where the correct word breaks appear in a text. Some of these applications are discussed in this segment.

Information Retrieval

One of the primary problems in information retrieval (IR) research is how best to index a huge amount of textual data such that the precision and relevancy of queries against the index are maximized. The idea is relatively simple: you want to choose the most important, most meaningful words and phrases in the text so that subsequent queries will retrieve these documents. The trick is finding the meaningful units. Much of the research in Chinese segmentation has been done for IR (see for example Nie *et al.* (1996), Wu and Tseng (1995), and Chen (1997).)

Traditionally the most common method of indexing for Chinese (as well as Korean and Japanese) text has been through the use of *n*-grams: the input text is divided into as many overlapping *n* character sequences as possible, and each of these are put into the index. The value for *n* can range between 1 and 4, though most commonly for Chinese bigrams (or 2-grams) are used. For example, the simple sentence

(a)	我	不	是	中国人
	wǒ	bù	shì	Zhōngguórén
	I	not	be	Chinese

I am not Chinese

contains the following bigrams:

(b) 1. 我不
2. 不是
3. 是中
4. 中国
5. 国人

Of these, only two (4 and 5) are actual words.[2] A possible segmentation of this sentence is

(c) 1. 我
2. 不
3. 是
4. 中国人

Of these four segments, only the last has any real semantic value: the rest could be discarded and not indexed. Several studies have compared the performance of bigramming versus segmentation and found that, on average, bigramming out performs segmentation (Kwok 1997). However, to date no in-depth analysis has been performed analyzing the deficiencies in segmentation that lead to the improved performance of the simpler bigram methods.

Text Processing

When using a word processor users often want to navigate or make selections by "word," yet the traditional means of differentiating words in Latin writing systems is absent in Chinese. If the word processor maintains an internally segmented version of the text then it can offer this functionality.

Similarly, an automatic spelling corrector must have some way of finding "words" in the text.

Machine Translation

While statistical machine translation (MT) methods are becoming more advanced, most MT systems perform lexical analysis and generate parse trees on the input. Before either of these tasks can take place, however, the text must be broken into tokens; it must be segmented.

The machine translation task covers not only conversion into different languages, but also translating text between different forms of the same language. Consider the Chinese-to-Chinese conversion problem described in Halpern and Kerman (1999) and Liu (1995): to correctly convert between Simplified Chinese and Traditional Chinese (TC) one needs to account for orthographic and lexical differences within a single written language. These are issues unrelated to the already difficult problem of converting between encodings (e.g., GB2312 and Big5). For

[2] One can argue whether or not 不是 bùshì (not be) is a compound word or a phrase: convincing arguments can be made for either. However, for the purposes of this paper it is considered to be two words.

example, the Simplified character 征 maps to both 徵 and 征 depending on the context. One cannot rely on simple character conversion when dealing with compound words: the SC word 暗里 ànlǐ (secretly) could conceivably be converted into any of the following six TC versions: 暗裡, 暗里, 闇里, 闇裡, 暗裏, and 闇裏 (暗裡 is the correct version). By segmenting the text and identifying the tokens one can achieve a much more accurate conversion between SC and TC.

Text-to-Speech

It is very important to have a segmented text when doing speech synthesis, especially for Chinese. For example, in spoken Mandarin, when two third-tone syllables occur next to each other the first changes to second-tone in speech (e.g., 打倒 dǎdǎo is pronounced as dáodǎo.) Similarly, a small number single syllable words undergo tone modification in speech depending on their context, including 一 yī and 不 bù. In order to correctly replicate the sound of these words the speech generator must have knowledge of where the words in an utterance are.

Linguistic Analysis

There are several areas of linguistics where being able to find word boundaries is essential. Assigning part-of-speech (POS) information (or "tagging") to words in a text is a common task in corpus linguistics, and doing this requires that the text be segmented (though probabilistic taggers have had some success, the majority of POS taggers for Chinese are segmentation driven). Beyond POS tagging, marking syntactic units requires the input text to be segmented and tagged. (Chang 1993)

4. System Overview

4.1 Previous Work

This section first presents a brief overview of previous work on Chinese text segmentation in order to provide a context for the description of Basis' approach. This will be followed with a high-level overview of the CMA and a discussion of some of the interesting issues encountered when implementing our approach.

There are have been three approaches to the word-segmentation problem for Chinese proposed over the last fifteen years: statistical/probabilistic, lexical, and a hybrid of statistical and lexical methods.

Lexical Methods

At its simplest form lexical, or dictionary-based, segmentation utilizes a lexicon containing known words in the language being segmented. The input text is scanned (either left-to-right, right-to-left, or both) and matches are returned. More often than not the longest (or "maximal") match at any given point is returned, on the assumption that the longest sequence of consecutive *hanzi* that appear in the lexicon is probably the correct segmentation at a given point. For example, the sequence 中国人 Zhōngguórēn (Chinese person) can be segmented as:

1. 中|国|人
2. 中国|人
3. 中|国人
4. 中国人

Of which (4) is the correct match. If 中国人 is in the lexicon, then the segmenter will correctly segment this sequence (assuming the *hanzi* following 人 is unambiguous). If the sequence is not in the lexicon, then either (2) or (3) will be returned depending on whether the segmenter is scanning left-to-right or right-to-left. This AB+C vs. A+BC ambiguity is called an "intersecting ambiguity".

Another form of ambiguity is termed "combining ambiguity": two characters may be joined to form a word in some contexts, but not in others. For example, the *hanzi* 的 de is the most common character in Chinese and is predominantly used as a grammatical particle, where it stands alone. However, 的 is also used in multisyllablic words such as 目的 mùdì or 的确 díquè.

It was quickly seen that there are some significant problems with a purely dictionary-based approach:

- Dictionaries, by definition, capture the specific state of a language at a specific time. No dictionary is going to be complete because words are constantly entering the language and no single dictionary can capture the vocabulary across all domains. Mandarin has a productive derivational morphology that allows new words to be formed with ease.
- Some words can be derived through inflectional processes, such as the addition of the pluralizing suffix -们 men to pronouns and person nouns.
- Chinese personal names commonly consist of a monosyllabic surname followed by a bisyllabic given name.[3]
- Transliterated foreign names also present a problem: different Chinese locales will transliterate names differently. Further, the meaning of a transliterated name is almost certainly not derivable from the combined meaning of its component *hanzi.* For example, the name "Kennedy" is transliterated 肯尼迪 kěnnídí in the PRC and 甘迺迪 gānnǎidí in Taiwan.

Aside from these issues, the pragmatic concern of finding a machine-readable dictionary containing a large number of verified entries, with part-of-speech tags has been a major impediment to this method.

One relatively straightforward extension to the pure dictionary based approach is to encode some morphological rules into the segmenter. Examples include the recognition of the pluralizing suffix -们 men, the ordinal number marker 第 dì, and common single-*hanzi* demonstratives and pronouns such as 他 tā (he/him), 她 tā (she / her), 它 tā (it), 这 zhè (this),

[3] This is an over simplification: it is generally the case that a surname consists of a single syllable, (though there are exceptions such as 歐陽 Ōuyáng) and a given name can contain between one and four syllables, though two is by far the most common. (Jones, 1997)

and 那 nà (that), and cardinal numbers. These additions help to a degree, but are still prone to error unless done carefully.

Dictionary-based methods are relatively easy to understand and implement. Lexicons presented in the literature contain between 30000 and 100000 entries, some with POS information.

Statistical Methods

Statistical methods of segmentation are based on the probability of two (or more) characters appearing together. The "language model", or set of probabilities describing a language, is generated from the analysis of a large corpus of (possibly segmented, though this is not necessary) Chinese text, in the domain and locale the segmenter is targeted for.

One of the most common statistical models used for Chinese segmentation is a 1st-order hidden Markov model (HMM). A Markov model is a description of a sequence of random variables evolving over time. The value of a particular variable at time t+1 is dependent only on the state of the variable at time t. In the context of segmentation, the probability that the character at position p + 1 is part of a word depends on the character at position p. The order of a Markov model represents the number of previous states used to predict the next state. A Markov model can be "visible", or "hidden". A Markov model is said to be "visible" if all of the information related to the probability calculations is visible. A "hidden" Markov model is one where there is an assumption that there is an underlying, unseen structure to the probabilities (such as grammatical categories).

The acknowledged problem with Markov models is that they become computationally complex very quickly as their order increases: a 1st-order HMM operates on bigrams: extending this to 4-grams or higher is not tractable.

Another problem with purely statistical techniques is that the language model is heavily dependent on the quality and breadth of the training corpus/corpora, and they must be regularly retrained to handle changes in language use. There are also rather difficult to understand and describe.

However, statistical models do have advantages: the effect of unknown and transliterated words is lessened. They are relatively easy to create and deploy, providing sufficient training data is available.

Recently a number of hybrid methods have been developed which attempt to unify the lexical and statistical methods in an effort to garner the best of both techniques.

4.2 The Basis Solution

The Basis Chinese morphological analyzer is a dictionary-based system that makes use of grammatical (syntactic) and morphological knowledge in conjunction with word-frequency information. The core segmentation engine includes two different algorithms (one of them tunable) that allow the user to tradeoff segmentation accuracy for runtime speed. It is supplied as a C++ library and as a stand-alone tool. The remainder of this section describes each of the CMA's components in detail.

Segmentation Parameters

The Basis CMA follows the Chinese National Standard GB13415, 信息处理用现代汉语分词规范 xìnxī chǔlǐ yòng xiàndài Hànyǔ fēncí guīfàn (Modern Chinese Segmentation Standards for Information Processing). This is a minimal segmentation specification where very few tokens are joined during processing. For example, aspect markers are not attached to their verb. Measure words are not attached to either their preceding number/demonstrative or their following noun or noun phrase.

This behavior is fully customizable, however: this allows the user to customize how certain units are combined most effectively for any given application.

The Lexicon

The CMA has a lexicon consisting of almost 1 million Simplified and Traditional Chinese forms. Each entry includes part-of-speech, frequency, and other information. It includes common words as well as Chinese proper nouns (surnames, given names, countries, cities, rivers, mountains, organizations, etc.) and non-Chinese (mostly Western) proper nouns, including transliterated names. The dictionary is licensed from the CJK Dictionary Publishing Society (CDPS, http://www.cjk.org), a respected and well-known organization. The lexical data has been collected and edited over many years, and continues to be regularly updated in order to capture language change. This lexicon is unprecedented in its coverage and the quality of its data.

The dictionary includes information to aid the segmentation process, such as marking forms that are known to be productive (e.g., 人 rén, which besides its "standard" meaning of 'person' can also be suffixed to country and city names (as well as a number of other categories) to indicate that an individual comes from that location.[4])

Because of the size of our lexicon, many of the deficiencies related to dictionary based approaches are ameliorated. The CDPS data includes words from Mainland China, Taiwan, and Hong Kong as well as locale-specific transliterated names; human editors have proofed the part-of-speech and *Hanyu Pinyin* readings; and the lexicon is under constant scrutiny and improvement as new texts are segmented by the CMA.

Word-Formation Heuristics

The CMA has a collection of rules describing various word-formation (WF) processes in Chinese. For example, these allow the segmenter to recognize a variety of constructed forms:

- Han numeric expressions: ordinal numbers (第三 dì-sān, 3rd), fractions (三分之二 sān fēnzhī èr, 2/3) , and percentages (百分之三十点二五 bǎi fēnzhī sānshí diǎn èr wǔ, 30.25%).
- Productive morpheme affixation, including suffixes like 员 yuán (person), 士 shì (scholar), 家 jiā (expert), and 学 xué (subject of study) and prefixes 女 nǚ (female) and 男 nán (male).

[4] 美国 Měiguó (United States) +人 rén (person) yields 美国人 Měiguórén (American). Compare this with the –er suffix in German ("Amerikaner", "Berliner") or –lainen/-läinen in Finnish ("englantilainen", "helsinkiläinen".) The primary difference is that 人 rén is a free morpheme while –er and –lainen/-läinen are bound morphemes.

- Word-class changing affixes, such as the verb to adjective prefix 可 kě (-able).
- Recognition of certain reduplication patterns, such as AA for nouns (人人 rénrén) and verbs (教教 jiāojiāo), A-yi-A for monosyllabic verbs, ABAB for verbs (睡觉睡觉 shuìjiàoshuìjiào), and others.

Segmentation Core

The segmentation core is responsible for actually segmenting the input text, making use of the lexicon and WF heuristics described above. It limits itself to analyzing the smallest contiguous sequence of *hanzi* available, usually between the current position and the next punctuation character or script change (i.e., a transition from ideographs to Latin characters). For the purposes of segmentation Arabic numerals are considered to be *hanzi* since it is common in Chinese orthography for Arabic numerals to be mixed with Han numerals when writing large numbers, e.g., "20 千" instead of "二十千" for 20,000.

The segmentation core operates in one of two modes: maximal match segmentation or "best" segmentation. The maximal match segmentation attempts to minimize the number of words in a sequence of *hanzi* by finding the longest matches in the dictionary at each point in the input. This is the maximal match algorithm described earlier in this section with the additional application of the word-formation rules found above. The application of these rules can be restricted if necessary, in which case this degenerates into the basic longest-match algorithm.

The more interesting (and computationally intensive) algorithm is the "best" segmentation. The segmenter starts generating all of the possible segmentations for the sentence, pruning nonsensical possibilities dynamically as more segmentations are generated. The weighting function is biased towards segmentations containing the longest high-probability words and takes into account several factors:

- The length (number of *hanzi*) of each segment. The algorithm favors longer words: it is well known that the average word length in contemporary Mandarin is between 2 and 3 *hanzi*.
- The frequency of each segment (based on the value from the lexicon). An infrequent three character word *may* be less likely to be correct than a very frequent two character word followed by a one character word.
- The part-of-speech of the segment and the segments on either side. This allows affixes to be joined to their stems, if constraints allow. For example, the prefix 可 kě can only be joined with to a subclass of verbs.
- Some parts-of-speech rarely, if ever, occur next to each other (e.g., an adverb followed by a noun). Hence it is possible to reject some nonsensical segmentations by these bogus co-occurrences.

In essence the segmenter is serving as a POS tagger and a shallow parser, in an attempt to determine the most accurate segmentation available for a given sequence of *hanzi*.

When the sequence has been segmented, the best remaining possibilities are analyzed and the best one (that with the highest weight) is selected.

One of the most difficult issues in developing the segmenter is handling proper names. While our dictionaries contains tens of thousands of names, the nature of Chinese names means

that it is often difficult, or impossible, to determine when a sequence of characters is being used as a name, and when it is not. This form of ambiguity is difficult to solve without deeper contextual (viz. semantic) information. The approach currently taken by the CMA is to initially ignore the impact of proper names and postpone processing until no other possible segmentations are evident. This is a feasible approach because the set of characters used in names, particularly surnames, is relatively small. Similarly, when dealing with transliterated foreign names, the set of characters used for transliterating English names (for example) is quite small: a sequence of three or more of these characters, in isolation (i.e., not part of an existing compound, or part of a low-frequency compound) is an excellent clue that we're seeing a foreign name.

5. Unicode and Chinese Segmentation

The Basis CMA uses Unicode (UCS-2) throughout its implementation, including the lexicon. Standardizing on Unicode for the internal character representation within CMA enabled us to use a common dictionary lookup algorithm for both SC and TC. All of the WF heuristics are shared between SC and TC implementations. The ability to transparently handle Chinese text in any script and from any Chinese locale greatly simplifies the logic of the segmenter.

6. Conclusion

This paper describes a Chinese word segmentation system currently under development at Basis Technology. An overview of the issues related to segmentation was presented as well as an overview of how the Basis segmenter operates and the important role that Unicode plays in its implementation.

References

CHANG, Chao-Huang and Cheng-Der Chen. **A study on integrating Chinese word segmentation and part-of-speech tagging**. *Communications of COLIPS*, 3(2), December 1993.

CHAO, Yuen Ren. 1968. *A Grammar of Spoken Chinese*. University of California Press, 1968.

CHEN, Aitao, Jianzhang He, and Liangjie Xu. 1997. **Chinese text retrieval without using a dictionary**. *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Philadelphia, Philadelphia, July 1997.

HALPERN, Jack and Jouni Kerman. 1999. **The Pitfalls and Complexities of Chinese to Chinese Conversion.** *Proceedings of the 14th International Unicode Conference*, Cambridge, Massachusetts, March 1999.

JONES, Russell. 1997. *Chinese Names: The Traditions Surrounding the Use of Chinese Surnames and Personal Names*. Pelanduk Publications, Selangor Darul Ehsan, Malaysia, 1997.

KWOK, K.L. 1997. **Comparing representations of Chinese information retrieval**. *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Philadelphia, Philadelphia, July 1997.

NIE, Jian-Yun, Martin Brisebois, and Xiaobo Ren. 1996. **On Chinese text retrieval**. *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Zürich, Switzerland, August 1996.

PACKARD, Jerome L., editor. 1997. *New Approaches to Chinese Word Formation: Morphology, Phonology and the Lexicon in Modern and Ancient Chinese.* Mouton de Gruyter, Berlin, 1997.

Liu, Shing-Huan. 1995. **An automatic translator between Traditional Chinese and Simplified Chinese in Unicode**. *Proceedings of the 7th International Unicode Conference*, San Jose, California, September 1995.

WU, Zimin and Gwyneth Tseng. 1995. **ACTS: An automatic Chinese text segmentation system for full text retrieval**. *Journal of the American Society for Information Science*, 46(2):83 - 96, March 1995.