CHINESE SEGMENTATION PROBLEM

何萬順

Abstract

The correct segmentation of a sentence into words is essential in the computerized analysis of natural languages, and the generation of a sentence involves the proper composition of individually separate words. Unlike most languages written in phonetic alphabets, the Chinese texts do not indicate word boundaries as spacing remains constant between ideographic characters. We demonstrate the problem of segmenting a written Chinese sentence into words in the context of machine translation and present some of the previous partial solutions: pre-editing, 'maximal matching', frequency priority, and other heuristic strategies. Furthermore, we propose the use of domain-specific frequency and a 'no-widow' principle in the implementation of the maximal matching strategy in combination with other heuristic rules as a more thorough scheme for Chinese segmentation.

0. BACKGROUND

The task of segmenting a sentence into individual words arises in the computerized analysis of any natural language, as segmentation is a necessary step in all applications of natural language processing involving parsing or text analysis, such as automatic phonetic transcription of Chinese texts, query systems, and machine translation. This papers takes the current approach of machine translation systems as an example in discussing the Chinese segmentation problem.

The more recent 'indirect' translation approach distinguishes machine translation systems from those of the earlier so-called first generation 'direct' approach. In an indirect translation system, the analysis of the source language and the generation (or synthesis) of the target language are motivated independent of each other, and the interface between analysis and generation is either a transfer or interlingua component (e.g., Her et al 1989). While the transfer component is an ad hoc set of rules for a specific pair of languages, the interlingua is a universal set of rules.

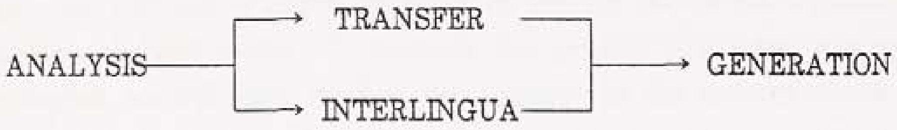


Figure. 1 Indirect Approach of Machine Translation

*作者現為政治大學語言研究所副教授

The correct analysis of a sentence is therefore the first phase towards a correct translation. In turn, the correct segmentation of the sentence into words is a crucial and necessary step towards a correct and efficient analysis; and the generation phase is in essence the proper composition of words in the target language. Similarly, a human translator, before proceeding to translate a sentence or to consult a dictionary, also must recognize what words constitute the sentence, and thus in effect, properly segment the sentence. Since words are stored in the lexicon, word identification in generating Chinese is rather straightforward. This paper, however, explores some of the problems of identifying words in a Chinese analysis system due to the peculiar characteristics of the Chinese texts. We will then discuss some of the previously proposed partial solutions in the literature. We will also suggest alternative strategies and supplementary measures. Ultimately, we propose that the implementation of a combination of some of these strategies is likely to be most successful.

1. SEGMENTATION

One of the very first steps in analysis is the segmentation of the input string. Assuming that the input string is a sentence, the primary task of segmentation is thus to locate words. Chinese, along with a few other languages such as Japanese and Korean, uses characters, or ideograms, as the basic, fundamental writing units. Although every character corresponds to a single syllable, it does not necessarily form a word.[1] In fact, a word very often consists of more than a monosyllabic character and the meaning of this word is usually not that of the composition of the meanings of its internal characters. In the Chinese written text, the spacing between any two characters is constant and the word boundary is never significant and thus never indicated, as shown in the following sample. The only typographical clues of word boundaries are from punctuation, for instance commas, semi-colons, and sentence-final marks like periods and question marks, again as the following short paragraph shows.

> 其實，互動的觀念對形式學派及功能學派都不是一個陌生的概念。前者有模塊式語法的觀念與互動相通；後者有動機競爭的看法與互動相似。

In fact, the current punctuation system of the western style was not adopted in Chinese until earlier this century. In a writing system like English where punctuation is well defined and word boundary is clearly indicated by spacing, the correct segmentation of a sentence is therefore straightforward. To illustrate this problem in Chinese, with an exaggerated example in English, we will show the first two sentences in the abstract of this paper, first with correct spacing between words and proper capitalization in A, and then without in B. Chinese segmentation problem is similar to that of the English text in B without word boundary.

A. The correct segmentation of a sentence into words is essential in
the computerized analysis of natural languages, and the generation

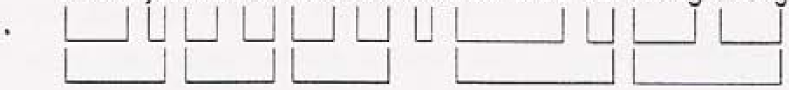
of a sentence involves the proper composition of individually separate words. Unlike most languages written in phonetic alphabets, the Chinese texts do not indicate word boundaries as spacing remains constant between ideographic characters.

B. thecorrectsegmentationofasentenceintowordsisessentialinthecomputeriz edanalysisofnaturallanguages,andthegenerationofasentenceinvolvesthe propercompositionofindividuallyseparatewords.unlikemostlanguageswri tteninphoneticalphabets,thechinesetextsdonotindicatewordboundariesa sspacingremainsconstantbetweenideographiccharacters.

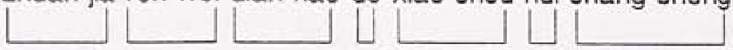
How to identify lexical items is thus the very first problem one encounters in translating Chinese in a machine translation system. A seemingly obvious solution is to build an adequate dictionary containing most Chinese words, and segmentation is achieved by simply matching all the possible segments of the input string with existing words in the dictionary. This would work fairly efficiently for European languages, but not for Chinese. Although an adequate dictionary is necessary for the correct identification of Chinese words, it is not sufficient. The problem is that a string of characters may be segmented into many possible word combinations. The considerable sum of ambiguities during word matching no doubt lead to extreme inefficiency and at times combinatorial explosion. We will look at the following example of a rather simple Chinese sentence. For typographical ease, we will use pinyin romanization to represent Chinese characters. Again, one syllable corresponds to one character, but not necessarily a word. Notice that while 1b shows the correct segmentation of words, the possible manners of segmenting this sentence are many, as shown in 1a.

1 a. ('Experts think the sales of computers will rise. ')

zhuan jia ren wei dian nao de xiao shou hui shang sheng

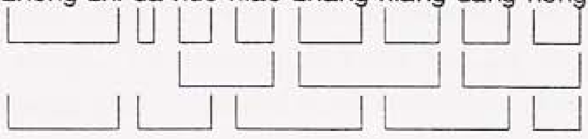


1 b. zhuan jia ren wei dian nao de xiao shou hui shang sheng

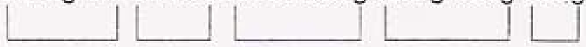


1 c. ('The pricipal of National Chengchi University is quite popular.')

zheng zhi da xue xiao zhang xiang dang hong[2]



1 d. zheng zhi da xue xiao zhang xiang dang hong



Both 1a and 1c are rather ordinary and in no way exaggerate the multiplicity of the different possibilities of segmentation. As a matter of fact, sentences in real texts are likely to be much longer and much more complex than this. It is therefore extremely inefficient to allow the parser to explore all the possible combinations of words matched. In other words, to allow indeterministic segmentation in analyzing Chinese sentences is simply impractical. Therefore, solutions to this problem are essentially designed to make the segmentation process more, if not completely, deterministic. We will now examine some of the solutions that have been suggested previously and discuss their effectiveness.

1.1 PRE-EDITING

The most simple and effective solution is certainly pre-editing. That is, the operator of the machine translation system indicates appropriate word boundaries in the input text, for instance, by spacing (e.g., Li and Ericson 1986). In other words, the pre-editing solution bypasses the segmentation stage by the computer; rather, when the input sentence enters the translation process, segmentation is completed by the human operator already. There is therefore no ambiguity left for the parser to deal with, in the respect of segmentation. This solution is simple and straightforward but not at all satisfactory, for it is one of the most highly pursued goals in machine translation to minimize human intervention. Given the necessity of post-editing in most machine translation systems, to further require the element of human pre-editing is no doubt another very serious compromise. Other feasible solutions without pre-editing rely on the use of heuristic strategies to rid the ambiguities in word matching.

1.2 MAXIMAL MATCHING

Ho (1984) proposed a partial solution that has been widely accepted. His strategy employs a simple heuristic rule of 'maximal matching': scanning the input string from left to right and select the longest possible segment matched as a word to be in the ultimate segmentation (e.g., Ho 1984). To illustrate, given the string <U V W X Y Z>, if every single character is matched as a word and UV and YZ are also match as words, then the string is segmented to be <UV W X YZ> and all the other combinations are ruled out. If, however, UVW is matched as a word, then the segmentation of the string will be <UVW X YZ>. The longest matching of a lexical item is always preserved, in other words. This strategy, though to a great extent it makes correct predictions, does not guarantee a most appropriate segmentation always. For example, it works well for 2a but not for 2b. Note that <and> enclose a sentence, and | marks word boundary. An asterisk indicates an incorrect segmentation.

2 a. (' Everybody is very interested in space. ')
Dui yu tai kong da jia hen you xing qu.
<Dui yu | tai kong | da jia | hen | you | xing qu | . >

2 b. (' People are very interested in space. ')
Dui yu tai kong ren men hen you xing qu.

*<Dui yu | tai kong ren | men | hen | you | xing qu | . >

In 2b, to select the longer segment *taikongren* 'spacemen' rather than *taikong* 'space' the shorter segment in the final segmentation is incorrect. To overcome flaws like this, three solutions, which are not at all mutually exclusive, may be implemented in addition.to or in lieu of the maximal matching strategy. The information of the technical domain of a word is to be implemented within the structure of the dictionary. The 'no widow condition' is a higher strategy which should override the maximal matching principle. And finally 'back-tracking' can occur, when the analysis fails.

1.3 THE 'NO-WIDOW' PRINCIPLE

A 'widow' being an unfound word, the 'no-widow' principle specifies that if a segmentation scheme leaves certain elements in the input unfound, i.e., unmatched with any entry in the dictionary (and also non-generateable by word formation rules, if such rules are implemented), then this segmentation is to be ruled out and an alternative one that leaves no widows should be preferred, unless it is the only possible segmentation. To illustrate, given a string <X Y Z> of Chinese characters and X, XY and YZ as words but not Z, the segmentation of <XY Z>, which leaves Z unmatched as a widow, is to be ruled out and thus the alternative <X YZ> is selected.

When implemented with 'maximal matching', given the completeness of the dictionary and word formation rules, the 'no-widow' condition should receive higher priority (e.g., Chen et al 1986). That is, if the longest matching leaves a widow, i.e. a character unmatched in the dictionary, alternatives are to be preferred. The problematic 2b thus would be rejected and the correct 2a selected.

2 b. (' People are very interested in space. ')
Dui yu tai kong ren men hen you xing qu.
widow
*N

2 a. <Dui yu | tai kong | ren men | hen | you | xing qu | . >

Another problem of maximal matching is that, mathematically speaking, it is possible to have more than one way of segmentation that are of equal length. For instance, given UV, W, VWX, XYZ, and YZ as words, maximal matching does not select between <UV W XYZ > and <U VWX YZ>. As the length of the input string increases, the possibility of such toss-up's also rises. In these cases, certainly the 'no widow condition' may help select a preferred segmentation. Between the above-mentioned two ways of segmentation <UV W XYZ > and <U VWX YZ>, the latter is ruled out because it leaves U an unmatched widow. However, if it so happens that U is a word, then we are again stuck. Certain other types of information may be useful in this case, e.g., the technical domain of the text to be translated and the dependency and adjacency constraints of words in the sentence. We will now explore the idea of technical domains.

1.4 TECHNICAL DOMAINS

A text to be processed often belongs to a particular technical domain, such as computer , law, finance, military, etc. To take advantage of such information, lexical items in the lexicon can be organized in a way that some words receive higher priority than others under certain circumstances. Yang et al (1984) rather briefly presented a scheme of resolving segmentation conflicts that assigns a measure of priority to every possible segment according to length and frequency (Yang et al 1984). (These principles are interrelated with the initial scanning of function words, which we will discuss in the next section concerning homonyms.) They did not implement 'maximal matching' in exactly the same manner we have just described; rather although they recognized that the length of a segment is a good indicator of its priority, a longer segment does not necessarily gets higher priority than a shorter one. Though not stated explicitly, it seems that the most important criterion they used to decide upon priority is frequency. Words that occur more frequently is given higher priority.

However, granted that frequency is certainly extremely useful in this type of weighing scheme to decide upon priority, it is certainly a measure that is bound to fail at times. More importantly, certain words, expressions, and grammatical constructions may be rather infrequent in their overall distribution in a language and yet very commonly used in texts of a particular knowledge domain, e.g., medical reports. Therefore, the specification of the technical domain of a word should be significantly relevant to the frequency of its use.[3]

The dictionary design in some of the machine translation system allows each lexical item to have its specific domain(s) specified (e.g., Her et al 1989 and Chen et al 1989). The dictionary reported in Her et al (1989) recognizes eighty-eight technical domains. Before translating a certain text, the specific domain(s) to which the text belongs is specified. Note that a lexical item may at the same time belong to more than one specific domain, e.g., computer and linguistics. If a text belongs to more than one domain, then the priority has to be specified, e.g. computer >linguistics>general, where 'A>B' means that A has higher priority than B.

This kind of hierarchy in the lexicon is helpful in resolving segmentation conflicts. For example, if the text to be parsed is specified as 'linguistics > general' and the word XYZ is specified in the dictionary as of linguistics and W, XY and Z are words of general use only and WXYZ is a word of the domain of zoology only, the preferred segmentation of <W X Y Z> should be <W XYZ> and not <W XY Z> nor <WXYZ>.

The specification of technical domains on all lexical items also allows the conceptualization that the entire lexicon is composed of various sub-lexicons each of which is devoted to a particular domain. Furthermore, it is easily perceivable that within any sub-lexicon certain criteria, such as frequency and length, can still be used to assign different priority to different lexical items. This type of hierarchical structure of lexicon may fully subsume Yang et al's implementation, but not vice versa.

So far we have presented several measures for selecting a 'best' way of segmentation other than pre-editing. It is evident from our discussion above that none of the strategies is adequate by itself, and that a combination of these strategies is necessary for the ultimate effectiveness in word identification, such as the scheme demonstrated in Chen and Liu (1992). However, there is no existing Chinese sentence analysis system to our knowledge that employs all of them. We recognize that the interaction among these measures can be rather complex and thus further research is needed as to which strategies in conjunction and what priority arrangement among them can provide the most efficient result. Nonetheless, we have suggested that the maximal matching principle be regulated by the 'no widow' condition, and that the frequency priority take into consideration the technical domain of the text.

1.5 BACKTRACKING

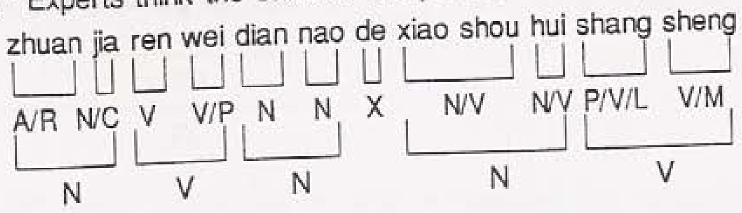
The last resort, if the analysis fails due to inappropriate segmentation, is backtracking, i.e., analyzing the same sentence again with the second most preferred way of segmentation. Backtracking can be repeated until a successful final parse is reached. The danger of this is that it is based on the assumption that all input strings are grammatical and that the rules of the parser are capable of handling all the grammatical input. Given an ungrammatical input or a grammatical string not accounted for by the rules, all possible combinations of segmentation will be tried, quite unnecessarily. It is therefore practical, if not necessary, to impose a limit on the number of times backtracking may apply.

Another possibility is to keep all the unresolved ambiguities around to allow the parser to try all paths. This approach avoids the necessity of repeated backtracking in case of a parse failure. Assuming the relative effectiveness of the strategies mentioned above, unresolved ambiguities should be rather infrequent and thus cause no serious combinatorial growth.

2. HOMONYMS

As a textbook example of isolated languages, Chinese has little inflectional morphology. A word may belong to more than one syntactic category at the same time without any morphological marking or any other kind of variation in form. Again, if we allow the analysis process to explore all the legal paths, we are likely to have combinatorial explosion or at least severely reduce parsing efficiency and speed. Let's look at a couple of examples. We will first repeat sentence 1 here with the possible categories indicated, before segmentation. And then we will show sentence 3 with the appropriate segmentation of words.

1 . (' Experts think the sales of computers will rise. ')


zhuan jia ren wei dian nao de xiao shou hui shang sheng
A/R N/C V V/P N N X N/V N/V P/V/L V/M
N V N N V

3 . (' The manager will explain the work plan ' .)

< Jing li | hui | jie shi | gong zuo | ji hua | . >

N/V X/N N/V V/N V/N

A = Adjective
R = Adverb
N = Noun
V = Verb

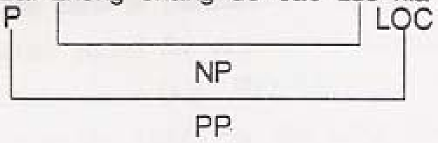
X = Auxiliary
C = Classifier
L = Locative noun
M = Measure noun

Again, both are rather simple sentences and in no way exaggerates the reality that how common it is in Chinese a word is of multiple syntactic categories. Almost all verbs and adjectives may also be nouns; most prepositions act like verbs as well; most adjectives may functions as adverbs post-verbally. To increase efficiency, the analysis process should avoid exploring every possible category of a word.

2.1 HEURISTIC RULES

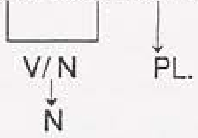
Yang et al (1984) first proposed the use of heuristic knowledge of function words to reduce ambiguities in segmentation and category identification, and Yang (1985) further described such an implementation in an experimental Chinese parsing system. Function words, as the term indicates, are a closed set of lexical items that have specific grammatical functions in a language. In Chinese they may include pronouns, determiners, prepositions, conjunctions, locatives, complementizers, auxiliaries, particles, classifiers, aspect makers, numbers, quantifiers, and perhaps degree adverbs, negation markers, punctuation marks as well. The strategy is to first scan through the input string and find the function words contained within. The idea is to identify these function words as early as possible, thus before loading the dictionary for segmentation, so that the heuristic knowledge is available throughout the entire analysis process including segmentation. Once function words are identified, they may provide useful information on what kind of syntactic constituents are around them. For instance, the locational preposition 'zai' often takes an NP that is marked by a locative marker such as *shang* 'on', *pangbian* 'by', or *waimian* 'outside'. Thus, if *zai* and a later *shang* are identified during the initial scanning, it may be predicted that the segment contained in between is an NP and the whole constituent is a PP.

4. ('Under normal operation computers seldom make mistakes.')
Zai zheng chang de cao zuo xia ji shuan ji hen shao chu cuo.



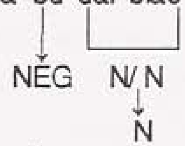
Likewise, if the human plural marker men is identified, the segment (or word) immediately preceding it must be a human noun. In the following sentence, *daibiao* can be both N 'representative' and V 'represent' but the presence of the plural marker *men* indicates that it is N.

5. ('The chairperson asks the representatives to vote.')
Zhu xi qing dai biao men tou piao.



On the other hand, in sentence 6, due to the negation adverb *bu* preceding *daibiao*, we know that here it cannot be an N.

6. ('He doesn't represent the chairperson and vote.')
Ta bu dai biao zhu xi tou piao.



It is easy to see that this kind of information can be conducive to word identification as well. For instance, when the negation adverb *bu* is identified, we know that the following segment, regardless of its length or technical domain, must be either a verb or adjective. In fact, a thorough research and implementation of this kind of constraints can drastically increase the efficiency of the entire analysis process.

2.2 GRAMMATICAL FRAMEWORKS

How and in what grammatical framework linguistic rules and generalizations are formulated of course play an important role in analysis. Better formulated, more general rules would simplify the analysis process, and a suitable grammatical framework felicitates more general and efficient formulations. An inappropriate theory, on the other hand, may obscure the otherwise straightforward process (e.g., Bresnan 1982: 282, Starosta and Nomura 1986, and Her et al 1989). Given the problems of word identification and multiple homonyms, it seems that a lexicalist framework like Lexical-Functional Grammar (LFG), where the construction of any non-terminal category is intrinsically dictated by the idiosyncratic dependency requirements of individual words, is best suited. In addition, as all computational systems must be formal systems, a formalism is thus the necessary device for the formalization of grammatical generalizations within a particular theory. Moreover, it is entirely possible that, for a certain theory, within the class of permissible formalisms; consequently, one particular formalism may stand out as better suited for expressing linguistic analyses or for expediting the computational process than others (e.g., Her 1991 and Shieber 1987). The variant LFG formalism, vLFG, developed in Her (1990) for the LFG theory is a good example, which enhances the lexicalist characteristic of the framework–unlike the conventional LFG formalism, the vLFG formalism allows c-structure categories only when they are functionally well-formed as well (Her 1991). Computationally, vLFG formalism is thus more efficient.

3. Conclusion

The proper segmentation of a Chinese sentence into words is often not simple and yet always imperative to the correct analysis. While a foolproof method in segmenting Chinese sentences into words is yet to be developed, several useful strategies have been found and proven to be relatively successful. We have discussed some of the previously suggested strategies, including pre-editing, maximal matching, frequency priority, backtracking, and the use of heuristic rules. We recommend the specification of technical domains in the lexicon and the 'no-widow' condition in word-identification for the enhancement of the maximal matching strategy. We further propose that to ensure a more thorough and effective method, a combination of some of these strategies could be implemented, together with a suitable grammatical framework, such as Lexical Functional Grammar, and one of its efficient formalism.

NOTES

1. A few exceptions to this one-to-one correspondence do exist, especially in Beijing dialect. However, they do not affect our discussion here. Furthermore, in recent history there has been a tendency for Chinese words to become di-syllabic.
2. An anonymous reviewer provided this example, which the author gratefully acknowledges along with other constructive comments.
3. Moreover, information of the technical domain may be very important in terms of translation selection, e.g. the translation of 'tank' as a container or vehicle, or 'bug' as an insect or a flaw (Her et al, to appear).

REFERENCES

Bresnan, J. 1982 (Ed.). *The Mental Representation of Grammatical Relations*. Cambridge, Mass.: MIT Press.

Chen, C.-G., et al. 1986. A Model for Lexical Analysis and Parsing of Chinese Sentences. Proceedings of 1986 International Conference on Chinese Computing. 33-40.

Chen, K.-J and S.-H. Liu. 1992. Word Identification for Chinese Sentences. COLING 1992.

Chen, S. C., et al. 1989. A Unification-based Approach to Lexicography for Machine Translation. *Journal of Information Science and Engineering*, Vol. 5, No. 4, October 1989, 437-48.

Her, O. 1991. *Grammatical Functions and Verb Subcategorization in Mandarin Chinese.* Taipei: Crane Publishing Co. Also as 1990. Ph.D. Dissertation. University of Hawaii.

Her, O., D. Higinbotham, and J. Pentheroudakis. 1989. An LFG-based English-Chinese Machine Translation System. Proceedings of 1989 International Symposium on Chinese Text Processing 8.3-7. Boca Raton: Florida Atlantic University.

Her, O., D. Higinbotham, and J. Pentheroudakis. (To appear). Lexical and Idiomatic Transfer in Machine Translation: An LFG Approach. In *Research in Humanities Computing*, S. Hockey and N. Ide (Eds.), Oxford: Oxford University Press.

Ho, W.-H. 1984. Automatic Recognition of Chinese Words, Master Thesis, National Taiwan Institute of Technology, Taipei, Taiwan.

Li, M-D. and G. G. Erickson. 1986. Special Parsing Techniques for the Chinese Language. Proceedings of 1986 International Conference on Chinese Computing. 41-45.

Shieber, S. 1987. Separating Linguistic Analyses from Linguistic Theories. In P. Whitelock, et al (Eds.). 1-36.

Starosta, S. and H. Nomura. 1986. Lexicase Parsing: A Lexicon-driven Approach to Syntactic Analysis. in M. Nagao (Ed.). Proceedings of the Eleventh International Conference on Computational Linguistics (COLING '86), Bonn: University of Bonn. 127-132.

Whitelock, P., M. Wood, H. Somers, R. Johnson, and P. Bennet. (Eds.) 1987. *Linguistic Theory and Computer Application.* London: Academic Press.

Yang. Y.-M., et al. 1984. Use of Heuristic Knowledge in Chinese Language Analysis. Proceedings, COLING 1984. 222-5.

Yang, Y.-M. 1985. Studies on an Analysis System for Chinese Sentences. Ph. D. Dissertation. Department of Information Sciences, Kyoto University, Kyoto, Japan.