WORD IDENTIFICATION FOR MANDARIN CHINESE SENTENCES

Keh-Jiann Chen **Shing-Huan Liu**

Institute of Information Science
Academia Sinica

Abstract

Chinese sentences are composed with string of characters without blanks to mark words. However the basic unit for sentence parsing and understanding is word. Therefore the first step of processing Chinese sentences is to identify the words. The difficulties of identifying words include (1) the identification of complex words, such as Determinative-Measure, reduplications, derived words etc., (2) the identification of proper names,(3) resolving the ambiguous segmentations. In this paper, we propose the possible solutions for the above difficulties. We adopt a matching algorithm with 6 different heuristic rules to resolve the ambiguities and achieve an 99.77% of the success rate. The statistical data supports that the maximal matching algorithm is the most effective heuristics.

1. Introduction

Chinese sentences are composed with string of characters without blanks to mark words. However the basic unit for sentence parsing and understanding is word. Therefore the first step of processing Chinese sentences is to identify the words(i.e. segment the character strings of the sentences into word strings).

Most of the current Chinese natural language processing systems include a processor for word identification. Also there are many word segmentation techniques been developed. Usually they use a lexicon with a large set of entries to match input sentences [2,10,12,13,14,21]. It is very often that there are many possible different successful matchings. Therefore the major focus for word identification were on the resolution of ambiguities. However many other important aspects, such as what should be done, in what depth and what are considered to be the correct identifications were totally ignored. High identification rates are claimed to be achieved, but none of them were measured under equal bases. There is no agreement in what extend words are considered to be correctly identified. For instance, compounds occur very often in Chinese text, but none of the existing systems except ours pay much attention to identify them. Proper name is another type of words which cannot be listed exhaustively in the lexicon. Therefore simple matching algorithms can not successfully identify either compounds or proper names. In this paper, we like to raise the problems and the difficulties in identifying words and suggest the possible solutions.

2. Difficulties in the Identification of Words

As we mentioned in the previous chapter, the basic technique to identify the words is by matching algorithms. It requires a well prepared lexicon with sufficient amount of lexical entries which covers all of the Chinese words. However such a large lexicon is never existing nor will be composed, since the set of words is open ended. Not only because the new words will be generated but because there are unlimited number of compounds. Most of the word identification systems deliberately ignore the problem of compounds and leave the problem unsolved until the stage of parsing. We don't agree their view points and believe that different type of compounds should be handled by the different methods at different stages. Some types of the compounds had better to be handled before parsing, for they require different grammatical representations and identification strategies compared with the parsing of phrase structures. On the contrary, if the morphological rules for compounds have the same representation as the phrase structure rules, it is better to be identified at parsing stage. We will discuss this issue in more details in the later sections.

The other problem is that ambiguous segmentations frequently occur during the processing of word matching. It is because that very often a multisyllabic word contains monosyllabic words as its components. We have to try the different strategies to resolve such ambiguities.

Many problems need to be solved, but first of all a lexicon should be composed for the matching algorithm.

2.1 What are the Set of Words

According to Liang's [14,15] definition, word is a smallest,meaningful, and freely used unit. It is the basic processing unit for Chinese natural language processing. Since there is no morphological features as word segmentation marks, we have to adopt such a vague definition of the Chinese word. Liang [15] also propose a word segmentation standard. However some of his view points are debatable and self contradictory. In fact it is almost impossible to define a standard for correct identification. Therefore instead of proposing a standard, we propose a criterion which should be followed by a good word segmentation algorithm. It is that a good segmentation algorithm should be able to produce a result which is suitable and sufficient for the purpose of later processing, such as parsing and understanding.

The set of words is open ended. Therefore the existing lexicons contain lexical entries which vary from 40 to 120 thousands. A large lexicon usually includes many compounds as well as many proper names, for it is hard to distinguish a word and a compound. For systems with a small lexicon, they might not perform worse than systems with a large lexicon, if they incorporate algorithms to identify compounds and proper names. However on the other hand there is no harm to have a large lexicon, once there is a way of handling the ambiguities,since statistically a large lexicon has the better chance to match words as well as producing ambiguous segmentations. Therefore we have the follow principle to collect the word set for the purpose of word segmentation .

Principle for composing a lexicon:

The lexicon should contain as many as possible words. If there is a doubt whether a string of characters is a word or a compound, you can just collect it as an entry.

Currently we have a lexicon of around 90 thousands entries, and keep updating for new words. A lexicon with such a size of course would still leave out many compounds and proper names. We use this lexicon to match the Chinese text, the result of the algorithm is a sequence of words defined in the lexicon. (1) is an instance of result.

(1)
a. jieshuoyuan Jau Shian-Ting yindau tamen
interpreter Jau Shian-Ting guide them
"The interpreter Shian-Ting Jau guided them."

b. jieshuo-yuan-Jau-Shian-Ting-yindau-tamen

However we can see that some of the compounds and proper names are not identified as shown in (1)b. They were segmented into words or characters. Therefore at later stage those pieces of segments should be regrouped into compounds and proper names. We will discuss the issue at next two sections.

2.2 Compounds

There are many different type of compounds in Chinese and should be handled differently [3, 6, 7, 11, 17, 19].

a. determinative-measure compounds (DM)

A determinative-measure compound is composed of one or more determinatives, together with an optional measure.

(2) je san ben
this three CL
"these three"

It is used to determine the reference or the quantity of the noun phrase that co-occurs with it. Despite the fact that both categories of determinatives and measures are closed, the combinations of them are not. However the set of DMs is a regular language which can be expressed by regular expressions and recognized by finite automata [19]. Mo [19] also point out that the structure of DMs are exocentric. They are hardly similar to other phrase structures which are endocentric and context-free and can be analyzed by head driven parsing strategies. Therefore we suggest that the identification of DMs should be done in parallel with the identification of common words. There are 76 rules for DMs which covers almost all of the DMs [19]. For word identification, those rules function as a supplement for the lexicon, which works as if the lexicon contains all of the DMs. We will show the test result in the section 3.3.

b. Reduplications

In Chinese many verbs can be reduplicated to denote an additional meaning of trying the actions gently and relaxedly(3)[7].

(3) tiau tiau wu
jump jump dance
"dance a little"

This kind of morphological construction will not change the argument structure of the verbs, but do change their syntactic behavior. For instance, the reduplications can not cooccur with the adjuncts of postverbal location, aspect marker, duration, and quantifier [17]. In [17], they derived 12 different reduplication rules which cover the reduplication construction of verbs. In addition, there are 3 rules for the reduplication of DMs and 5 rules for A-not-A questions formation.

The identification of the reduplication construction should be done after the words have been identified, since it is better to see the words and then check whether part of the words has been reduplicated. It is a kind of context dependent process, so a separated process other than the process for DMs or Phrase structures, should be incorporated.

c. A-not-A construction

A-not-A constructions are commonly used in Chinese to form a question [3,7]. As we mentioned before, for A-not-A construction, there are 5 different rules for reduplicating part of the verbs and coverbs [17].

(4) fang-bu- fangshin
put not stop-worry
"stop worrying or not"

A-not-A construction is a kind of reduplication constrution. Therefore the technique for the identification of reduplication is also applicable for the identification of the A-not-A construction.

d. Derived Words

A derived word is a compound which composed with a word of stem and a prefix or a suffix [11]. Derivative affixes are very productive in Mandarin Chinese.

(5) difang-shing
place quality
"localityr"

Those affixes usually are bound morphemes. In [11], we collect a set of most frequently occurred affixes and study their morphological behaviors. We found that there are syntactic and semantic restrictions between modifiers and heads. Such a morphological patterns can be represented in terms of Information-based Case Grammar[5], which is also the grammatical formalism adopted for representing Chinese phrase structures in our parsing system[5,11]. Following is an example of representation.

(6) shing
Semantic:meaning: equivalent of "-NESS", "-ITY", for expressing abstract notation
Syntactic: category: Nad
feature: bound; [+N,-V]
constraints: MR: {Vh1, V[+transitive], N} < < *

Since the grammatical representation of derived words is the same as the representation of phrase structures, we suggest that the identification of the derived words are better to be done at the parsing stage. Furthermore, the boundaries of the derived words are syntactically ambiguous. They can not be identified without checking the contextual information.

2.3 Proper Names

Proper names occur very frequently in all kinds of articles. The identification of proper names become one of the most difficult problems in Chinese natural language processing, since we can not exhaustly list all of the proper name in the lexicon. Also there is no morphological nor punctuation makers to denote a proper name. Besides that a proper name may contain a substring of common words. It makes the identification of the proper names even harder. The only clue might be useful in identifying proper names is the occurrences of bound morphemes. Usually each Chinese character is a meaningful unit. Some of them can be used freely as a word. Some are not; they have to be combined with other characters to form a word. Such characters can not be used freely as words, are named bound morphemes. If bound morphemes occur after word matching process, it means that there are derived words or proper names occurred in the text and have not been identified. The semantic classification of morphemes can be utilized to identify the different type of proper names. For instance,in [1], the set of surnames were used as a clue to identify people's names and titles. There is no general solutions so far to handle the different types of proper names. The only suggestion is that mark the proper names before identification process or treat the unknown strings as proper names.

2.4 Ambiguities

For Chinese character strings,they might have many different well formed segmentations, but usually there is only one grammatically and semantically sound segmentation for each sentence (7).

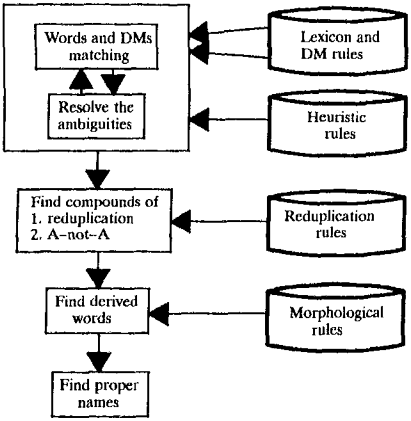
(7) yijing jenglichu jieguo
already arrange-out result
"The result has come out."
yijing-[jengli-chu]-jieguo
yijing-[jeng-lichu]-jieguo

Therefore many algorithms were proposed and heuristic or statistical preference rules were adopted for resolving ambiguities. However none of those rules has been thoroughly tested and provided their success rates. In the next section, we will state our algorithm as well as the heuristic rules and also provides the experiment results in section 3.2 to show the success rate of each individual rule.

3.Word Identification Algorithm

According to the discussion of the chapter 2, the picture of the word identification algorithm should be clearly as follows.
(8)

In fact not all of the above processes were thoroughly studies, but more or less some of them were studied and have successful results [2, 8, 12, 13, 14, 16, 19, 20 ,21]. Our word identification system adopt the above sequence of algorithms, but we defer the second last process of finding derived words until parsing stage and the last process of finding proper names is tempo-



rary ignored for not having a feasible identification algorithm.

3.1 Matching algorithm and disambiguation rules

The first two steps of word identification algorithm are the word matching and disambiguation. These two processes were performed in parallel. Once an ambiguous match occurs, the disambiguation process is invoked immediately. The algorithm reads the input sentences from left to right. Then match the input character string with lexemes as well as DMs rules. If an ambiguous segmentation do occur, then the matching algorithm looks ahead two more words, then apply the disambiguation rules for those three word chunks. For instance,in (9), the first matched word could be 'wan' or 'wancheng'. Then the algorithm will look ahead to take all of the possible combinations of three word chunks, as shown in (10), into consideration.

(9) wanchengjianding baugau
complete authenticate report
"complete the report about authenticating"

(10) wan-cheng-jianding
wancheng-jianding-bau
wancheng-jianding-baugau

The disambiguation algorithm will select the first word of the most plausible chunk as the solution. In this case, it is the word 'wancheng'. The algorithm then proceeds to process the next word until all the input text been processed.

The most powerful and commonly used disambiguation rule is the heuristic of maximal matching [12, 13, 14, 21]. There are a few variations of the sense of maximal matching, but after we have done the experiments with each of different variations, we adopt the following maximal matching rules.

Heuristic rule 1:

The most plausible segmentation is the three word sequence with the maximal length.

This heuristic rules achieves as high as 99.69% accuracy and 93.21% of the ambiguities were resolved by this rule. We will see the detail statistics in the next section. However there are still about 6.79% of ambiguities still can not be resolved by the maximal matching rule. Therefore we adopt the next heuristic rule.

Heuristic rule 2:

Pick the three word chunk which has the smallest standard deviation in the word length. This is equivalent to find the chunk with the minimal value on (L(W1) - Mean) **2 + (L(W2) - Mean) **2 + (L(W3) - Mean)**2 ,where W1,W2,and W3 are three words in a chunk; Mean is the average length of W1,W2,and W3; L(W) denotes the length of the word W.

Heuristic rule 2 simply says that the word length are usually evenly distributed. For instance in (11), the segmentation of (11a) has the value 0, but (11b) has value 2. Therefore according to the heuristic rule number 2, the (11a) will be the selected solution and it is the correct segmentation.

(11) yianjiou shengmin chiyuan
research life origin
"to investigate the origin of life"

a. [yianjiou-shengmin]-chiyuan
b. [yianjiousheng-min]-chiyuan

However it may happen that there are more than two chunks with the same length and variance, we need a further resolution.

Heuristic rule 3:

Pick the chunk with fewer bound morphemes.

Heuristic rule 4:

Pick the chunk with fewer characters in DMs.

That is to say the normal words get higher priority than the bound morphemes and DMs. For instances examples (12,13) were resolved by the rule 3 and 4 respectively. (12a) and (13a) are right choices.

(12) shietiau shang shoushiu jiau mafan
negotiate up procedure more troublesome

"In negotiation, the process is more complicated."

a. shietiau-[shang-shoushiu]-jiau-mafan
b. shietiau-[shangshou-shiu]-jiau-mafan

(13) ta benren
he self
"he himself"

a. ta-benren
b. taben-ren

The heuristic rules 2,3,and 4 only resolve 1.71% of the ambiguities as shown in the table 2 of the next section. After we observe the remaining ambiguities we found that many ambiguities were occurred due to the occurrences of monosyllabic words. For instances, the character string in (14a) can be segmented as (14b) or (14c), but none of the above resolution rules can resolve this case.

(14) ganran de chiuanshr renshu shieshialai
infect DE real number write
down
"write down the precise number of the infected"

ganran-[de-chiuanshr]-renshu-shieshialai
ganran-[dechiuan-shr]-renshu-shieshialai

If we compare the correct segmentations with the incorrect segmentations, we find out that almost all of the monosyllabic word in the correct answer are function words, such as prepositions,conjunctions, as well as a few high frequent adverbs. And that the monosyllabic words in the incorrect segmentations are lower frequency words. The set of such frequently occurred monosyllabic words are shown in appendix 1. We then have the following heuristic rule.

Heuristic rule 5:

Pick the chunk with the high frequently occurred monosyllabic words.

This rule contributes 3.46% of the success of the ambiguity resolution. The remaining unsolved ambiguities are about 1.62% of the total input words. They usually should be resolved by applying real world knowledge or by checking grammatical validity. However it is almost impossible to apply real world knowledge nor to check the grammatical validity at this stage, so applying Markov model is a possible solution[21]. The other solution is much simpler ,i.e. to pick the chunk with the highest accumulated frequency of words[22]. It requires the frequency counts for each words only instead of word bigram or trigram which required by the Markov model.

Heuristic rule 6:

Pick the chunk with the highest probability value. The probability value of the sequence of words ' W1 W2 W3' can be estimated by either

a) Markov model with the bigram approximation
$P = P(W0|W1) * P(W1|W2) * P(W2|W3) * P(W3)$; or

b) Word probability accumulation
$P = P(W1) + P(W2) + P(W3)$

Heuristic rule 6a might not be feasible, since it requires word bigram a matrix of size in the order of 10**10. But heuristic 6b) might not produce a satisfactory resolution. According to our experiment the success rate for 6b) is less than 70%. Therefore the other solution is to retain the ambiguities and resolve at the parsing stage.

3.2 Experiment Results

We designed a word identification system to test the matching algorithm and the above mentioned heuristic rules. The lexicon for our system has about 90 thousands lexical entries plus unlimited amount of the DMs generated from 76 regular expressions. The 90 thousands lexemes form a word tree data structure in order to speed up the word matching [4,10].For the same reasons, DM rules are compiled first to produce a Chomsky Normal Form like parsing table. The parsing table will then be interpreted during the word matching stage[19]. Two sets of test data are randomly selected from a Chinese corpus of 20 million characters. We summarize the testing result in Table 1. Table 2 shows the success rates and applied rates for each heuristic rule.

The recall rate and recognition rate in the above table are defined as follows. Let
N1 = the number of words in the input text,
N2 = the number of words segmented by the system for the input text.
N3 = the number of words were correctly identified.
Then the recall rate is defined to be N3/N1 and the precision rate is N3/N2. The definition of the other statistical result are obviously followed the convention.The above testing algorithm do not include the process of handling derived words. Therefore the above statistics do not count the mistakes occurred due to the existence of derived words, or proper names.

We can see that the maximal matching algorithm is the most effective heuristics. There are 10311 number of ambiguities out of 17494 occurrences of the segmentations. It counts 58.94% of the total segmentations and 93.21% of ambiguities were resolved by this heuristics.

4.Discussions and Concluding Remarks

From the statistical results shown in table 2, it is clear that the maximal matching algorithm is the most useful heuristic method. Most of the mistakes caused by this heuristic are due to the occurrences of the words which are composed by two subwords. Those words are needed to have further investigations. If we want to further improve our system's performance, it seems that employing lexically dependent rules is unavoidable.

The errors caused by the heuristic rule 2 are due to the cases of a three character word followed by a monosyllabic word and which can be divided into two bisyllabic words, for instance (15).

(15) tzai shia san jou
at down three week
"in the following three week"

[tzai-shiasanjou]
[tzaishia-sanjou]

Such mistakes can be avoided by giving the second bisyllabic words a lexically dependent marker which denotes that a low priority is given to this word when the heuristic rule 2 is applied.

Table 1. Testing results

	Sample 1	Sample 2	Total
# of sentences	833	1968	2801
# of characters	8455	20879	29334
# of words	5085	12409	17494
# of words identified by the system	5076	12399	17475
# of correct identifications	5064	12370	17434
recall rate	99.58%	99.69%	99.66%
precision rate	99.76%	99.77%	99.77%

Table 2. The success rates of the heuristic rules

	Sample 1			Sample 2			Total		
	# of identifications	# of errors	success rate	# of identifications	# of errors	success rate	# of identifications	# of errors	success rate
Heuristic Rule 1	2875	13	99.55%	6938	17	99.75%	9813	30	99.69%
Heuristic Rule 2	36	4	88.89%	74	3	95.95%	110	7	93.64%
Heuristic Rule 3	0	0	100%	5	0	100%	5	0	100%
Heuristic Rule 4	18	0	100%	32	1	96.86%	50	1	98.00%
Heuristic Rule 5	104	2	98.08%	238	12	94.96%	342	14	95.90%
Heuristic Rule 6	48	15	68.75%	109	36	66.97%	157	51	67.52%

The heuristic rules #3 and #4 are the most reliable disambiguation rules. However they only contribute 0.53% of the disambiguation processes.

The heuristic rule #5 is useful, but the priority values for each high frequent monosyllabic word has to be carefully rearranged in order to reduce possible mistakes [18].

The heuristic rule #6 needs to be further studied. It will be much more easier to use the bigram or trigram based on grammatical categories instead of the word bigram or the simple accumulation of the word frequencies. It will be the future study.

About the identification of the proper names, it requires a further investigation on the results of the proper names after segmentation algorithm is applied.

5.References

[1] J. S. Chang, "A Multiple-Corpus Approach to Identi cation of Chinese Surname-Names," Proc. of Natural Language Processing Pacific Rim Symposium, Singapore, 1991

[2] J. S. Chang, J. I. Chang and S. D. Chen, "A Method of Constraint Satisfaction and Statistical Optimization for Chinese Word Segmentation," Proc. of the 1991 R. O. C Computational Linguistics Conference, Taiwan, 1991

[3] Y. R. Chao, A Grammar of Spoken Chinese, University of California Press, California, 1968

[4] K. J. Chen, C. J. Chen and L. J. Lee, "Analysis and Research in Chinese Sentences — Segmentation and Construction," Technical Report, TR-86-004, Nankang, Academia Sinica, 1986

[5] K. J. Chen and C. R. Huang, " Information-based Case Grammar," COLING - 90, Vol 2, p.54 - p.59

[6] K. J. Chen et al, "Compounds and Parsing in Mandarin Chinese," Proc. of National Computer Symposium, 1987

[7] G. Y. Chen, " A-not-A Questions in Chinese," manuscript, CKIP group, Academia Sinica, Taipei, 1991

[8] C. K. Fan and W. H. Tsai, " Automatic Word Identification in Chinese Sentences by the Relaxation Technique," Computer Processing of Chinese and Oriental Languages, Vol.4, No.1, November 1988

[9] R. Garside, G. Leech and G. Sampson, " The Computational Analysis of English -- a Corpus-based Approach," Longman Group UK Limited, 1987

[10] W. H. Ho, " Automatic Recognition of Chinese Words," Master Thesis, National Taiwan Institute of Technology, Taipei, Taiwan, 1983

[11] W. M. Hong, C. R. Huang, T. Z. Tang and K. J. Chen," The Morphological Rules of Chinese Derivative Words," To be presented at the 1991 International Conference on Teaching Chinese as a Second Language, December, 1991, Taipei

[12] C. Y. Jie, Y. Liu and N. Y. Liang, " On Methods of Chinese Automatic Segmentation," Journal of Chinese Information Processing, Vol.3, No.1, 1989

[13] B. I. Li, S. Lien, C. F. Sun and M. S. Sun, "A Maximal Matching Automatic Chinese Word Segmentation Algorithm Using Corpus Tagging for Ambiguity Resolution," Proc. of the 1991 R. O. C Computational Linguistics Conference, Taiwan, 1991

[14] N. Y. Liang, "Automatic Chinese Text Word Segmentation System — CDWS", Journal of Chinese Information Processing, Vol.1, No.2, 1987

[15] N. Y. Liang, "Contemporary Chinese Language Word Segmentation Standard Used for Information Processing," 1989, a draft proposal

[16] N. Y. Liang, "The Knowledge of Chinese Words Segmentation," Journal of Chinese Information Processing, Vol.4, No.2, 1990

[17] M. L. Lin, " The Grammatical and Semantic Properties of Reduplications," manuscript, CKIP group, Academia Sinica, 1991

[18] I. M. Liu, C. Z. Chang and S. C. Wang, "Frequency Count of Frequently Used Chinese Words," Taipei, Taiwan, Lucky Book Co., 1975

[19] R. P. Mo, Y. J. Yang, K. J. Chen and C. R. Huang, " Determinative-Measure Compounds in Mandarin Chinese : Their Formation Rules and Parser Implementation," Proc. of the 1991 R.O.C Computational Linguistics Conference, Taiwan, 1991

[20] R. Sproat and C. Shih, "A Statistical Method for Finding Word Boundaries in Chinese Text," Computer Processing of Chinese and Oriental Languages, Vol.4, No.4, March 1990

[21] C. L. Yeh and H. J. Lee, "Rule-based Word Identification for Mandarin Chinese Sentences -- A Unification Approach," Computer Processing of Chinese and Oriental Languages, Vol.5, No.2, March 1991

Appendix

為的不於向已是有去最就到會在了當應較自並很大中小至受被把將有拿幫替像待朝害閉對要以能可般似往
與除從連同和用藉仗憑假因據視隨趁等趁趕距離打逮隔迄上趕靠挨下前後內裡外旁間東西南北底末初邊方
側端來時該須得別還跟休甭勿免依按照比如擬