Ambiguity Resolution in Chinese Word Segmentation*

Sun Maosong
Tsinghua University and City University of Hong Kong
E-mail: dcsjpf@tsinghua.edu.cn, ctmssun@cityu.edu.hk

Benjamin K T'sou
City University of Hong Kong
E-mail: rlbtsou@cpccux0.cityu.edu.hk

ABSTRACT

A new method for Chinese word segmentation named Conditional F&BMM (Forward and Backward Maximal Matching) which incorporates both bigram statistics (i.e., mutual information and difference of t-test between Chinese characters) and linguistic rules for ambiguity resolution is proposed in this paper. The key characteristics of this model are the use of: (i) statistics which can be automatically derived from any raw corpus, (ii) a rule base for disambiguation with consistency and controlled size to be built up in a systematic way.

1. Ambiguities in Chinese Word Segmentation

In Chinese, there are no delimiters, such as spacing in English, to explicitly indicate boundaries between words. Chinese word segmentation, viewed as the first step in any Chinese information processing system, has been intensively studied by the Chinese language computing community in the last decade. Unfortunately, a satisfactory segmentation procedure remains elusive.

Ambiguity is one of two main obstacles to progress in Chinese word segmentation research (another is *Unknown Word*) [1]. We have to face two kinds of fundamental ambiguities:

- Type I — *Overlapping Ambiguity* (*OA*)

(1) 在这个项目的研究上

At least two possible segmentations which overlap in position exist for the sequence "项目的": "项目\的" and "项\目的", if we simply match it with a Chinese dictionary.

- Type II — *Categorial Ambiguity* (*CA*)

(2)a. 廖晖一行于昨日抵达香港

b. 一行白鹭上青天

Note the sequence "一行" in (2): the constituents should be combined as a single word in (a), but they should be separated in (b) because of the productivity of expressions such as 两行, 三行 ...

Basic methods for dealing with segmentation ambiguities so far can be either rule-based [2,3] or statistics-based [4,5]. The former employs the conventional *maximal matching strategy*, either forward or backward (referred here as *FMM* and *BMM* respectively), or both , as a detector of ambiguities, and applies relevant rules from the rule base to solve them. The problems in this approach are (i) the constructions of *OAs* in texts to be processed are nearly unpredictable, resulting in unwieldy complexity in rule base establishment and maintenance, and (ii) it always fails in finding *CAs*. The latter gives segmentation possibilities exhaustively by dictionary lookup as a

* This research is supported in part by the Youth Science Foundation of Tsinghua University, Beijing, and by the Language Information Sciences Research Centre, City University of Hong Kong

candidate space to the input sentence first, then makes use of statistics (typically word frequencies) to prune the candidates for the solution. But what is confusing in this approach is the acquisition of statistical data. It can be estimated that to obtain more reliable word frequencies, a pre-segmented corpus with at least 20M words is needed because there exist about 50,000 words in a medium sized Chinese dictionary. This is as yet an impractical mammoth task.

Focusing on the problems identified above, a new method named *Conditional F&BMM* which incorporates both *statistics* and *rules* into ambiguity resolution is proposed in this paper. The task of *F&BMM* is to find *OA*s whereas an additional "*Conditional*" mechanism is responsible for finding *CA*s. A Chinese character bigram model, having been trained automatically from a very large corpus, is utilized as the means to disambiguate *OA*s. The disambiguation of *CA*s, however, will be carried out by some general rules invoked by internal constructions of *CA*s accordingly.

2. Ambiguity Detection & Resolution

2.1. Why F&BMM Can Work?

The idea of segmenting Chinese sentences with *F&BMM* simultaneously is not something new. But till now no quantitative analysis appears available in the literature to validate the feasibility of this strategy. We test this by applying *F&BMM* to a set of randomly selected Chinese texts with 55230 characters, or 3680 sentences. The following is observed from the result:

case 1: The output of *FMM* and *BMM* are different, but both are incorrect

(3) 将以新的姿态出现在世界东方

FMM: 将\以\新\的\姿态\出现\在世\界\东方 ==> incorrect

BMM: 将\以\新\的\姿态\出\现在\世界\东方 ==> incorrect

Total in number: 2 sentences (0.054%)

case 2: The output of *FMM* and *BMM* are different, but only one is correct

(4) 使节约粮食进一步形成风气

FMM: 使节\约\粮食\进一步\形成\风气 ==> incorrect

BMM: 使\节约\粮食\进一步\形成\风气 ==> correct

Total in number: 340 sentences (9.24%)

case 3: The output of *FMM* and *BMM* are identical but incorrect

(5) 反映了一个人的精神面貌

FMM&BMM: 反映\了\一\个人\的\精神\面貌 ==> incorrect

Total in number: 15 sentences (0.41%)

case 4: The output of *FMM* and *BMM* are identical and correct

(6) 美国加州大学的科学家发现 ...

FMM&BMM: 美国\加州\大学\的\科学家\发现 ... ==> correct

Total in number: 3323 sentences (90.30%)

Case 2 and case 3 are examples of *OAs* and *CAs* respectively. Such ambiguities can be successfully captured by *F&BMM* (the *Conditional* mechanism needs to be elaborated for case 3, see section 2.3) . Case 1, on the other hand, is really a "blind area" for *F&BMM*, but as can be seen, it only accounts for a very small part of the whole texts (0.054%), and may be omitted in practical considerations. This suggests that *F&BMM* is quite appropriate for the purpose of Chinese word segmentation.

2.2. Resolving Ambiguities of Type I

Mutual information and *t-test*, two important concepts in information theory and statistics, have been exploited to measure the degree of association between two words in an English corpus[6]. We adopt these measures almost completely here, with one major modification: the variables in two relevant formulae are no longer *words* but *Chinese characters*.

Definition 1 Given a Chinese character string 'xy', the *mutual information* between characters x and y is defined as:

$$I(x:y) = \log_2 \frac{p(x,y)}{p(x)p(y)}$$

where p(x,y) is the co-occurence probability of x and y, and p(x), p(y) are the independent probabilities of x and y respectively.

Definition 2 Given a Chinese character string 'xyz', the *t-test* of the character y relevant to characters x and z is defined as:

$$t_{x,z}(y) = \frac{p(z|y) - p(y|x)}{\sqrt{\text{var}(p(z|y)) + \text{var}(p(y|x))}}$$

where p(y|x) is the conditional probability of y given x, and p(z|y), of z given y, and var(p(y|x)), var(p(z|y)) are variances of p(y|x) and of p(z|y) respectively.

Note that $t_{x,z}(y)$ is attached to a character y, whereas $I(x:y)$ is attached to the location between two adjacent characters x and y. This inconsistency may cause some inconvenience if we try to incorporate both. We initially introduce a new measure instead:

Definition 3 Given a Chinese character string 'vxyw', the *difference of t-test* between characters x and y is defined as:

$$\Delta t(x:y) = t_{v,y}(x) - t_{x,w}(y)$$

Like $I(x:y)$ now, $\Delta t(x:y)$ is also allocated to the location between characters x and y. Furthermore, the domain that $\Delta t(x:y)$ covers is 4 characters, larger than that of $t_{x,z}(y)$ (3 characters). These two features make $\Delta t(x:y)$ more suitable for our needs.

Both $I(x:y)$ and $\Delta t(x:y)$ serve as estimates to measure the combinatorial propensity between characters x and y: the higher the value, the stronger the combinatorial propensity. For example, in the following sentence:

(7)	标	志	着	我	国	公	民
$I(x:y)$	6.2	5.2	2.6	4.4	0.0	2.2	1.6
$\Delta t(x:y)$	25.5	2.8	-86.7	164.2	-106.6	36.8	-52.2
	法	律	意	识	的	增	强
$I(x:y)$	7.3	1.3	7.2	0.2	-0.3	5.2	
$\Delta t(x:y)$	73.4	-55.7	20.8	8.2	-39.5	60.1	

I(法:律) is very high (7.3), so "法律" should be absolutely combined; I(的:增) is very low (-0.3),thus "的增" should be separated. Similarly, "我国" should be combined but "国公" separated, because of their respective Δt values: Δt (我:国) = 164.2, Δt (国:公) = -106.6.

In some complicated cases however, $I(x:y)$ and $\Delta t(x:y)$ should be both used judiciously and jointly, allowing them to complement each other. Sorting character pairs collected in (7) according to $I(x:y)$ in descendent order, we get:

法律	意识	标志	增强	志着	我国	着我	公民	民法	律意	识的	国公	的增
7.3	7.2	6.2	5.2	5.2	4.4	2.6	2.2	1.6	1.3	0.2	0.0	-0.3
A				B				C				

For character pairs in area A or area C, $I(x:y)$ allows judgment to be properly made by itself. But for those in area B, it seems not so clear: *I*(志:着)(5.2) is equivalent to *I*(增:强)(5.2), and further, *I*(我:国)(4.4) is lower than *I*(志:着) , *I*(公:民)(2.2) is also lower than *I*(志:着), even lower than *I*(着:我)(2.6), why should "增强", "我国", "公民" be grouped together while "志着", "着我" must be separated? There will be no explanation based solely on $I(x:y)$.

We sort character pairs by their $\Delta t(x:y)$ this time:

我国	法律	增强	公民	标志	意识	识的	志着	的增	民法	律意	着我	国公
164.2	73.4	60.1	36.8	25.5	20.8	8.2	2.8	-39.5	-52.2	-55.7	-86.7	-106.6
A				B				C				

Consider the above character pairs again. The distinction between Δt (增:强)(60.1) and Δt (志:着)(2.8) is quite significant given *I*(增:强) equals *I*(志:着), indicating that "增强" should be combined and "志着" separated. The situation becomes even more apparent if "我国" and "着我" are included: "我国" goes forward to the head of the queue from area B to area A (Δt (我:国) = 164.2), and "着我" moves almost to the tail from area B to area C (Δt (着:我) = -86.7) . That is really what we are expecting!

The algorithm integrating these two kinds of statistics for *OA* resolution is sketched below:

step 1. for an input sentence *S*, segment it with *FMM* and *BMM* respectively, deriving two candidate segmentations *SEG1* and *SEG2*.

step 2. if *SEG1* and *SEG2* are identical, then output it as the result; otherwise

step 3. if number of words in *SEG1* and *SEG2* are different, then output one which will result in less words ; otherwise

step 4. for any two successive words *FRAG1* in *SEG1* and their counterparts *FRAG2* in *SEG2* composed of also two words, if the location between two words in *FRAG1* denoted as *pt1* is different from that in *FRAG2* denoted as *pt2*, then

step 4.1. if $I(pt2) - I(pt1) \geqslant \alpha$ then output *FRAG1* for the corresponding part of *S*;
if $I(pt1) - I(pt2) \geqslant \alpha$ then output *FRAG2* for the corresponding part of *S*;

step 4.2. if $|I(pt1) - I(pt2)| < \alpha$ then
if $\Delta t(\text{pt2}) - \Delta t(\text{pt1}) \geqslant \beta$ then output *FRAG1* for the corresponding part of *S*;
if $\Delta t(\text{pt1}) - \Delta t(\text{pt2}) \geqslant \beta$ then output *FRAG2* for the corresponding part of *S*;

step 4.3. if $|I(pt1) - I(pt2)| < \alpha$ and $|\Delta t(\text{pt1}) - \Delta t(\text{pt2})| < \beta$ then
if $I(pt2) - I(pt1) \geqslant 0$ then output *FRAG1* for the corresponding part of *S*;
otherwise output *FRAG2* for the corresponding part of *S*.

/* The constants α , β are to be determined by experimentation. */

Applied to 340 *OA*s found by *FMM&BMM* mentioned in section 2.1, 292 (i.e., 85.9%) are correctly segmented. There is a 35.9% improvement over the average accuracy of 50.0% when *FMM* and *BMM* are applied separately.

Some of the experimental results are listed below to illustrate the performance of this algorithm:

(4) 使节约粮食进一步形成风气

FMM: 使节 \ 约 (\粮食\进一步\形成\风气)

$I(x:y)$ 8.8

$\Delta t(x:y)$ 51.6

BMM: 使 \ 节约 (\粮食\进一步\形成\风气) ==> output

$I(x:y)$ 4.6

$\Delta t(x:y)$ -29.0

This situation can be processed by the algorithm easily (step 4.1).

(8) 伪造者缺乏书画的基本功底

FMM: (伪造者\缺乏\书画\的\) 基本功 \ 底

$I(x:y)$ 5.5

$\Delta t(x:y)$ 4.9

BMM: (伪造者\缺乏\书画\的\) 基本 \ 功底 ==> output

$I(x:y)$ 5.7

$\Delta t(x:y)$ -81.7

In this case, the difference between I(功:底) and I(本:功) is not large enough, so the difference of Δt (功:底) and Δt (本:功) must be taken into account. *OA* in (8) can be successfully solved now (step4.2). It is obvious that if a decision is made by comparing the values of I(功:底) and of I(本:功) only, the results will be misleading.

(9) 她的艺德也成了大家称道的话题

FMM: (她\的\艺德\也\成\了\大家\称道\) 的话 \ 题

$I(x:y)$ 5.3

$\Delta t(x:y)$ 13.5

BMM: (她\的\艺德\也\成\了\大家\称道\) 的 \ 话题 ==> output

$I(x:y)$ 4.3

$\Delta t(x:y)$ 15.0

Neither the difference between I(话:题) and I(的:话) nor that between Δt (话:题) and Δt (的:话) would be significant this time. We shall have to return to the starting point, that is, making use of mutual information solely (step4.3).

A key characteristic of this algorithm is that the relevant statistics (Chinese character bigram) can be trained automatically from any raw corpus of unlimited size without any requirement of human monitoring. In our experience of training such statistics with a news corpus of 20M Chinese characters, we have found it very easy and fast to adapt parameters of the model to any new application domain of word segmenter under the condition that the electronic corpus is available accordingly.

2.3. Resolving Ambiguities of Type II

Definition 4 A word w is said to have feature '*ck*' iff w can be further split into n words $w_1,\ldots,w_i,\ldots,w_n$ (n $\geqslant$ 2), and there exists at least one Chinese sentence involving w in which $w_1,\ldots,w_i,\ldots,w_n$ can be syntactically and semantically justified if w is simply viewed as a character string.

By definition, words with '*ck*' feature can be pre-defined and enumerated in the dictionary . In our dictionary for word segmentation, 4171 out of 61039 entries are marked with '*ck*'. "一行" and "个人" , as illustrated in (2) and (5), are typical examples.

In section 1 of this paper, we regard our approach as *Conditional F&BMM*. *Conditional* means that if a word with '*ck*' is encountered in the process of segmentation, *F&BMM* will no longer be an absolute ruler; the possibility of splitting them into words with smaller units needs to be reserved for further accounting.

In our experience, words of this sort can be divided into some single groups according to their shared parts of speech and a sequence of smaller words with different parts of speech. For instance, words like "一行", "两头", "两手", "两厢", "三角", "三轮", "三副", "四则", "五行", "八股", "八方", "十位" can be regarded as a single group, because they are both 'noun' and 'numeral +

classifier' construction.

Then, a generalization rule for dealing with this *CA* group can be set up:

#RULE [*w* | POS: *NOUN*, WC: *NUMERAL+CLASSIFIER*]

if *w* can take the position of 'numeral + classifier' in the sentence to be segmented

then segment it into two words as 'numeral' and 'classifier'; otherwise output *w* as a 'noun'

One point deserving attention is that the construction of *CA*s can be predicted in general as contrasted with the unpredictability of *OA*s. In other words, a rule base for the disambiguation of *CA*s with consistency and controlled size can be conducted in a more systematic way.

3. Conclusion

The design principle for the presented model should benefit some practical applications of word segmentation in, for example, post-processing of OCR, speech recognition and synthesis systems, in which a trade off among accuracy, cost and speed must be made. The limitation of the model is mainly caused by the fact that the character bigram only approximates the unigram model of word in nature. In addition, ambiguities associated with unknown words are not taken into account at all. Improvement on these will be our future work.

Reference

[1] M.S. Sun and B.K. T'sou, "Theoretical Aspects of Chinese Word Segmentation", *Applied Linguistics* (Beijing), No.4, 1995

[2] T.S. Yao, G.P. Zhang and Y.M. Wu, "A Rule-based Chinese Word Segmentation System", *Journal of Chinese Information Processing*, Vol.4, No.1, 1990

[3] K.K. He and H. Xu, "Design Principles of an Expert System for Automatic Word Segmentation of Written Chinese Texts", *Journal of Chinese Information Processing*, Vol.5, No.2,1991

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

[5] J.S. Zhang, Z.D. Chen and S.D. Chen, "A Method of Word Identification for Chinese by Constraint Satisfaction and Statistical Optimization Techniques", *Proc. of ROCLING-IV*, Kenting, Taiwan, 1991

[6] K.W. Church, P. Hanks and D. Hindle, "Using Statistics in Lexical Analysis", in *Lexical Acquisition: Exploiting On-line Resources to Build a Lexicon*, edited by U. Zernik, Hillsdale, N.J.:Erlbaum, 1991