Chinese Word Segmentation without Using Lexicon and Hand-crafted Training Data

Sun Maosong, Shen Dayang*, Benjamin K Tsou**
State Key Laboratory of Intelligent Technology and Systems, Tsinghua University, Beijing, China
Email: lkc-dcs@mail.tsinghua.edu.cn
* Computer Science Institute, Shantou University, Guangdong, China
** Language Information Sciences Research Centre, City University of Hong Kong, Hong Kong

Abstract

Chinese word segmentation is the first step in any Chinese NLP system. This paper presents a new algorithm for segmenting Chinese texts without making use of any lexicon and hand-crafted linguistic resource. The statistical data required by the algorithm, that is, mutual information and the difference of t-score between characters, is derived automatically from raw Chinese corpora. The preliminary experiment shows that the segmentation accuracy of our algorithm is acceptable. We hope the gaining of this approach will be beneficial to improving the performance(especially in ability to cope with unknown words and ability to adapt to various domains) of the existing segmenters, though the algorithm itself can also be utilized as a stand-alone segmenter in some NLP applications.

1. Introduction

Any Chinese word is composed of either single or multiple characters. Chinese texts are explicitly concatenations of characters, words are not delimited by spaces as that in English. Chinese word segmentation is therefore the first step for any Chinese information processing system[1].

Almost all methods for Chinese word segmentation developed so far, both statistical and rule-based, exploited two kinds of important resources, i.e., lexicon and hand-crafted linguistic resources(manually segmented and tagged corpus, knowledge for unknown words, and linguistic rules)[1,2,3,5,6,8,9,10]. Lexicon is usually used as the means for finding segmentation candidates for input sentences, while linguistic resources for solving segmentation ambiguities. Preparation of these resources (well-defined lexicon, widely accepted tag set, consistent annotated corpus etc.) is very hard due to particularity of Chinese, and time consuming. Furthermore, even the lexicon is large enough, and the corpus annotated is balanced and huge in size, the word segmenter will still face the problem of data incompleteness, sparseness and bias as it is utilized in different domains.

An important issue in designing Chinese segmenters is thus how to reduce the effort of human supervision as much as possible. Palmer(1997) conducted a Chinese segmenter which merely made use of a manually segmented corpus(without referring to any lexicon). A transformation-based algorithm was then explored to learn segmentation rules automatically from the segmented corpus. Sproat and Shih(1993) further proposed a method using neither lexicon nor segmented corpus: for input texts, simply grouping character pairs with high value of mutual information into words. Although this strategy is very simple and has many limitations(e.g., it can only treat bi-character words) , the characteristic of it is that it is fully automatic -- the mutual information between characters can be trained from raw Chinese corpus directly.

Following the line of Sproat and Shih, here we present a new algorithm for segmenting Chinese texts which depends upon neither lexicon nor any hand-crafted resource. All data necessary for our system is derived from the raw corpus. The system may be viewed as a stand-alone segmenter in some applications (preliminary experiments show that its

This work was supported in part by the National Natural Science Foundation of China under grant No. 69433010.

accuracy is acceptable); nevertheless, our main purpose is to study how and how well the work can be done by machine at the extreme conditions, say, without any assistance of human. We believe the performance of the existing Chinese segmenters, that is, the ability to deal with segmentation ambiguities and unknown words as well as the ability to adapt to new domains, will be improved in some degree if the gaining of this approach is incorporated into systems properly.

2. Principle

2.1. Mutual information and difference of t-score between characters

Mutual information and *t-score*, two important concepts in information theory and statistics, have been exploited to measure the degree of association between two words in an English corpus[4]. 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(or equally, the *mutual information* of the *location* between x and y) is defined as:

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

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

As claimed by Church(1991), the larger the mutual information between x and y, the higher the possibility of x and y being combined together. For example:

经济合作将是对目前世界经济趋势的
一个适当回答。 (1)

The distribution of *mi(x:y)* for sentence (1) is illustrated in Fig. 1(where "◆" denotes x, y should be combined and "■" be separated in terms of human judgment. This convention will be effective throughout the paper). The correct segmentation for (1) can be achieved when we decide that every location between x and y in the sentence be treated as 'combined' or 'separated' accordingly if its *mi* value is greater than or below a threshold(suppose the threshold is 3.0 for this example):

经济 | 合作 | 将 | 是 |
economy cooperation will be
对 | 目前 | 世界 | 经济 | 趋势 |
for current world economy trend
的 | 一个 | 适当 | 回答
of an appropriate answer
(Economic cooperation will be an appropriate answer to the trend of economics in current world.)

It is evident that x and y are to be strongly combined together if *mi(x:y)*>>0 and to be separated if *mi(x:y)*<<0. But if *mi(x:y)* ≈ 0, the association of x and y becomes uncertain.

Observe the *mi* distribution for sentence (2) in Fig. 2:

法国网球公开赛今天在巴黎西郊拉开
战幕。 (2)

In the region of 2.0 ⩽ *mi* < 4.0, there exist some confusions: we have *mi(球:公)=mi(公:开)*> *mi(开:赛)*, *mi(天:在)*> *mi(赛:今)*> *mi(法:国)*, and *mi(在:巴)*> *mi(拉:开)*, however, "球:公""天:在""赛:今""在:巴"should be separated and "公:开""开:赛""法:国""拉:开" be combined by human judgment -- the power of *mi* is somewhat weak in

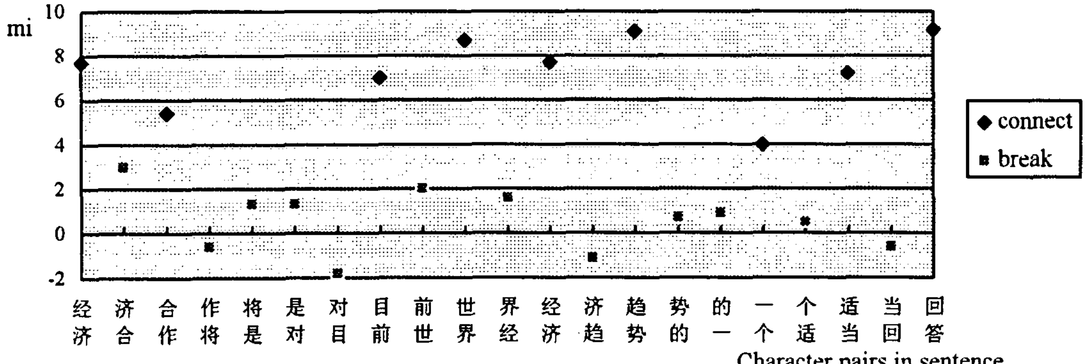

Fig.1 The distribution of mi(sentence 1)

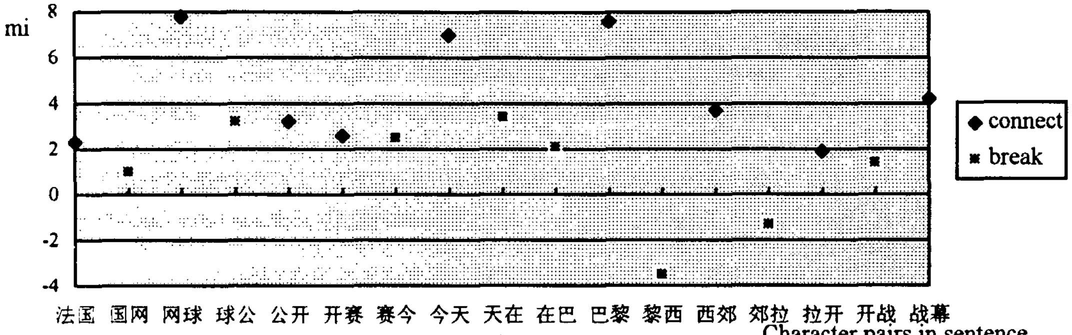

Fig.2 The distribution of mi(sentence 2)

the 'intermediate' range of its value. To solve this problem, we need to seek other ways additionally.

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

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

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.

Also as pointed out by Church(1991), $ts_{x,z}(y)$ indicates the binding tendency of y in the context of x and z:

if p(z|y) > p(y|x), or $ts_{x,z}(y) > 0$
 then y tends to be bound with z rather than with x
if p(y|x) > p(z|y), or $ts_{x,z}(y) < 0$
 then y tends to be bound with x rather than with z

A distinct feature of *ts* is that it is context-dependent (a relative measure), along with certain degree of flexibility to the context, whereas *mi* is context-independent (an absolute measure). Its drawback is it attaches to a character rather than to the location between two adjacent characters. This may cause some inconvenience if we want to unify it with *mi*. We initially introduce a new measure *dts* instead of *ts*:

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

$$dts(x{:}y) = ts_{v,y}(x) - ts_{x,w}(y)$$

Now $dts(x{:}y)$ is allocated to the location between x and y, just like $mi(x{:}y)$. And the context of $dts(x{:}y)$ becomes 4 characters, 1 character larger than that of $ts_{x,z}(y)$.

The value of $dts(x{:}y)$ reflects the competition results among four adjacent characters v, x, y and w:

(1) $ts_{v,y}(x) > 0$ $ts_{x,w}(y) < 0$

(x tends to combine with y, and y tends to combine with x) ==> $dts(x{:}y) > 0$

ⓥ ⓧ ⇄ ⓨ ⓦ

In this case, x and y attract each other. The location between x and y should be bound.

(2) $ts_{v,y}(x) < 0$ $ts_{x,w}(y) > 0$

(x tends to combine with v, and y tends to combine with w) ==> $dts(x{:}y) < 0$

ⓥ ⟵ ⓧ ⓨ ⟶ ⓦ

In this case, x and y repel each other. The location between x and y should be separated.

(3a) $ts_{v,y}(x) > 0$ $ts_{x,w}(y) > 0$

(x tends to combine with y, whereas y tends to combine with w)

ⓥ ⓧ ⟶ ⓨ ⟶ ⓦ

(3b) $ts_{v,y}(x) < 0$ $ts_{x,w}(y) < 0$

(x tends to combine with v, whereas y tends to combine with x)

ⓥ ⟵ ⓧ ⟵ ⓨ ⓦ

In cases of (3a) and (3b), the status of the location between x and y is determined by the competition of $ts_{v,y}(x)$ and $ts_{x,w}(y)$:

if $dts(x{:}y) > 0$ then it tends to be bound
if $dts(x{:}y) < 0$ then it tends to be separated

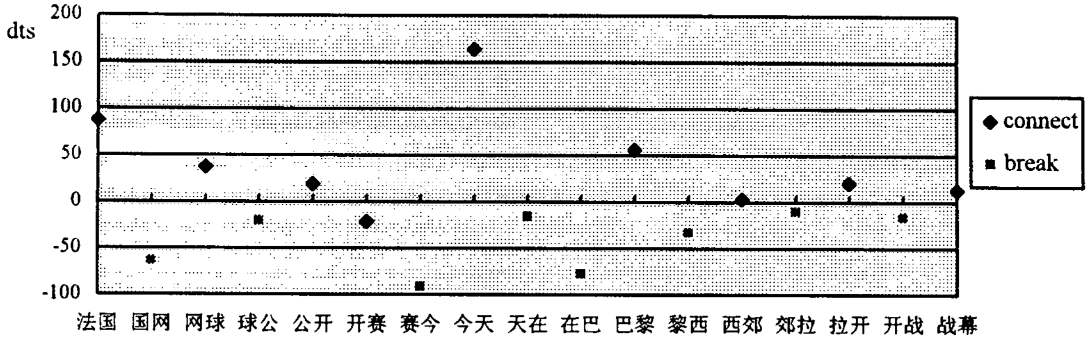

Fig.3 The distribution of dts(sentence 2)

The general rule governing *dts* is similar as that governing *mi*: the higher the difference of t-score between x and y, the stronger the combination strength between them, and vice versa. But the role of *dts* is somewhat different from that of *mi*: it is capable of complementing the 'blind area' of *mi* on some occasions.

Consider sentence (2) again. The distribution of *dts* for it is shown in Fig. 3. Return to the character pairs whose *mi* values fall into the region of $2.0 \leqslant mi < 4.0$ in Fig. 2, compare their *dts* values accordingly: $dts(\text{公}:\text{开}) > dts(\text{球}:\text{公}) > dts(\text{开}:\text{赛})$, $dts(\text{法}:\text{国}) > dts(\text{天}:\text{在}) > dts(\text{赛}:\text{今})$, and $dts(\text{拉}:\text{开}) > dts(\text{在}:\text{巴})$ -- the conclusion drawn from these comparisons is very close to the human judgment.

2.2. Local maximum and local minimum of *dts*

Most of the character pairs in sentence (2) have got satisfactory explanations by their *mi* and *dts* so far. "球:公" "开:赛" are two of few exceptions. We have $mi(\text{球}:\text{公}) > mi(\text{开}:\text{赛})$ and $dts(\text{球}:\text{公}) > dts(\text{开}:\text{赛})$, however, the human judgment is the former should be separated and the latter be bound. Aiming at this, we further proposed two new concepts, that is, local maximum and local minimum of *dts*.

Definition 4 Given 'vxyw' a Chinese character string, $dts(x{:}y)$ is said to be a local maximum if $dts(x{:}y) > dts(v{:}x)$ and $dts(x{:}y) > dts(y{:}w)$. And, the height of the local maximum $dts(x{:}y)$ is defined as:

$$h(dts(x{:}y)) = \min\{\, dts(x{:}y) - dts(v{:}x),\ dts(x{:}y) - dts(y{:}w) \,\}$$

Definition 5 Given 'vxyw' a Chinese character string, $dts(x{:}y)$ is said to be a local minimum if $dts(x{:}y) < dts(v{:}x)$ and $dts(x{:}y) < dts(y{:}w)$. And, the depth of the local minimum $dts(x{:}y)$ is defined as:

$$d(dts(x{:}y)) = \min\{\, dts(v{:}x) - dts(x{:}y),\ dts(y{:}w) - dts(x{:}y) \,\}$$

Two basic hypotheses can be easily made as the consequence of context-dependability of *dts*(note: *mi* has not such property):

Hypothesis 1 x and y tends to be bound if $dts(x{:}y)$ is a local maximum, regardless of the value of $dts(x{:}y)$(even it is low).

Hypothesis 2 x and y tends to be separated if $dts(x{:}y)$ is a local minimum, regardless of the value of $dts(x{:}y)$(even it is high).

In Fig. 3, $dts(\text{球}:\text{公})$ is a local minimum whereas $dts(\text{开}:\text{赛})$ isn't. At least we can say that "球:公" is likely to be separated, as suggested by the hypothesis 2(though we still can say nothing more about "开:赛").

2.3. The second local maximum and the second local minimum of *dts*

We continue to define other four related concepts:

Definition 6 Suppose 'vxyzw' is a Chinese character string, and $dts(x{:}y)$ is a local maximum. Then $dts(y{:}z)$ is said to be the *right* second local maximum of $dts(x{:}y)$ if $dts(y{:}z) > dts(v{:}x)$ *and* $dts(y{:}z) > dts(z{:}w)$.And, the distance between the local maximum and the second local maximum is defined as:

$$dis(locmax, y{:}z) = dts(x{:}y) - dts(y{:}z)$$

Definition 7 Suppose 'vxyzw' is a Chinese character string, and *dts(x:y)* is a local minimum. Then *dts(y:z)* is said to be the *right* second local minimum of *dts(x:y)* if $dts(y:z) < dts(v:x)$ and $dts(y:z) < dts(z:w)$. And, the distance between the local minimum and the second local minimum is defined as:

$dis(locmin, y:z) = dts(y:z) - dts(x:y)$

The *left* second local maximum and the *left* second local minimum of *dts(x:y)* can be defined similarly.

Refer to Fig. 3. By definition, *dts(开:赛)* is the *left* second local minimum of *dts(赛:今)*, and *dts(天:在)* is the *right* second local maximum of *dts(今:天)* meanwhile the *left* second local minimum of *dts(在:巴)*.

These four measures are designed to deal with two common construction types in Chinese word formation: "2 characters + 1 character" and "1 character + 2 characters". We will skip the discussion about this due to the limited volume of the paper.

3. Algorithm

The basic idea is to try to integrate all of the measures introduced in section 2 together into an algorithm, making best use of the advantages and bypassing the disadvantages of them under different conditions.

Given an input sentence S, let

μ_{mi} : the mean of *mi* of all locations in S;

σ_{mi} : the standard deviation of *mi* of all locations in S;

μ_{dts} : the mean of *dts* of all locations in S; (in fact, $\mu_{dts} \equiv 0$)

σ_{dts} : the standard deviation of *dts* of all locations in S

we divide the distribution graphs of *mi* and *dts* of S into several regions(4 regions for each graph) by μ_{mi}, σ_{mi}, μ_{dts} and σ_{dts}:

region A $dts(x:y) > \sigma_{dts}$

region B $0 < dts(x:y) \leqslant \sigma_{dts}$

region C $-\sigma_{dts} < dts(x:y) \leqslant 0$

region D $dts(x:y) \leqslant -\sigma_{dts}$;

region a $mi(x:y) > \mu_{mi} + \sigma_{mi}$

region b $\mu_{mi} < mi(x:y) \leqslant \mu_{mi} + \sigma_{mi}$

region c $\mu_{mi} - \sigma_{mi} < mi(x:y) \leqslant \mu_{mi}$

region d $mi(x:y) \leqslant \mu_{mi} - \sigma_{mi}$

The algorithm scans the input sentence S from left to right two times:

```
The first round for S
For any location (x:y) in S, do
1. in cases that <dts(x:y), mi(x:y)> falls into:
1.1  Aa or Ba or Ca or Da or Ab
       mark (x:y) 'bound'
1.2  Ad or Bd or Cd or Dd or Dc
       mark (x:y) 'separated'
1.3  Ac or Cb
       if dts(x:y) is local maximum then
           if $h(dts(x:y)) > \delta_1$
             then mark (x:y) 'bound' else '?'
       if dts(x:y) is local minimum then
           if $d(dts(x:y)) > \xi_2$
             then mark (x:y) 'separated' else '?'
1.4  Bc or Db
       if dts(x:y) is local maximum then
           if $h(dts(x:y)) > \delta_2$
             then mark (x:y) 'bound' else '?'
       if dts(x:y) is local minimum then
           if $d(dts(x:y)) > \xi_1$
             then mark (x:y) 'separated' else '?'
1.5  Cc
       if (dts(x:y) is local maximum) and
          ($h(dts(x:y)) > \delta_3$)
           then mark (x:y) 'bound' else '?'
       if dts(x:y) is local minimum
           then mark (x:y) 'separated' else '?'
1.6  Bb
       if dts(x:y) is local maximum
           then mark (x:y) 'bound' else '?'
       if (dts(x:y) is local minimum) and
          ($d(dts(x:y)) > \xi_3$)
           then mark (x:y) 'separated' else '?'
2. For (x:y) unmarked so far, mark it as '?'
   except that:
   if dts(x:y) is the second local maximum
     then if $dis(locmax, x:y) <$
                   $0.5 \times lrmin(loc,x:y)$
   /* Refer to the notations in definition 6&7.
   $lrmin(loc,x:y) = \min\{|dts(x:y) - dts(v:x)|,$
                   $|dts(x:y) - dts(z:w)|\}$ */
```

```
            then mark (x:y) '←' if
                (x:y) is the right second local max
                      or '→' if
                (x:y) is the left second local max
    if $dts(x:y)$ is the second local minimum
        then if $dis(locmin, x:y) < 0.5 \times lrmin(loc,x:y)$
            then mark (x:y) '←' if
                (x:y) is the right second local min
                      or '→' if
                (x:y) is the left second local min
The second round for S
if (x:y) is marked '?'
    then if $mi(x:y) \geqslant \theta$
        then mark (x:y) 'bound' else 'separated'
if (x:y) is marked '←'
    then the status of (x:y) follows that of
        the adjacent location on the left side
if (x:y) is marked '→'
    then the status of (x:y) follows that of
        the adjacent location on the right side
```

(The constants δ_1, δ_2, δ_3, ξ_1, ξ_2, ξ_3 are determined by experiments, satisfying:

$$\delta_1 < \delta_2 < \delta_3 ; \quad \xi_1 < \xi_2 < \xi_3$$

and $\theta = 2.5$)

Generally speaking, the lower the *<dts(x:y), mi(x:y)>* in distribution graphs, the more restrictive the constraints. Take 'bound' operation as example: there is not any additional condition in case 1.1; in case 1.6 however, the existence of a local maximum is needed; in case 1.3, a requirement for the height of local maximum is added; in case 1.4, the height required becomes even higher; and in case 1.5, which is the worst case for 'bound' operation, the height must be high enough.

Case 2 says if the second local maximum is pretty near to the local maximum corresponded, then its status ('bound' or 'separated') would be likely to be consistent with that of the local maximum. So does the second local minimum.

Finally, for locations marked '?' with which we have no more means to cope, simply make decisions by the value of *mi*(we set it to 2.5, same as that in the system of Sproat and Shih(1993)).

Recall sentence (2). The character pair "天:在" is regarded as 'separated' successfully by following "在:巴"(local minimum) with the rule in case 2 although its *mi* value is rather high(3.4). "开:赛" is marked '?' in the first round and treated properly by θ in the second round.

The algorithm outputs the correct segmentation for sentence (2) at last:

法国 | 网球 | 公开赛 | 今天 |
France tennis competition today
在 | 巴黎 | 西郊 |
in Paris the western suburbs
拉开 | 战幕
open curtain

(The Tennis Competition of France opened in the western suburbs of Paris today.)

Note that there exist two ambiguous fragments "公开赛"("公开 | 赛" or "公开赛") and "拉开战幕"("拉开 | 战幕" or "拉 | 开战 | 幕"), as well as two proper nouns "France" and "Paris" in sentence (2).

4. Experimental results

We select 100 Chinese sentences, consisting of 1588 characters(or 1587 locations between character pairs) randomly as testing texts. The statistical data required by calculating *mi* and *dts*, in fact it is character bigram, is automatically derived from a news corpus of about 20M Chinese characters. The testing texts and training corpus are mutually excluded.

Out of 1587 locations in the testing texts, 1456 are correctly marked by our algorithm.

We define the accuracy of segmentation as:

$$\frac{\text{\# of locations being correctly marked}}{\text{\# of locations in texts}}$$

Then, the accuracy for testing texts is 1456/1587 = 91.75%.

The distribution of local maximum, local minimum and other types of *dts* value(involving the second local maximum and the second local minimum) of the testing texts over *<dts, mi>* regions is summarized in Fig. 4 (Fig. 5 is the same distribution in percentage representation). This would be helpful for readers to understand our algorithm.

Future work includes: (1) enlarging the size of experiments; (2) refining the algorithm by studying the relationship between *mi* and *dts* in depth; and (3) integrating it as a module with the existing Chinese segmenters so as to improve their performance (especially in ability to cope with unknown words and ability to adapt to various domains). -- it is indeed the ultimate goal of our research here.

5. Acknowledgments

This work benefited a lot from discussions with Professor Huang Changning of Tsinghua University, Bejing, China. We would also like to thank anonymous COLING-ACL'98 reviewers for their helpful comments.

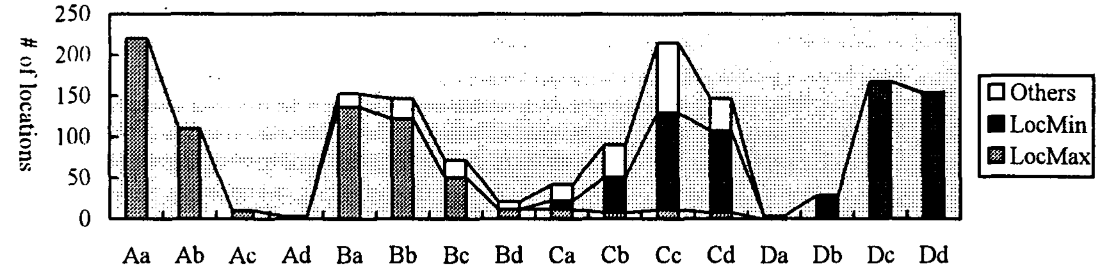

Fig.4 The distribution of dts types in testing texts

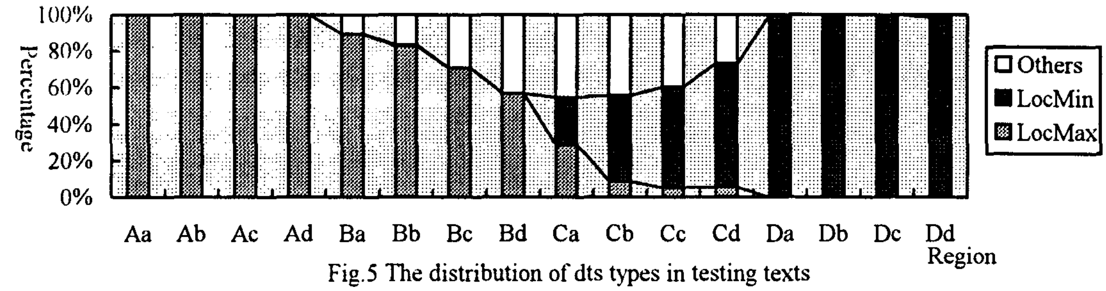

Fig.5 The distribution of dts types in testing texts

References

[1] Liang N.Y., "CDWS: An Automatic Word Segmentation System for Written Chinese Texts", *Journal of Chinese Information Processing*, Vol.1, No.2, 1987 (in Chinese)

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

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

[4] Church K.W., Hanks P., Hindle D., "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

[5] Chan K.J., Liu S.H., "Word Identification for Mandarin Chinese Sentences", *Proc. of COLING-92*, Nantes, 1992

[6] Sun M.S., Lai B.Y., Lun S., Sun C.F., "Some Issues on Statistical Approach to Chinese Word Identification", *Proc. of the 3rd International Conference on Chinese Information Processing*, Beijing, 1992

[7] Sproat R., Shih C.L., "A Statistical Method for Finding Word Boundaries in Chinese Text", *Computer Processing of Chinese and Oriental Languages*, No.4, 1993

[8] Sproat R. *et al*, "A Stochastic Finite-State Word Segmentation Algorithm for Chinese", *Proc. of the 32nd Annual Meeting of ACL*, New Mexico, 1994

[9] Palmer D.D., "A Trainable Rule-based Algorithm for Word Segmentation", *Proc. of the 35th Annual Meeting of ACL and 8th Conference of the European Chapter of ACL*, Madrid, 1997

[10] Sun M.S., Shen D.Y., Huang C.N., "CSeg&Tag1.0: A Practical Word Segmenter and POS Tagger for Chinese Texts", *Proc. of the 6th ANLP*, Washington D.C., 1997