Available online at www.sciencedirect.com
SCIENCE DIRECT®
Information Sciences 162 (2004) 275–285
INFORMATION SCIENCES
AN INTERNATIONAL JOURNAL
www.elsevier.com/locate/ins

A Chinese word segmentation based on language situation in processing ambiguous words

Mao-yuan Zhang [a,*], Zheng-ding Lu [a], Chun-yan Zou [b]

[a] *Department of Computer Science and Technology, HuaZhong University of Science and Technology, Wuhan 430074, PR China*
[b] *Foreign Language School, Central China Normal University, Wuhan 430079, PR China*

Received 17 July 2003; received in revised form 16 September 2003; accepted 25 September 2003

Abstract

While the processing of natural language is beneficial to the text mining, Chinese word segmentation is an important step in the processing of Chinese natural language. In this paper, the convergence essence of the segmentation process is analyzed, and a theory of Chinese word segmentation based on language situation is deducted. Based on the segmentation theory, an algorithm of Chinese word segmentation is presented. Both in theory and from the experiment results, the algorithm is efficient.

Keywords: Language situation; Chinese word segmentation; Ambiguous word

1. Introduction

The explosive growth of web pages on the Internet has created a great demand for new and powerful tools to acquire useful information. In order to take advantage of the information, lots of text mining technologies have been researched on. Text mining has been gaining popularity in the knowledge

* Corresponding author. Tel.: +86-27-87544285-0; fax: +86-27-87544285.
E-mail address: coolphenix@163.com (M.-y. Zhang).

0020-0255/$ - see front matter
doi:10.1016/j.ins.2003.09.010

discover field, particularly with the increasing availability of digital documents in various languages from all around the world [1]. To some extent, the processing of natural language is helpful for the text mining.

Chinese text consists of strings of Chinese characters and punctuations. While English text use space as word delimiter, Chinese text has no delimiters to mark word boundaries other than the punctuations. Hence, it is believed that the word segmentation is a necessary step in processing Chinese texts. Chinese words are composed of one or more characters and are not explicitly delimited. Because of the ambiguous words, it is difficult to segment the string of Chinese characters into words with high accuracy [2].

Although word segmentation is trivial for languages such as English where space is used as word delimiter [3], it is the fundamental task of information processing for languages such as Chinese [4]. Some researchers have proposed some methods of word segmentation. The N-best searching algorithm of Japanese word segmentation [5] and the augmented algorithm of Chinese word segmentation [6], take syntax into consideration, but ignore the statistical rules of morphology. The Chinese language model based on multi-knowledge sources [7] takes the rules of syntax into account, but it is of less robustness and expansibility. The maximum matching algorithm of Chinese word segmentation [8] based on morphology does not consider syntax, and its accuracy is not much high. The fuzzy clustering algorithm of Chinese word segmentation [9] considers the rules of syntax, but its analysis on the features of Chinese is insufficient.

In this paper, a theory of Chinese word segmentation based on language situation is proposed, after analyzing the convergence essence of the segmentation process. The segmentation theory indicates that the Chinese word segmentation should be researched from the angle of language situation. Based on the segmentation theory, an algorithm of Chinese word segmentation is presented in this paper. While the clear words are segmented in virtue of morphology, the ambiguous words are recognized and segmented from the angle of language situation. Both in theory and from the experiment results, the segmentation algorithm is of high accuracy.

2. The segmentation theory based on language situation

Assume a string of Chinese characters is $C_1C_2\cdots C_k$. If the sequence C_iC_{i+1} and the sequence $C_iC_{i+1}C_{i+2}$ are both words in Chinese vocabulary, the two sequences are of the combinatorial ambiguity. If the sequence $C_{i-1}C_i$ and the sequence C_iC_{i+1} are both words in Chinese vocabulary, the two sequences are of the intersectional ambiguity. In 1983, John Barwise and John.Perry established the situation semantics. All contexts exist under the condition of language situation [10], and language situation should be considered for computer to understand the natural language [11]. As follows, the processing of Chinese

word segmentation is analyzed, and the theory of Chinese word segmentation based on language situation is deducted.

2.1. The description based on Markov chain

The probability denotes the happening degree of the phenomena, so it is generally used for the decision-making. The Markov chain that is a series of probabilities on the time dimension, describes the features of transition between two states, and can be used to simulate the random process [12].

Because of the ambiguous words, the Chinese segmentation is not easily solved. In order to express whether a currently processed Chinese character is segmented with the preceding remained Chinese characters into a word or not, two phases are introduced. If the current Chinese character can be segmented with the preceding remained Chinese characters, the phase transits to the phase $s0$, otherwise, the phase transits to the phase $s1$. Because of the ambiguous words, the segmentations in the whole segmentation process of the Chinese characters are uncertain problems of decision-making, and the transition between phase $s0$ and phase $s1$ is uncertain. The Markov chain can be used to describe the features of the transition between two phases, therefore, it is applied to simulate the whole segmentation process.

In the whole segmentation process, all segmentation probabilities are not equal, and may be acquired difficulty. Hence the total average transition probabilities are used for the analysis of the whole segmentation process. Different segmentation methods have different segmentation judgments of ambiguous words, and then have different total average transition probabilities. Assume the total average probabilities of the transitions from $s0$ to $s0$, from $s0$ to $s1$, from $s1$ to $s0$ and from $s1$ to $s1$, are q, $1-q$, p and $1-p$ respectively, so the one-step matrix of the transition denotes as $A = \begin{bmatrix} q & 1-q \\ p & 1-p \end{bmatrix}$.

2.2. The convergence analysis of the Markov chain

Theorem 1. *Let the matrix A be* $\begin{bmatrix} q & 1-q \\ p & 1-p \end{bmatrix}$ *where* $p, q \in [0,1)$. A^n *converges at the matrix* $\begin{bmatrix} \frac{p}{1+p-q} & \frac{1-q}{1+p-q} \\ \frac{p}{1+p-q} & \frac{1-q}{1+p-q} \end{bmatrix}$ *steadily, when n tends towards the infinitude.*

Proof. Assume A^n is the matrix $\begin{bmatrix} x_1(n) & x_2(n) \\ x_3(n) & x_4(n) \end{bmatrix}$, then $A^{n+1} = \begin{bmatrix} x_1(n+1) & x_2(n+1) \\ x_3(n+1) & x_4(n+1) \end{bmatrix}$.

Because $A^{n+1} = A^{n*}A$, it can be obtained as follows that

$$x_1(n+1) = q^* x_1(n) + p^* x_2(n), \tag{1}$$

$$x_2(n+1) = (1-q)^* x_1(n) + (1-p)^* x_2(n), \tag{2}$$

$$x_3(n+1) = q^* x_3(n) + p^* x_4(n), \tag{3}$$

$$x_4(n+1) = (1-q)^* x_3(n) + (1-p)^* x_4(n). \tag{4}$$

Substitute (1) into (2), so that

$$x_1(n+1) + x_2(n+1) = x_1(n) + x_2(n).$$

And then it can be obtained that

$$x_1(n+1) + x_2(n+1) = x_1(1) + x_2(1) = 1. \tag{5}$$

Substitute (5) into (1), so that

$$x_1(n+1) = p + (q-p)^* x_1(n). \tag{6}$$

Assume

$$x_1(n+1) - m = (q-p)(x_1(n) - m). \tag{7}$$

By integrating (7) with (6), it can be gain that

$$m = p/(1+p-q). \tag{8}$$

From (7), it can be obtained that

$$x_1(n+1) - m = (q-p)^n (x_1(1) - m),$$

that is

$$x_1(n+1) = m + (q-p)^n (x_1(1) - m) = m + (q-p)^n (q-m). \tag{9}$$

And thus

$$x_2(n+1) = 1 - x_1(n+1) = 1 - m - (q-p)^n (q-m). \tag{10}$$

Because

$$p, q \in [0,1), \quad |q-p| \leqslant \max(|p|, |q|) < 1. \tag{11}$$

From (9)–(11), it can be obtained that $x_1(n+1)$ and $x_2(n+1)$ converge steadily at m and $1-m$ respectively when n tends towards the infinitude. It is deducted similarly that $x_3(n+1)$ and $x_4(n+1)$ converge steadily at m and $1-m$ respectively when n tends towards the infinitude. Therefore, A^n converges at the matrix $\begin{bmatrix} \frac{p}{1+p-q} & \frac{1-q}{1+p-q} \\ \frac{p}{1+p-q} & \frac{1-q}{1+p-q} \end{bmatrix}$ steadily, when n tends towards the infinitude.

From the above deduction, A^n converges speedily when $|q-p|$ is smaller, while A^n converges slowly when $|q-p|$ is larger. Hence, the convergence speed of A^n and $|q-p|$ are in inverse ratio, and $|q-p|$ is the speed coefficient of the convergence of A^n.

The larger p is, the larger $p/(1+p-q)$ is. So the increase of p increases the convergence values of $x_1(n)$ and $x_3(n)$, and decreases the convergence values of $x_2(n)$ and $x_4(n)$. Similarly, the increase of q increases the convergence values of $x_1(n)$ and $x_3(n)$, and decreases the convergence values of $x_2(n)$ and $x_4(n)$.

2.3. The convergence analysis of the Chinese word segmentation

The Markov chain can be applied to simulate word segmentation process of Chinese sentences. In the process, the matrix A denotes the one-step matrix of the transition, and then the matrix A^n describes the n-step matrix of the transition.

The whole segmentation process consists of many segmentation processes of single sentence, so the whole process can be analyzed from the angle of single sentence. If the end phase of the segmentation process of single sentence is the phase $s0$, the word segmentation is correct, otherwise, the word segmentation is wrong, and the interlunation between two sentences is used to transform the end phase into the phase $s0$. If the convergence value of $x_1(n)$ increases, the probability that the end phase of single sentence is $s0$ increases, and then the accuracy of the word segmentation of single sentence is improved. Hence, p and q are required to be large enough to improve the accuracy of the Chinese word segmentation.

From the angle of the whole segmentation process, if a Chinese sentence is wrongly segmented, a new word segmentation of the sentence is needed. So the number of segmentation processes of single sentence increases, and then the speed of the segmentation process decreases. Therefore, in order to speed the whole segmentation process, the convergence speed of $x_1(n)$ is needed to increase, that is decreasing the value of $|q-p|$.

To sum up, it is required that the value of $|q-p|$ is small enough while p and q are large enough, in order to obtain the right and speedy segmentation.

2.4. The segmentation theory based on language situation

p and q are two kinds of total average transition probabilities. While q denotes the probability that single Chinese character can be segmented into a word, p denotes the probability that the Chinese character can be segmented into a word with its preceding characters. Hence, q describes the segmentation probability at the position after the first character in a word, while p describes the segmentation probability at any positions except the position after the first character in a word. Thus it can be seen that p and q reflect not only the structure of the words, but also the happening frequency of the words in the society. So p and q should be restricted by both morphology and language situation in the society, where language situation in the society is thought as the

globe language situation, and then the values of p and q cannot exceed their maximum.

On one hand, p and q have the difference. They embody the probabilities that the different parts of a word can be segmented, so that they should be constrained with morphology and language situation in the society. On the other hand, p and q have the consistency. Both of them embody the probability that the word can be segmented in a document, so they reflect language situation in the document. Hence, while the difference makes p and q are restricted by language situation in the society, the consistency causes them to be restricted by language situation in the document.

In order to obtain the large values of p and q, and the small difference between them, the consistency between p and q should be maintained when the values of p and q increase. So the values of p and q are restricted by morphology, language situation in the society and language situation in the document. Therefore the excellent method of Chinese word segmentation should be based on morphology, language situation in the society and language situation in the document. Morphology is the basis of clear word segmentation, while language situation in the society and language situation in the document are the foundation of ambiguous word segmentation. The algorithm of word segmentation based on language situation takes the function of language situation, which is proposed in what follows, as the criterion for ambiguous word segmentation.

2.5. The function of language situation

In the process of Chinese word segmentation, a character may be segmented into a word with its preceding characters, or with its preceding characters and successive characters, because of ambiguous words. To obtain right segmentation, a function of language situation is applied. On account of the segmentation theory based on language situation, the function of language situation takes both language situation in the society and language situation in the document into account.

As is known from the theory of communication, the inter-information can describe the compact degree of two codes. Assume w is a Chinese word in the document i, the first Chinese character of w is x, and the Chinese character series after x is y. Language situation in the document can be expressed with the inter-information between x and y as follows.

$$I_l(x:y) = \begin{cases} \log[p_l(xy)/p_l(x)p_l(y)], & \text{if } y \text{ is vacant} \\ \log[p_l(x)], & \text{otherwise.} \end{cases} \tag{12}$$

The happening probability of x in document i is calculated with $p_l(x) = N_i(x)/L_i$, where $N_i(x)$ is the happening times of x in document i, and L_i denotes the number of all characters in document i.

Assume the document i belongs to the directory d, which is a set of many documents. Language situation in the society can be expressed as follows,

$$I_g(x:y) = \begin{cases} \log[p_g(xy)/p_g(x)p_g(y)], & \text{if } y \text{ is vacant} \\ \log[p_g(x)], & \text{otherwise.} \end{cases} \tag{13}$$

The happening probability of x in directory d is calculated with $p_g(x) = N_d(x)/L_d$, where $N_d(x)$ is the happening times of x in directory d, and L_d denotes the number of all characters in directory d.

Based on Eqs. (12) and (13), the function of language situation can be calculated by

$$I(x:y) = \alpha I_g(x:y) + (1-\alpha) I_l(x:y), \tag{14}$$

where parameter $\alpha \in [0,1]$. Because the directory d is the subsets of all documents in the whole society, the inter-information $I_g(x:y)$ reflects language situation in the society to some extent, and thus the value of parameter α depends on the directory d.

3. The framework and the algorithm

3.1. The framework

The framework of Chinese word segmentation consists of two parts. The first part is the preprocessing, and the other is the word segmentation. The former extracts the character series of web documents and preprocesses the character series. The latter segments the character series into words with the segmentation method based on language situation.

In the preprocessing, while a HTML parser is used to extract the character series from the web pages, the segmentation delimiters are added to the character series at some places including both the front and the rear of non-Chinese characters. The segmentation delimiters shorten the character series, and are useful to the word segmentation.

The ambiguous segmentation exists in the segmentation of Chinese word, because of the features of Chinese. Hence, the word segmentation adopts the function of language situation to eliminate the ambiguous segmentation. As is shown in Fig. 1, the word segmentation consists four layers. After processing the character series $A_1 \cdots A_{i-1}$, the rear part of the character series $A_1 \cdots A_{i-1}$ may not be segmented into a word, and is remained. Assume the remained character series is variable wordleft. The first layer is to pick out the currently processed character A_i and its three successive characters that are A_{i+1}, A_{i+2} and A_{i+3}, and set them as x, y, z and u respectively. The second layer is to construct some combinatorial character series from wordleft, x, y, z and u, such as wordleft + x. The second layer is also to select the combinatorial character

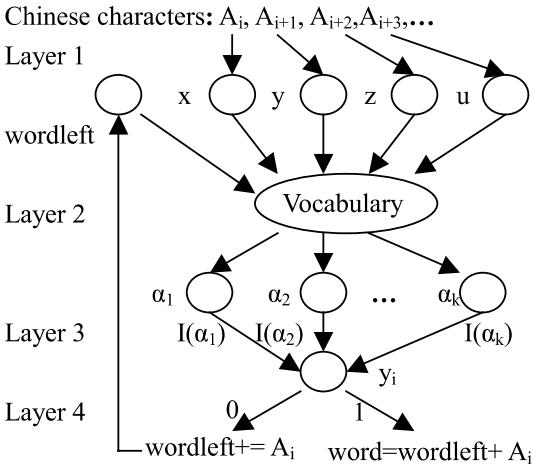

Fig. 1. The segmentation of Chinese word based on language situation.

series that are words in the vocabulary, and sets them as $\alpha_1, \ldots, \alpha_k$, in which α_1 is set as wordleft + x. The third layer is to calculate the function $I(\alpha_1), \ldots, I(\alpha_k)$ for $\alpha_1, \ldots, \alpha_k$ by Eq. (14). The last layer is to compare the function values, and if $I(\alpha_1)$ is the greatest value, set $y_i = 1$, otherwise, set $y_i = 0$. While $y_i = 1$ denotes wordleft + x can be segmented into a word, $y_i = 0$ shows wordleft + x is not fit to be segmented.

3.2. The algorithm

The number of Chinese characters in a word (NCCW) is variable. For example, the NCCW of some words is two while that of other words is three. As is shown in Table 1, the statistic result [13] makes it clear that the ratio of the Chinese words, of which the NCCW exceeds four, to all Chinese words is about 0.28%. Hence, the algorithm of Chinese word segmentation is used for the ambiguous words, of which the NCCW is not greater than four.

The segmentation algorithm based on language situation is shown as follows. (1) Let variable wordleft be vacant, where wordleft is defined in above

Table 1
The distribution of NCCW

	NCCW						
	1	2	3	4	5	6	7
Quantity	2606	33 527	3693	3622	83	36	3
Ratio (%)	5.98	76.95	8.48	8.31	0.19	0.08	0.007

section; (2) if all the characters have been processed, go to step 10, otherwise, process the next character; (3) assume the currently processed character is x, and its three successive characters are y, z, u respectively; (4) while wordleft + x is set into $w1$, search those combinatorial character series from wordleft, x, y, z and u, in the vocabulary, and then set them into $w2, \ldots, wk$; (5) calculate the function values from the word $w1$ to the word wk by Eq. (14); (6) compare those function values, and assume the word that has the maximal function value is the word $w$max; (7) if the word $w$max is wordleft + x, go to step 9; (8) wordleft = wordleft + x, go back to step 2; (9) segment the character series wordleft + x into a word, and then let wordleft be vacant and go back to step 2; (10) the algorithm is end.

4. Experiments and results

Select 8730 web pages and classify them into five document sets. Those document sets include news, sport, fun, technology and other sets, and are used to acquire language situation in the society. From those web pages, select 610 web page as testing documents, which is used to obtain language situation in the document. Pick out 1360 sentences that have ambiguous words from the testing documents, and segment them with three methods. The methods are the reverse maximal match method, the statistic method of word frequency and the method based on language situation, and they are named as method A, method B and method C respectively. The results of segmentation are shown in Table 2.

Compare the result of method A with that of method C, it is found out that the latter is higher than the former by about 30%. Hence, the reverse maximal match method is of low accuracy, and is less adapted for the ambiguous segmentation than the method based on language situation.

Compare the result of method B with that of method C, it is clear that the latter is higher than the former by 8%. While the difference between the

Table 2
The segmentation results of three methods

Class	Quantity	Method A		Method B		Method C	
		Right	Accuracy (%)	Right	Accuracy (%)	Right	Accuracy (%)
News	370	230	62.2	300	81.1	331	89.5
Sports	230	144	62.6	188	81.7	206	89.6
Fun	290	182	62.8	239	82.4	261	90.0
Technol.	190	118	62.1	153	80.5	169	88.9
Others	280	175	62.5	228	81.4	250	89.3
Total	1360	849	62.4	1108	81.5	1217	89.5

maximal accuracy and the minimal accuracy of method B is 1.9%, that of method C equals to 1.1% and is smaller. Therefore, the method based on language situation has not only higher accuracy, but also steadier accuracy.

Through the above analysis on the results, the method based on language situation has the high accuracy, which is near to 90%, so the method is one of the best methods of Chinese word segmentation at present.

5. Conclusion

In this paper, a theory of Chinese word segmentation based on language situation is proposed by analyzing the process of Chinese word segmentation, and a segmentation method based on the segmentation theory is put forward. The segmentation method deals with the ambiguous words from the angle of language situation, and is of high accuracy. Therefore the segmentation method based on the segmentation theory is one of the best methods of Chinese word segmentation at present.

References

[1] C.-H. Lee, H.-C. Yang, Text mining of bilingual parallel corpora with a measure of semantic similarity, in: Proceedings of 2001 IEEE International Conference on Systems, Man, and Cybernetics, 2001, pp. 470–475.
[2] H. Zhu, T. Ruan, Q.-X. Yu, Studies on text segment algorithms' influence on Chinese-based information filtering, Computer Engineering and Application (13) (2002) 62–65.
[3] W.-S. Lo, P.-F. Wong, M.-H. Siu, Maximum likelihood algorithm on Chinese word segmentation, in: Proceedings of the 6th International Conference on Signal Processing, 2002, pp. 468–471.
[4] K.-Y. Liu, J.H. Zheng, Research of automatic Chinese word segmentation, in: Proceedings of 2002 International Conference on Machine Learning and Cybernetics, 2002, pp. 805–809.
[5] M. Nagata, A stochastic Japanese morphological analyzer using a forward-DP backward-A^*N-Best search algorithm[C], in: Proceedings of the 15th International Conference on Computational Linguistics, Kyoto, Japan, 1994, pp. 201–207.
[6] S.-H. Bai, Chinese word segmentation and auto tagging of syntactical functions, in: Proceedings of 1995 Advances and Applications on Computational Linguistics, 1995, pp. 56–61.
[7] B.-Q. Liu, X.-L. Wang, Y.-Y. Wang, Research on a Chinese language model based on multi knowledge sources and its implementation, Journal of Computer Research & Development 39 (2) (2002) 231–235.
[8] H. Guo, Z.-Y. Shu, W. Wang, J. Chui, An augmented MM algorithm of word segmentation, Microcomputer Application 18 (1) (2002) 13–15.
[9] J.-F. Li, Y.-F. Zhang, J.-J. Lu, Application of fuzzy clustering algorithm in Chinese document clustering, Computer Engineering 28 (4) (2002) 15–16.
[10] T.H. Lecky, Language and Context: A Functional Linguistics Theory of Register, Pinter, London, 1995.
[11] Q.-S. Gao, Intelligent Technology and System Basis, Peking University Press, Peking, 1990.

[12] A.-S. Rodionov, H. Choo, H.-Y. Youn, Process simulation using randomized Markov chain and truncated marginal distribution, The Journal of Supercomputing 22 (1) (2002) 69–85.
[13] Y. Li, The Information Processing with the Criterion of Chinese Word Segmentation and the Method of Auto Segmentation, Tsinghua University Press, Peking, 1994.