Combing Different Statistical Evidence For Chinese Word Segmentation

Joon Ho Lee
School of Computing, Soongsil University
1-1 Sangdo-dong, Dongjak-gu
Seoul 156-743, Korea
joonho@computing.soongsil.ac.kr

Hyun Jung Lee
Search Solutions Co., Ltd.
7th FL. Samhueng Bldg., 705-9, Yoksam-dong, Kangnam-gu
Seoul 135-080, Korea
juvenile@searchsolutions.co.kr

1. Introduction

Information Retrieval(IR) deals with the problems of retrieving useful documents from a large number of stored documents. An essential component of IR systems is to generate index terms from documents and queries, and the indexing produce a crucial effect on the quality of retrieval results. Word-based indexing has been widely used in many languages such as English, Spanish et al. to represent documents and queries. The conventional word-based indexing includes identification of words with some delimiters such as space, comma and so on.

Chinese is an agglutinating language, that is, there is no delimiters or white spaces to mark word boundaries. Hence, the first step toward word-based indexing is to break a sequence of characters into words, which is often called word segmentation. Many Chinese word segmentation methods have been investigated in the literature (Chen & Liu, 1992; Leung & Kan, 1996; Nie, Hannan, & Jin, 1995; Sproat & Shih, 1990). In this paper we investigate previous statistical segmentation methods using mutual information (Chen et al., 1997) or head-tail information (Ogawa & Matsuda, 1997), and propose to combine multiple evidence from the statistical segmentation methods. Experimental results show that the proposed method might provide better performance than the previous methods.

2. Using Mutual Information

The statistical segmentation using mutual information is one of the widely used statistical segmentation methods. It first calculates the association strength of two adjacent characters called the mutual information value of the two characters. Suppose that $f(c)$ and $f(c_1c_2)$ are occurrence frequency values in the collection for each Chinese character c and each bigram c_1c_2, respectively,

appearing at least once in the collection, and N is the number of characters in the collection. For each Chinese bigram c_1c_2, the mutual information value $M(c_1c_2)$ is computed by the following formula.

$$M(c_1c_2) = \log_2 \frac{f(c_1c_2) \bullet N}{f(c_1) \bullet f(c_2)} \quad (1)$$

After getting mutual information values, we can segment the written Chinese text using the purely statistical method. First, the written text is broken into phrases – a consecutive sequence of Chinese characters is considered as a phrase. Second, each phrase is segmented with the following procedure: (i) Treat the bigram of the largest mutual information value as a word and then remove it from the phrase. The removal of the bigram may result in one or two shorter phrases, (iii) Perform step 1 on each of the shorter phrases until all phrases consist of one or two characters.

3. Using Head-Tail Information

The statistical segmentation using head-tail information exploits segmentation probabilities of all bigrams occurring in the collection. The segmentation probability of a bigram c_1c_2 is estimated by using statistical information about its constituent characters. That is, the segmentation probability of a bigram c_1c_2, $S(c_1c_2)$ is a product of the tail probability of the first character c_1, $P_{tail}(c_1)$ and the head probability of the second character c_2, $P_{head}(c_2)$:

$$S(c_1c_2) = P_{tail}(c_1) \times P_{head}(c_2) \quad (2)$$

$$P_{tail}(c_1) = \frac{\#(c_1 \text{ appeared at the tail of words})}{\#(c_1 \text{ appeared at any place})} \quad (3)$$

$$P_{head}(c_2) = \frac{\#(c_2 \text{ appeared at the head of words})}{\#(c_2 \text{ appeared at any place})} \quad (4)$$

where $\#(x)$ represents the number of x's occurrences. Computing these probabilities such as $P_{tail}(c_1)$ and $P_{head}(c_2)$ requires a morphologically analyzed corpus in which words are identified manually or automatically. Using the segmentation probabilities determined as explained above, text is segmented into disjoint parts at points whose segmentation probabilities are greater than a segmentation threshold T_{seg}.

4. Combining Multiple Statistical Evidence

It has been known that retrieval effectiveness can be significantly improved by combining multiple evidence from different query or document representations, or multiple retrieval techniques. We adapt this idea to develop a new statistical segmentation method in that we combine the mutual information with the head-tail statistical information. In the remainder of this section we describe the proposed method in detail.

1. *Derivation of the mutual information*: First of all, compute the mutual information $M(c_1c_2)$ for each Chinese bigram c_1c_2. Second, normalize every mutual information value into a value between 0 and 1. That is, when the minimum and maximum mutual information values are given as M_{min} and M_{max}, respectively, compute the normalized mutual information value of a bigram c_1c_2 as follows:

$$M_{norm}(c_1c_2) = \frac{M(c_1c_2) - M_{min}}{M_{max} - M_{min}} \tag{5}$$

   Finally, transform each normalized mutual information value by subtracting the normalized mutual information value from 1.

$$M_{trans}(c_1c_2) = 1 - M_{norm}(c_1c_2) \tag{6}$$

2. *Derivation of the head-tail information*: The head-tail information is another statistical evidence being incorporated in our new statistical segmentation method. We calculate the head and tail probabilities of each character and compute the segmentation probability of each bigram as mentioned in section 3. We, then, normalize the segmentation probability into a value between 0 and 1.

$$S_{norm}(c_1c_2) = \frac{S(c_1c_2) - S_{min}}{S_{max} - S_{min}} \tag{7}$$

3. *Combination of the two statistical information*: We combine M_{trans} resulting from the mutual information with S_{norm} derived from the head-tail information. The new statistical evidence S_{comb} is the weighted summation of M_{trans} and S_{norm} as follows:

$$S_{comb} = M_{trans} + \alpha \bullet S_{norm} \tag{8}$$

   where α is a constant.

4. *Segmentation method*: The segmentation method using mutual information treats the bigram of the largest mutual information value as a word and removes it from the string recursively, until all strings consist of one or two characters. Whereas, the segmentation method using head-tail information segments the string into disjoint parts at the point whose segmentation probability is greater than a threshold T_{seg}. In the proposed method, we rank the value S_{comb} in decreasing order and exploit the top-ranked S_{comb} value to segment the string into two disjoint parts. The segmentation is continued until the string of each resulting segment consists of one or two characters.

5. Performance Evaluation

In order to evaluate the performance of Chinese word segmentation methods, we need a

written Chinese text and a list of words included in the text. We created a test collection consisting of 442 newspaper articles and a list of words extracted by a human expert from the articles. It is customary to compute values of the recall and precision to evaluate the effectiveness of an IR system. We adapt the recall and precision as follows:

$$recall = \frac{number\ of\ words\ segmented\ by\ the\ program}{number\ of\ words\ segmented\ by\ the\ human} \tag{9}$$

$$precision = \frac{number\ of\ words\ segmented\ correctly}{number\ of\ words\ segmented\ by\ the\ program} \tag{10}$$

Table 1 shows the precision and recall given by the statistical Chinese segmentation methods. The segmentation method using head-tail information provides the best performance when the threshold T_{seg} is equal to 0.25, and the segmentation method combining multiple statistical evidence gives the best performance when the weight α is equal to 0.5. We can see that the proposed method gives slightly better segmentation performance than the previous methods in terms of the recall. We did not get as much improvement as we expected. Further investigation should be done to figure out the practical meaning of the statistical information.

Table 1. Segmentation performance of the statistical Chinese segmentation methods

	Precision	Recall
Mutual Informaiton	73.4900	73.9096
Head-Tail Information (T_{seg}=0.25)	60.1621	61.7320
Combination (α=0.5)	74.0087	79.1532

REFERENCE

Chen, A., He, J., Xu, L., Gey, F.C. & Meggs, J. (1997). Chinese text retrieval without using a dictionary. In: *Proceedings of ACM SIGIR Conference*, Philadelphia, PA, pp. 42-49.

Chen, K.J. & Liu, S.H. (1992). Word identification for mandarin Chinese sentences. In: *Proceedings of COLING*, pp. 23-28.

Leung, C.H. & Kan, W.K. (1996). A statistical learning approach to improving the accuracy of Chinese word segmentation. *Literary and Linguistic Computing*, Vol. 4, No. 2, pp. 87-92.

Nie, J.Y., Hannan, M.L. & Jin, W. (1995). Combining dictionary, rules and statistical information in segmentation of Chinese. *Computer Processing of Chinese and Oriental Languages*, Vol. 9, No. 2, pp. 125-143.

Ogawa, Y. & Matsuda, T. (1997). Overlapping statistical word indexing: A new indexing method for Japanese text. In: *Proceedings of ACM SIGIR Conference*, Philadelphia, PA, pp. 226-234.

Sproat, R. & Shih, C. (1990). A statistical method for finding word boundaries in Chinese text. *Computer processing of Chinese and Oriental Languages*, Vol. 4, pp. 336-351.