A Simple Heuristic Approach for Word Segmentation

Wing-Kwong Wong, Chenming Hsu, Jien-Iao Chen, Jien-Chi Yu
Dept. of Electronic Eng., Natl. Yunlin Inst. of Tech., Touliu, Yunlin, Taiwan.
wongwk@el.yuntech.edu.tw

Abstract

Previous approaches to Chinese word segmentation includes maximal matching heuristic, morphological rules, and POS tag statistics. This paper proposes to estimate the word occurrence probabilities with some "unlikelihood" scores based only on word lengths. Also, the problem of maximizing likelihood is shown to be equivalent to the graph problem of shortest path, whose edges stands for words with their corresponding unlikelihood scores.

Introduction

Chinese sentences have no white spaces to delimit words, unlike English. Word segmentation has been a fundamental problem in Chinese text processing since almost all applications in this area must deal with this problem. Many solutions have been published before. Almost all solutions use a word dictionary to determine whether a given character sequence is a legal word. The earliest proposal is probably the maximal matching approach, which favors long words during segmentation (e.g., [Li et al. 91], [Lochovsky & Chung 94]). Since more frequently used words are more likely to occur, word frequencies, which estimate independent occurrence probabilities, are used in ([Chang et al. 91], [Nie et al. 95]). Furthermore, independent word probabilities are not as powerful as co-occurrence statistics of grammatical categories in predicting word sequence. Thus, morphological or POS (part of speech) tagging statistics (e.g., bigrams and trigrams), are used in [Li et al. 91], [Chiang et al. 92], [Lin et al. 93], [Pan & Chang 93], [Luk 94]. Morphological rules, which use POS categories of words to predict the formation of compound words, are used in [Lin et al. 93]. When there is no unknown words in the tested sentences, all the above approaches perform quite well, with accuracy rates ranging from 95% to 99%. When many unknown words exist in the tested data, which is generally the case in real-life texts, however, performance degrades down to 60%. Some approaches to detecting unknown words are proposed in [Chiang et al. 92], [Lin et al. 93], and [Luk 94].

Model

The word segmentation problem can be stated formally in the following way:

Given a character sequence $C_1C_2C_3 \ldots C_n$, determine the optimal word sequence $W_1W_2W_3 \ldots W_m$, such that each W_i, where i=1,…,n, spans a subsequence of C_j's, and all such subsequences do not overlap and the entire word sequence spans the entire character sequence.

Intuitively, the optimal word sequence is one that would make the most sense out of the character sequence in the given communication context. All approaches to word segmentation attempt to give an operational, i.e., computational, approximation to this intuitive definition of optimality. For statistical approach, this is equivalent in choosing the word sequence that maximizes the conditional probability of word sequence $W_1W_2 \ldots W_m$ given the character sequence $C_1C_2 \ldots C_n$ (e.g., [Pan & Chang 93]):

$$\max P(W_1 W_2 \ldots W_m | C_1 C_2 \ldots C_n).$$

The simplest and roughest computational model is one that assumes the words are all independent of each other and independent of the given character sequence:

$$\max P(W_1 W_2 \ldots W_m | C_1 C_2 \ldots C_n)$$
$$\cong \quad \max P(W_1) \ldots P(W_m) \quad \cong \quad \max \prod_i P(W_i)$$

The maximal matching approach, which favors the longest words first met when scanning the characters in a sentence either from left to right or vice versa. Its basic intuition is that the longer a word is, the greater is its occurrence probability when its character sequence is found in a sentence. It provides an effective heuristic and can achieve a high accuracy rate. However, it is only a local heuristic and does not achieve a global optimum. Therefore, it is very efficient but can fail in cases where the first met longest word produces an incorrect segmentation. For example, 【省長上行政院】 is segmented as 【省長 | 上行 | 政 | 院】, since 上行 is the longest word scanned from the left at the character 上. However, the correct segmentation is 【省長 | 上 | 行政院】. If more global information is used, i.e., the later but longer word 行政院 is considered, then a correct segmentation should result.

The basic flaw of maximal matching is that it is only a local heuristic. We propose to keep its basic intuition that the likelihood of a longer word is greater but to use it with global optimization, similar to the above baseline model $\max \prod_i P(\mathrm{W}_\mathrm{i})$. To achieve this goal, several mathematical transformations are needed. First, to maximize the product of probabilities is equivalent to maximize the sum of the log of these probabilities ([Chiang et al. 92]). Second, some likelihood function can be used to estimate the log of probabilities. Third, likelihood maximization is equivalent to "unlikelihood" minimization and some unlikelihood function can be selected to reflect the likelihood function:

$$
\begin{aligned}
& \max \prod_i P(\mathrm{W}_\mathrm{i}) \\
\cong\ & \max \sum_i \log P(\mathrm{W}_\mathrm{i}) \\
\cong\ & \max \sum_i \mathit{likelihood}(\mathrm{W}_\mathrm{i}) \\
\cong\ & \min \sum_i \mathit{unlikelihood}(\mathrm{W}_\mathrm{i})
\end{aligned}
$$

Thus, we propose to assign some unlikelihood scores to words based on their length:

In dictionary?	Word Length	Unlikelihood Score
No	Don't care	∞
No	>=5	∞
Yes	1	7
Yes	2	4
Yes	3	2
Yes	4	1

If a character sequence is not in the dictionary, then its unlikelihood score is infinity. To keep the analysis simple, word length is limited to four. This would also reduce the dictionary size and the running time for checking whether a character sequence is in the dictionary or not. Thus, there is no need to check character sequence whose length exceeds four. For words that are in the dictionary, the scores are 7, 4, 2, 1 respectively for words with lengths 1, 2, 3, and 4. This means that longer words are less likely to occur as random character sequences in natural texts and is the basic heuristic for the maximal matching method. This score assignment, however, asserts more than this basic heuristic. For example, this assignment says that a two bi-character words 【 AB | CD 】, whose unlikelihood score is 4+4=8, is more likely than a three-character word followed by a single-character word 【 ABC | D 】 or a single-character word followed by a three-character word 【A | BCD】, whose unlikelihood scores are 7+2=9. An example is 【台北 | 市民】 is better than 【台北市 | 民】. Similarly, 【AB | CDE】, with score 6, is more better than 【A | BCDE】 or 【ABCD | E】, with score 8; 【ABC | DEF】, with score 4, is more better than 【AB | CDEF】 or 【ABCD | EF】, with score 5. These assumptions are subject to further empirical testing.

Algorithm

Word segmentation problem is commonly portrayed as an optimization problem. [Fan & Tsai 87] uses a relaxation algorithm. [Chang et al. 91] considers segmentation as a constraint satisfaction problem and employs a dynamic programming method called arc consistency. [Nie et al. 94] presents an algorithm that seems to be a recursive version of the Viterbi algorithm (e.g., [Allen 95], [Bertsekas 87]). All these algorithms are actually variations of solutions for the shortest path problem, which is a fundamental graph problem (e.g., see [Ahuja et al. 93]). Here is how we transform word segmentation into such a graph problem:

1. Put the number 0 at the front of the sentence in question and move past the first character.
2. Put the next number between the last character and the next character.
3. Repeat Step 2 until the last character of the sentence is encountered.
4. Then put the next number following the last character.

The resulting row of symbols becomes:

$$0 \quad C_1 \quad 1 \quad C_2 \quad 2 \quad C_3 \quad \ldots \quad n\text{-}1 \quad C_n \quad n$$

where C_i is the ith character in the sentence, and n is the number of characters in the sentence. Each number x in the row of symbols is considered as a node called node_x in a graph. Then any legal word $C_iC_{i+1}\ldots C_j$ found in dictionary, where i<=j, in the sentence is a directed edge from node_i to node_{j+1}. The distance of the edge from node_i to node_{j+1} is the unlikelihood score of the word. Therefore the segmentation problem becomes the graph problem of finding the shortest path from node_0 to node_n---a path is a series of edges connecting the source node and the destination node and the path distance is the sum of distances of all the edges on the path.

With the proposed scheme of unlikelihood score assignment, the unlikelihood score of each edge (or character sequence) and that of each path (or segmentation) are given below (remember the node numbering scheme: 0 台 1 北 2 市 3 民 4.):

Character sequence	Edge	Unlikelihood score	Legal word?
台, 北, 市, 民	0-1, 1-2, 2-3, 3-4	7	Yes
台北	0-2	4	Yes
北市	1-3	4	Yes
市民	2-4	4	Yes
台北市	0-3	2	Yes
北市民	1-4	∞	No
台北市民	0-4	∞	No

	Segmentation	Path	Unlikelihood score
1	【台北市民】	0-4	∞
2	【台｜北市民】	0-1-4	1 + ∞ = ∞
3	【台北市｜民】	0-3-4	2 + 7 = 9
4	【台北｜市｜民】	0-2-3-4	4 + 7 + 7 = 18
5	【台｜北市｜民】	0-1-3-4	7 + 4 + 7 = 18
6	【台｜北｜市｜民】	0-1-2-3-4	7 + 7 + 7 + 7 = 28
7	【台北｜市民】	0-2-4	4 + 4 = 8
8	【台｜北｜市民】	0-1-2-4	7 + 7 + 4 = 18

According to the above data, the shortest path is 0-2-4, which corresponds to Segmentation 7 【台北｜市民】, since edge 0-2 stands for the word 台北 and edge 2-4 stands for the word 市民. Comparing automatic segmentation results to human segmented texts indicates an accuracy rate of 98.5%.

Acknowledgment

The authors thank Dr. Wen-Lian Hsu for discussing ideas on Chinese text processing and for providing a dictionary. This research is supported by the NSC with grant NSC86-2511-S-224-001-ICL.

References

Ahuja, R. K., T. L. Magnanti & J. B. Orlin, “Network Flows,” Prentice Hall, New Jersey, 1993.

Allen, J. “Natural Language Understanding,” Benjamin-Cummins, Redwood City, CA, 1995, p.202.

Bertsekas, D. P., “Dynamic Programming,” Prentice-Hall, N.J., 1987, p.30.

Chang et al. 張俊盛、陳志達、陳舜德，限制式滿足及機率最佳化的中文斷詞方法， ROCLING IV, 1991, pp. 147-165.

Chiang, T. H., J. S. Chang, M. Y. Lin and K. Y. Su. “Statistical models for word segmentation and unknown word resolution.” ROCLING V, 1992, pp. 123-146.

Fan, C. K. & W. H. Tsai, “Automatic word identification in Chinese sentences by the relaxation technique,” Proc. of National Computer Symposium, 1987, pp. 423-431.

Li et al. 黎邦洋、藺蓀、孫朝奮、孫茂松，一種主要使用語料庫標記進行歧義校正的、最大匹配漢語自動分詞算法設計，ROCLING IV, 1991, pp. 147-165.

Lin, M. Y., T. H. Chiang and K. Y. Su, “A preliminary study on unknown word problem in Chinese word segmentation,” ROCLING VI, 1993, pp. 119-141.

Lochovsky, A. F. & K. H. Chung, “Word segmentation for Chinese phonetic symbols,” 1994 International Computer Symposium, Vol. 2, 1994, pp. 911-916.

Luk, W. P. R., “Chinese-word segmentation based on maximal-matching and bigram techniques,” ROCLING VII, 1994, pp. 273-282.

Nie, J. Y., X. Ren and M. Brisebois, “A unifying approach to segmentation of Chinese and its application to text retrieval,” ROCLING VIII, 1995, pp. 175-190.

Pan & Chang 彭載衍、張俊盛，中文辭彙歧義之研究——斷詞與詞性標示，ROCLING VI, 1993, pp. 173-193.