Research on Word Segmentation for Chinese Sign Language

Yinchao Cheng[1], Baocai Yin, and Yanfeng Sun

Key Laboratory of Multimedia and Intelligent Software, Beijing University of Technology, Beijing, 100022, China
haohuanc@emails.bjut.edu.cn, {ybc,yfsun}@bjut.edu.cn

Abstract. It remains to be a difficult issue to convert Chinese language into Chinese sign language, which makes it hard to implement an obstacle free Chinese sign language information service under pervasive environment. This paper presents an improved algorithm of forward maximum match approach (MM) and backward maximum match approach (FMM), by taking the characteristics of Chinese sign language into consideration. Besides, a method to reorganize the Chinese sign language dictionary is presented. This paper also proposes a novel strategy of disambiguation based on the statistical information of the context and the mutual information. The experiment results indicate that the accuracy and the efficiency of word segmentation can improve significantly compared to conventional algorithms.

Keywords: Word Segmentation; Chinese Sign Language; Disambiguation; Mutual Information

1 Introduction

In human languages, a combination of multiple continuous words, or phrases, is usually a minimum meaningful unit, and word segmentation (WS) is one of the major issues of information processing in character-based languages. Because there are no explicit word boundaries in these languages, WS is important for information retrieval, machine translation, lexicon construction, digital libraries, and Chinese sign language. The conventional WS is different from the one in Chinese sign language environment. It should not be based on the subjective evaluation of the views, but be evaluated by whether it could contribute to improving the accuracy and precision of Chinese sign language synthesis.

The basic approaches of word segmentation in character-based languages can be partitioned into two categories: statistic-based [8] [9] and dictionary-based [1]. Statistic-based approaches make use of statistical properties, such as frequencies of characters and character sequences in the corpus [2]. In practice, however, to choose a corpus which is big enough and includes all categories is impossible, a statistic dictionary that contains all possible words is unfeasible, costly and unnecessary [3]. Dictionary-based approaches use a dictionary to identify words. When matched in the dictionary, a sequence of characters will be extracted as a word. There are many match criteria in literature, such as maximum match, minimum match and hybrid approach. The maximum match approach can be further divided into MM method and FMM method.

However, WS under Chinese sign language environment is quite different from conventional WS. Firstly, speed and accuracy must be considered for the approach of WS under Chinese sign language environment. Each word is translated into a specific series of sign language action, so the accuracy of WS has direct impact on the accuracy of animation synthesis for Chinese sign language. Secondly, Chinese sign language has its particularity in following aspects: (1) Two dictionaries, basic word dictionary (BWDIC) and finger word dictionary (FWDIC); (2) Low vocabulary in BWDIC, only about 6,000 items; (3) Distribution disparity in BWDIC, words made up of single and double characters take up a considerable proportion (total 90.98%). Because of the particularity, new problem will be appeared in WS. Because of the particularity, new problem will be appeared in WS, such as "大学生" , "主流程" ,etc. Each of them is a word in the natural language dictionary, so it will not generate ambiguity in WS.

[1] Project supported by the National Natural Science Foundation of China (No.60375007、60533030), the Beijing Natural Science Foundation(No.4061001).

But these phrases do not exist in sign language dictionary, it has different segmentation for these phrases: "/大学/生/", "/大/学生/", "/主/流程/", "/主流/程/".

In summary, because of the strict requirements of WS for Chinese sign language, the conventional segmentation system can not be suitable for dealing with Chinese sign language directly. This paper proposed a novel segmentation algorithm, through reorganizing the structure of dictionary and improving the WS algorithm. It applies an algorithm based on the statistical information of the context and mutual information (MI) which help for disambiguation. Experiment results show that both efficiency and accuracy of the proposed method has been improved greatly.

2 Statistical Language Model and Mutual Information

In the proposed algorithm, it applies both uni-gram of characters and mutual information of adjacent characters. We'll refer the phrase "target text" below to the text currently being segmented. Parameters in the model can be calculated from the target text and a manually tagged corpus.

2.1 Statistical Language Model

For an N-grams [4] [5], suppose W is one sequence of N characters in a given field (Fig.1): $W = w_1 w_2 w_3 w_4 \cdots w_n$, and the occurrence probability of any w_i is only related to its previous N-1 words(N-gram)[6], namely:

$$P(w_i \mid w_1 w_2 w_3 \cdots w_{i-1}) = P(w_i \mid w_{i-N+1} \cdots w_{i-1}) \tag{1}$$

Then,

$$\begin{aligned} P(W) &= P(w_1)P(w_2 \mid w_1)P(w_3 \mid w_2 w_1)\cdots P(w_i \mid w_{i-N+1} \cdots w_{i-1})\cdots\cdots \\ &= \prod P(w_i \mid w_{i-N+1} \cdots w_{i-1}) \quad (i = 1、2、3\cdots n) \end{aligned} \tag{2}$$

When N=l or 2**,** the statistical language model is called uni-gram and bi-gram respectively.

According to (1) and (2), when N =1, (2) can be simply written as:

$$\begin{aligned} P(W) &= P(w_1)P(w_2)P(w_3)\cdots P(w_i)\cdots\cdots \\ &= \prod_n^1 P(w_i) \quad (i = 1、2、3\cdots n) \end{aligned} \tag{3}$$

$P(w_i)$ is calculated by the following formula, and parameters in the formula can be calculated from the target text.

$$p(w_i) = \frac{frequency\_word(w_i)}{\sum_n^1 frequency\_word(w_i)} \times 100\% \tag{4}$$

Here $frequency\_word(w_i)$ is times of occurrence for w_i and $\sum_n^1 frequency\_word(w_i)$ indicates the times of occurrence for all words in the target text.

2.2 Mutual Information

Mutual information (MI) [7] can be used to measure the coherence of the adjacent characters and is applied widely in statistic-based WS, where the adjacent characters with high MI score are identified as a word. In our approach, similarly, we identify the adjacent characters as a word if its MI score is higher than a predefined threshold.

Consider a sequence of characters: $c_1c_2c_3\cdots c_ic_{i+1}\cdots c_n$, the MI of characters c_i and c_{i+1} is computed by equation 5:

$$F_{mi}(c_ic_{i+1}) = \log_2 \frac{P(c_ic_{i+1})}{P(c_i)P(c_{i+1})} \tag{5}$$

Where $P(c_ic_{i+1})$ is the occurrence probability of the character sequence c_ic_{i+1}, which is estimated by the number of times that c_i is followed by c_{i+1}, normalized by N which is the total number of words in the corpus. $P(c_i)$ is the probability of character c_i which is estimated by the total occurrences of the word c_i normalized by N, namely:

$$P(c_ic_{i+1}) = \frac{freq_{corpus}(c_ic_{i+1})}{N}, P(c_i) = \frac{freq_{corpus}(c_i)}{N} \tag{6}$$

Therefore, equation 5 is represented as follows:

$$F_{mi}(c_ic_{i+1}) = \log_2\left(\frac{N \times freq_{corpus}(c_ic_{i+1})}{freq_{corpus}(c_i) \times freq_{corpus}(c_{i+1})}\right) \tag{7}$$

3 WS under Chinese Sign Language Environment

3.1 Pre-processing

In the course of WS, efficiency will be reduced gradually with the increase of the length of sentence, so it performs the preprocessing to the target text got from the Web. Besides removing some useless symbols, the most important thing is to divide the text into some shorter fields.

First of all, divide the text into some sentences according to the symbol of pause, such as: comma, full stop, etc. Then, divide the sentence into some fields according to some special symbols, figure, character, etc. Word segmentation, recognition of ambiguity and disambiguation will deal with these fields.

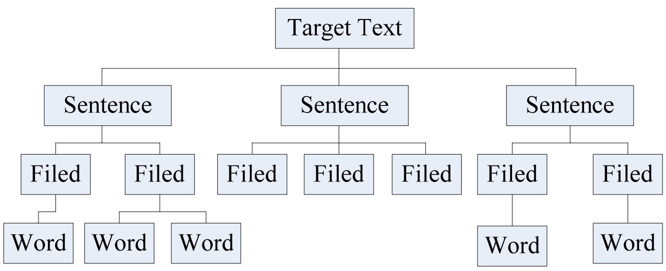

Fig. 1. Text hierarchy structure

3.2 Proposed WS Algorithm

Characteristic of the BWDIC. According to the statistical information of the dictionary, as shown in Table 1: Single character and double characters are counted in the majority, so we redesign the structure of the BWDIC, divided into four layers and formed a tree-like structure.

Table 1. Statistics of BWDIC

Length of words	One	Two	Three	Four	More than five
Number	815	4265	388	93	5
Proportion	14.60%	76.38%	6.90%	1.70%	0.09%

Proposed MM Algorithm. The length of longest word in dictionary is usually longer than the length of word segmented from the target text; therefore the conventional MM algorithm will waste a lot of time for matching. Furthermore, the characteristic of our sign language dictionary is very specific, as shown in Table 1. So, we proposed a gradational match and length first algorithm (GMALF algorithm).

Suppose W is a sequence of n characters in a given field (The method to obtain field is mentioned in **3.1**) : $W = C_1C_2C_3 \cdots C_n$.

Step 1: Get one character C_i from field W (when the first time, i=1). Match the second layer of reconstructed BWDIC and check whether C_i exists or not. If not exists go to Step2, otherwise go to Step 3.

Step 2: Match C_i in the FWDIC, records C_i as a word and the basic word of Chinese sign language of word C_i , then go to Step 4.

Step 3: Match the words in sub-tree of C_i separately. The principle is that the length has priority. If exist, record it and the basic word of Chinese sign language of it, then return to Step4. Otherwise, go to Step2.

Step 4: Whether W is null or not. If null, the filed of W is finished and obtains the other fields. Otherwise, adjust the value of i and go to Step1.

Proposed FMM Algorithm. The proposed FMM method and proposed MM method are very similar in principle, so we no longer go into details here. The description of proposed FMM algorithm can refer to the MM method above.

3.3 Recognition of Ambiguity and Disambiguation

Recognition of Ambiguity. This paper discerns the ambiguous field by bidirectional scanning. We segment the field by the MM method and FMM method separately. If the two segmentation results are different, then this field is regarded ambiguous.

Disambiguation. Both uni-gram of characters and MI are applied to achieve disambiguation in this paper.

Suppose W is an ambiguous field of overlap type: W： $a_1a_2 \cdots a_m b_1b_2 \cdots b_t w_1w_2 \cdots w_n$, which has two different segmentations:

F-Seg： $\backslash\, a_1a_2 \cdots a_m b_1b_2 \cdots b_t \,\backslash\, w_1w_2 \cdots w_n \,\backslash$.

B-Seg： $\backslash\, a_1a_2 \cdots a_m \,\backslash\, b_1b_2 \cdots b_t w_1w_2 \cdots w_n \,\backslash$.

N-gram in Target Text Space. In the target space, we generate all possible character sequences and statistic the frequency of occurrence of them from the target text.

From formula (3), the forward segmentation can be represented as equation 8.

$$P_{forward}(W) = P(a_1 a_2 \cdots a_m b_1 b_2 \cdots b_t) P(w_1 w_2 \cdots w_n) \quad (8)$$

The backward segmentation can be represented as equation 9.

$$P_{backward}(W) = P(a_1 a_2 \cdots a_m) P(b_1 b_2 \cdots b_t w_1 w_2 \cdots w_n) \quad (9)$$

Then:

If $P_{forward}(W) > P_{backward}(W)$, select F-Seg; otherwise select B-Seg.

MI in Corpus Space. In the corpus space, the occurrences of the words will be considered. For all the ambiguous fields, we compute their MI scores in the corpus space. Here N is the total number of words in corpus and the $freq_{corpus}(c)$ is estimated by the times of word c appears in the corpus.

From formula (7), the forward segmentation can be represented as equation 10.

$$F_{mi-forward}(b_t w_1) = \log_2 \left(\frac{N \times freq_{corpus}(b_t w_1)}{freq_{corpus}(b_t) \times freq_{corpus}(w_1)} \right) \quad (10)$$

The backward segmentation can be represented as equation 11.

$$F_{mi-backward}(a_m b_1) = \log_2 \left(\frac{N \times freq_{corpus}(a_m b_1)}{freq_{corpus}(a_m) \times freq_{corpus}(b_1)} \right) \quad (11)$$

Then:

If $F_{mi-forward}(b_t w_1) \geq F_{mi-backward}(a_m b_1)$, select B-Seg.

If $F_{mi-forward}(b_t w_1) < F_{mi-backward}(a_m b_1)$, select F-Seg.

The Algorithm of Disambiguation. The method MI is sensitive to data sparseness, so it is not suitable to sparse data set. In this paper, if $freq_{corpus}(a_m b_1)$ or $freq_{corpus}(b_t w_1)$ is less than the threshold ($\delta_{threshold}$), the uni-gram method is employed. Coefficients α and β in the model can be calculated from the target text.

Then:

If $freq(b_t w_1) > \delta_{threshold}$ and $freq(a_m b_1) > \delta_{threshold}$ then:

If $\alpha F_{mi-forward}(b_t w_1) < \beta F_{mi-backward}(a_m b_1)$, select F-Seg.

If $\alpha F_{mi-forward}(b_t w_1) \geq \beta F_{mi-backward}(a_m b_1)$, select B-Seg.

Else, then:

If $P_{forward}(W) > P_{backward}(W)$, select F-Seg.

If $P_{forward}(W) \leq P_{backward}(W)$, select B-Seg.

Where α is the times of occurrence for $b_1 b_2 \cdots b_t w_1 w_2 \cdots w_n$ and β indicates the times of occurrence for $a_1 a_2 \cdots a_m b_1 b_2 \cdots b_t$ in the target text.

4 System Structure and Implementation

The whole system includes several basic components, Web text download component, Web text parse component, text pre-processing component, text scanning component, word segmentation component, recognition of ambiguity component, disambiguation component and result output component. It parses the Web text, executes the task of segmentation and then translates natural language into Chinese sign language, as illustrated in Fig.2.

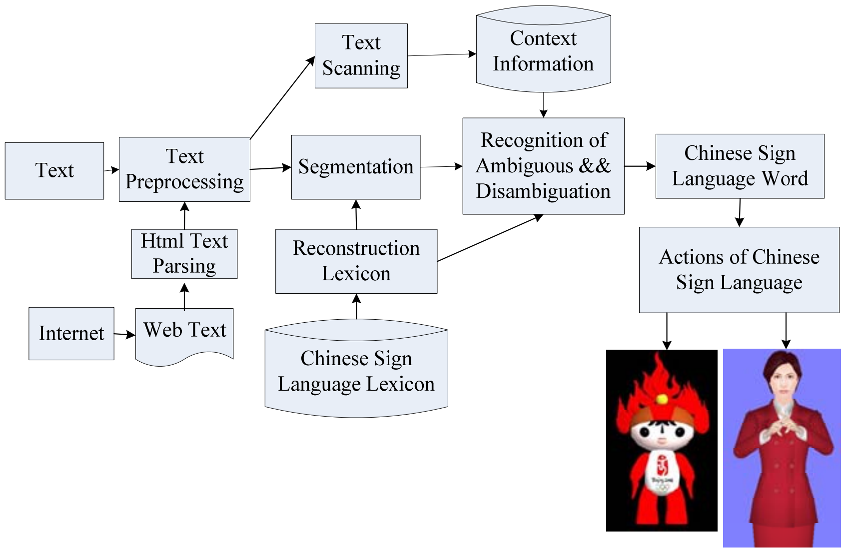

Fig. 2. The structure of WS system

5 Experiment Results

We design and implement the WS system for Chinese sign language, and test in the news of campus network of Beijing University of Technology and corpus of Beijing University. The results are as follows:

The segmentation results of news of the campus network:

Table 2. Segmentation results of news

File Size(KB)	Number of Word	Number of Sign Language Word	Time(Second)
28KB	631	472	0.1 Second
60KB	13655	9621	1.3 Second

The performance of disambiguation is shown in Table 3, where NSD stands for Number of Successful Disambiguation.

Table 3. Segmentation results of ambiguous fields

Category	Ambiguous Fields	MM Method		FMM Method		This Paper	
		NSD	Precision	NSD	Precision	NSD	Precision
News	381	165	43.3%	216	56.7%	335	87.9%
Sports	31	13	41.9%	18	58.1%	26	83.9%
Others	25	12	48.0%	13	52.0%	21	84.0%

6 Conclusions

This paper has discussed word segmentation of Chinese sign language, and how to transform natural language into Chinese sign language, involving word segmentation, recognition of ambiguous fields, disambiguation etc.. The experiment results show that significant improvement in performance of segmentation has been achieved compared to conventional methods of word segmentation. For further work, we need to focus on translating natural language into Chinese sign language more accurately with following aspects: (1) Improve the algorithm of recognition of ambiguity and disambiguation. (2) Expand the dictionary of Chinese sign language, in order to improve the accuracy of segmentation. 3) Design and implement the recognition of China NER (Chinese Named Entity Recognition), in order to eliminate the peculiar ambiguity caused by names.

References:

1. Nie, J., Briscbois, M., Ren, X.: On Chinese text retrieval. In: Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM Press (1996) 225–233.
2. Lua, K., Gan, G.: An application of information theory in Chinese word segmentation. Computer Processing of Chinese and Oriental Languages (1994) 115–124.
3. S.Foo, H.Li.: Chinese word segmentation and its effects on information retrieval. Information Processing and Management 40 (2004) 161–190.
4. Cheng-Ning Huang. What can the Statistical modeling do?. Applied Linguistics, 2002(1):77-84.
5. Shi-Wen Yu, Hui-Ming Duan, Xue-Feng Zhu, Bin Sun. Fundamental Rules Constituted by Beijing University for the Modem Chinese Linguistic Corpus processing, Journal of Chinese Information Processing, 2002,16 ,(5):49-65.
6. Yuan-Yuan Wang, Zhong-Shi He．A new approach to Chinese person names recognition based part-of speech detecting. Proceedings of the Third International Conference on Machine Learning and Cybernetics．2004:2969-2972.
7. Jingfang Xu, Shaozhi Ye, and Xing Li. Query Based Chinese Phrase Extraction for Site Search. Lecture Notes in Computer Science, 2004: 125-134.
8. Shimohata, S., Sugio, T.: Retrieving collocations by co-occurrences and word order constraints. In: Proceedings of the eighth Conference on European Chapter of the Association for Computational Linguistics. (1997) 476–481.
9. De-Gen Chang, Yuan-Sheng Yang, Sheng Wang,Yan-Li Zhang, Wan-Xie Zhong．Chinese Person Names Recognition Algorithm based on Statistics．Journal of Chinese Information Processing.01,15(2):31-44.