ERROR ANALYSIS AND CONFIDENCE MEASURE OF CHINESE WORD SEGMENTATION

Chih-Chung Kuo and Kun-Yuan Ma

Computer & Communication Research Laboratories, Industrial Technology Research Institute
E000, Bldg.51, 195-11 Sec.4 Chung Hsing Rd.,
Chutung, Hsinchu, Taiwan 310, R.O.C.

ABSTRACT

Word segmentation for a Chinese sentence is essential for many applications in language and speech processing. There's no perfect method that could achieve word segmentation without any errors. We propose a confidence measure for the segmentation result to cope with the problem caused by the errors. The effective method depends mainly on the error analysis of the word segmentation. With the confidence measure the suspected errors can be identified such that manual inspection loads can be largely reduced for non-real-time applications. A soft-decision method and a composite-word approach for prosody generation are also designed for text-to-speech systems by exploiting the confidence measure, such that the wrong prosody caused by wrong word boundaries can be alleviated.

1. INTRODUCTION

Each Chinese character itself is also a word. But most useful words are composed of two or three characters. Certainly there are also words composed of more than three characters. A Chinese sentence has no word delimiters, like white space in English, between words. Readers who don't understand Chinese could imagine an English text written without any space characters between words, such as "atopposition". It could be segmented into "a top position ", "at opposition", "at op position", or "atop position". Of course, not all of them are valid or meaningful in English. Similarly, except for some rare cases each Chinese sentence could be segmented into only one word sequence in light of context.

Although word segmentation is not a problem to a human reader, it is a not-trivial job for computers. This is because the possible word segmentation could be very large in number and could be highly ambiguous. Figure 1 shows an extreme case in which a sentence could be segmented into 22400 possible word sequences. The word segmentation is essential in the stage of text analysis for Chinese text-to-speech (TTS) system. The stages following text analysis, for example, the prosody generation and waveform synthesis are strongly affected by the result of word segmentation. There are also many other applications need word segmentation, such as corpora establishment for research in natural language and speech processing.

Many methods have been proposed to accomplish word segmentation automatically [1][2]. A lexicon defining words is necessary for every method. No matter what method is used, there's no perfect method that could achieve word segmentation without any errors. The errors could cause wrong prosody or even wrong enunciation in a TTS system, or could produce an erroneous corpus, which in turn result in wrong statistic data for many applications.

When the job is not real-time, such as corpora construction, the word segmentation error could be corrected by manual inspection, which however is labor intensive, time consuming, and error-prone. When the application is real-time, such as on-line TTS system, the errors are just ignored. In this paper we propose a method to cope with this problem. Instead of designing a new method for word segmentation, we propose a confidence measure for the segmentation result. The effective confidence measure depends mainly on the error analysis of the word segmentation. With the confidence measure the suspected segmentation errors can be identified such that manual inspection loads can be largely reduced for non-real-time applications. A soft-decision method and a composite word approach for prosody generation are also designed for TTS systems by exploiting the confidence measure, such that the wrong prosody caused by wrong word boundaries can be alleviated.

2. DESIGN OF CONFIDENCE MEASURE

With any input Chinese text the word segmentation method would generate one best word sequence which will be denoted as *word segmentation result* (WSR) in the following discussion for convenience. The confidence measure can be designed by the following general procedure:

1. Produce the other competitors of the WSR.

Sentence: "經證管會核准辦理公開發行公司財務查核簽證的聯合會計師出具保留意見"

- 經|證管會|核准|辦理|公開|發行|公司|財務|查核|簽證|的|聯合|會計師|出|具|保留|意見
- 經|證管會|核准|辦|理|公開|發行|公司|財務|查核|簽證|的|聯合|會計師|出|具|保留|意見
- 經|證管會|核准|辦理|公開|發行|公司|財務|查|核|簽證|的|聯合|會計師|出|具|保留|意見
- 經|證管會|核准|辦理|公|開發|行|公司|財務|查核|簽證|的|聯合|會計師|出|具|保留|意見

....

Figure 1: Some examples of the 22400 possible word sequences of a Chinese sentence.

2. Determine *ambiguity grade* (AG) of the WSR in comparison with its competitors.
3. Compute or determine the confidence measure for the WSR according to the AG.

Selecting the word sequences that are different from the WSR only in few places could generate the competitors of the WSR. The competitors can also be selected according to scores if the WSR is generated through scoring each possible word sequence and choosing the one with top score. The procedure shown above is a general concept, which can be realized by various implementations. It is also related to the used method for word segmentation.

2.1. Word Segmentation Method

Assume the input character sequence, $C = \{c_1 c_2 \cdots c_N\}$, is segmented into the word sequence, $W = \{w_1 w_2 \cdots w_M\}$, with the correspondent POS (Part Of Speech) tag sequence, $T = \{t_1 t_2 \cdots t_M\}$. Then the score, $S(W,T)$, is given as:

$$\begin{aligned} S(W,T) = & L^2(w_1) + \eta_1 \log P(w_1) + \eta_2 \log P(t_1) \\ & + \sum_{i=2}^{M} \left[L^2(w_i) + \eta_1 \log P(w_i) + \eta_2 \log P(t_i | t_{i-1}) \right] \end{aligned} \qquad (1)$$

where we need the information as follows:

w_i and t_i: Chinese word and its POS tag stored in a lexicon,
$L(\mathrm{w_i})$: word length of w_i,
$P(\mathrm{w_i})$: occurring probability of the word w_i,
$P(t_i)$: occurring probability of the POS t_i,
$P(t_i | t_{i-1})$: transitional probability of the POS,
η_1, η_2: weighting factors.

The score is defined based on the rule that longer word is preferred and on statistics of word, POS, and POS transition. Through Viterbi search and back tracing, the word sequence and POS tagging with top score could be found efficiently [3].

2.2. Difference Type and Ambiguity Grade

The word segmentation difference between the WSR and a competitor can be classified into three types as follows:

- Type 1 (Over-Segmented): one word in the WSR is segmented into multiple words in the competitor.
- Type 2 (Miss-Combined): Multiple words in the WSR are combined into one word in the competitor.
- Type 3 (Mismatch): Multiple words in the WSR are combined and re-segmented into different multiple words in the competitor.

Figure 2 shows an example sentence, which contains all the three types of difference between the competitors and the WSR. Notice that the competitor 3 (Cmp3) contains both type 2 and type 3 differences.

WSR: 高|價|採用|公司|*出品*|一| 高|價|採用|公司|
Cmp1: 高價|採價|公司|出品|一| 高|價|採用|公司| Type 2
Cmp2: 高|價|採價|公司|出品|一| 高|價採|用|公司| Type 3
Cmp3: 高價|採價|公司|出品|一| 高|價採|用|公司| Type 2,3
Cmp4: 高|價|採價|公司|*出|品*|一| 高|價|採用|公司| Type 1

Figure 2: Examples of the three types of word segmentation difference.

After analysis and statistics of segmentation errors, we found that the degree of ambiguity in word segmentation depends on the difference type. Therefore we define the ambiguity grade (AG) as follows:

- AG0: no competitor.
- AG1: competitors contain only type 1 difference.
- AG2: competitors contain type 3 difference, but no competitor contains type 2 difference.
  - — AG2-1: only one competitor contains type 3 difference.
  - — AG2-2: more than one competitor contains type 3 difference.
- AG3: competitors contain type 2 difference.

The higher the grade, the more uncertain the WSR is. The AG0 contains no any ambiguity because only one word sequence exists, i.e., only one-character words exist in the sentence.

2.3. Determining Confidence Measure

Confidence measure could be determined according to ambiguity grade, and then the confidence measure would be a discrete function with finite and discontinuous values. It might be further modified into a continuous function with the help of other information like relative score between the WSR and its competitors. No matter what function is designed, the real value of confidence measure comes from statistics over a specific database using a specific word segmentation method. Some information of the training database used in our experiment are listed as follows:

- total count of characters: 35226
- total count of words: 23346
- total count of sentences: 3779
- average characters per sentence: 9.3215
- average words per sentence: 6.1778
- average characters per word: 1.5089
- average candidates of word segmentation per sentence: 64.8500

Table 1 and Figure 3 show the result of our experiment. The confidence measure is defined as the correct rate of word segmentation for each ambiguity grade in this case. It's clear from Figure 3 that the definition of confidence measure in this experiment is useful because the less confident cases (with AG greater than 1) seldom occur.

AG	0	1	2-1	2-2	3	Sum
Correct	263	3005	348	47	43	3706
Error	0	3	26	10	34	73
Total	263	3008	374	57	77	3779
Occurring Rate (%)	7.48	85.55	10.64	1.62	2.19	100.00
Confidence Measure (%)	100.00	99.90	93.05	82.46	55.84	

Table 1: Confidence measure according to ambiguity grade (AG) and the statistics of the training database

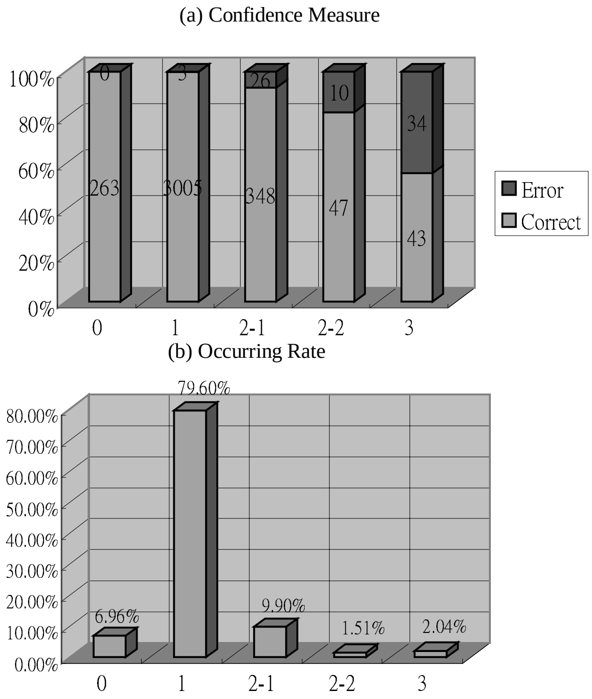

Figure 3: (a) Confidence measure and (b) occurring rate based on the statistics of the training database.

3. APPLICATIONS

3.1. Real-Time Application

Soft-Decision Method. Former design of Chinese TTS system use word segmentation to determine word boundary, which in turn generates proper prosody information for synthesis. Because the word segmentation is a hard-decision, so is the generated prosody. This produces an unacceptable synthesized prosody when the word segmentation is wrong. With the help of confidence measure, a soft-decision could be made with the word segmentation. The prosody can be produced by linearly combining different prosody parameters due to different word segmentation. Weighting of the linear combination is based on the confidence measure.

Take for example the pause between characters, which is one of the important prosody parameters. The pause parameter of a character can be produced as follows:

$$p = m \cdot p_R + \frac{1-m}{K}\sum_{k=1}^{K} q_k \qquad (2)$$

where

p : final pause parameter

p_R : pause parameter of WSR

q_k : pause parameter of k-th competitor that is different in word segmentation at this character boundary.

m : confidence measure.

If the pause took only two values: a large value for word boundary and a small value for the intra-word character boundary, the equation (2) would be simplified as a linear combination of two terms.

Composite-Word Approach. Another robust prosody generation approach is even simpler. Any character sequence with high ambiguity could be combined as a composite word. In our TTS system, the prosody generator is a recurrent neural network (RNN) [3], which can map linguistic parameters into acoustic parameters for each character. The input to the RNN includes information about word boundary and character order within a word. Therefore, a composite word would generate a smoother prosody so as to avoid improper prosody when the word segmentation is wrong.

The CDROM proceedings contain two sound files that demonstrate the effect of this approach. The first file, [SOUND 1078_01.WAV], is synthesized with poor prosody due to wrong word segmentation (wrong: 讓｜高｜消｜受不了； correct: 讓｜高｜消受｜不了). The second file, [SOUND 1078_02.WAV], is synthesized based on composite word approach (讓｜高｜消受不了), which obviously produces more correct and natural prosody.

3.2. Non-Real-Time Application

Referring to Table 1, if we inspect only those sentences with AG greater than 1, the amount of loading will shrink into only 14.45% at the cost of 3 misses, which is 0.079%. If only the sentences with AG greater than 2-1 are inspected, the loading will be further reduced into 3.81%, but the missed error will also increase by 3 to 29, which is 0.767%. Therefore, given a tolerable cost (error percentage), a threshold of confidence measure can be defined, which could largely reduce the amount of sentences which need manual inspection.

To verify the usefulness of this mechanism, we randomly selected 2085 sentences from newspaper to form a testing database and process it with our word segmentation method. We first let a person check the word segmentation result of all 2085 sentences, and also correct any found errors. Then a computer-aided method is used by another person to check those sentences with AG greater than 1 and to correct any found errors by selecting the correct one from few competitors. Computer also marks the different places of the competitors from the WSR. This will help the user to concentrate only on the different points so as to accelerate the inspection process. The number of found and corrected errors and the consumed time are summarized in Table 2. We found that with computer-aided method not only the time is saved but also the accuracy is increased. The manual method missed 21 errors and made 2 mistakes in correcting errors.

Method	Checked Sentences	Consumed Time	Found Errors	Corrected Errors
manual	2085	> 8 hour	73	71
computer-aided	330	< 50 min.	94	94

Table 2: Comparison of word segmentation inspection between manual method and computer-aided (with confidence measure) method.

For reference purpose, the statistic detail of the testing database is given in Table 3 and illustrated in Figure 4. The trend in training and testing database is the same. The confidence measure for high ambiguity grades descends to far lower values in the testing database than that in the training database. This difference is caused by insufficient data.

AG	0	1	2-1	2-2	3	Sum
Correct	136	1619	198	35	3	1991
Error	0	0	30	37	27	94
Total	136	1619	228	72	30	2085
Occurring Rate (%)	6.52	77.65	10.94	3.45	1.44	100.00
Confidence Measure (%)	100.00	100.00	86.84	48.61	10.00	

Table 3: Confidence measure according to ambiguity grade (AG) and the statistics of the testing database

4. CONCLUSIONS

There is no word segmentation method that can segment all sentences without any error. The segmentation error would cause unaccepted effect in some applications. We propose the concept of confidence measure and realize it by analyzing the relation of difference type to error. By defining three types of word segmentation difference and several ambiguity grades, the confidence measure could be effectively determined.

Applications of the confidence measure for word segmentation were also presented. Two new prosody generation methods in cooperation with the confidence measure were proposed. A practical experiment on constructing a corpus with the help of the confidence measure and computer-aided display/correction was conducted. The result showed the usefulness of the methods presented in this paper.

5. ACKNOWLEDGEMENT

This paper is a partial result of the project No.3P11100 conducted by the CCL/ITRI under sponsorship of the Ministry of Economic Affairs, Taiwan, R.O.C. The authors also like to thank their colleagues, Dr. Chiang Tung-Hui and Dr. Chang Sen-Chia for their kindly help and illuminating discussion.

6. REFERENCES

1. Tung-Hui Chiang, Ming-Yu Lin, and Keh-Yih Su, “Statistical Models for Word Segmentation and Unknown Word Resolution,” *Proceedings of 1992 R.O.C. Computational Linguistics Conference (ROCLING V)*, pp.121—146, Taipei, Taiwan, 1992.

2. Ming-Yu Lin, Tung-Hui Chiang and Keh-Yih Su, “A Preliminary Study on Unknown Word Problem in Chinese Word Segmentation,” *Proceedings of 1993 R.O.C. Computational Linguistics Conference (ROCLING VI)*, pp.119—141, Taipei, Taiwan, 1993.

3. Chih-Chung Kuo, “A Chinese Text-to-Speech System with Text Preprocessing and Confidence Measure for Practical Usage,” *Proceeding of TENCON’97*, Brisbane, Australia, Dec. 2-4, 1997.

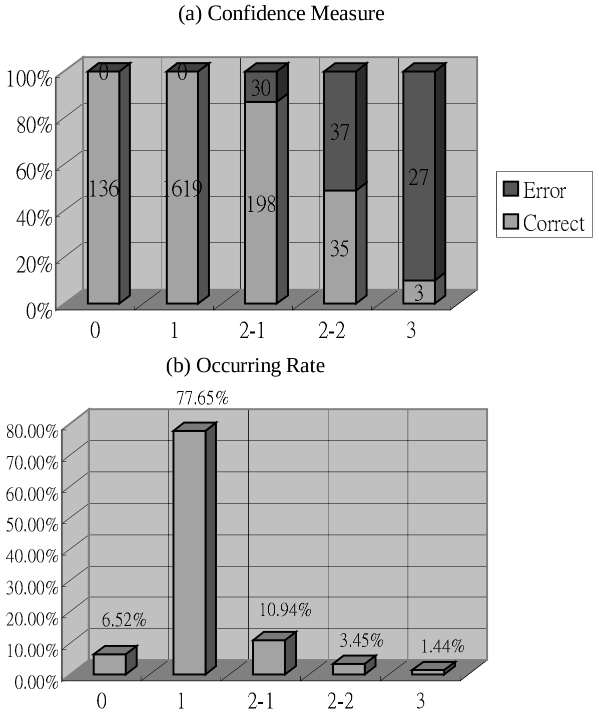

Figure 4: (a) Confidence measure and (b) occurring rate based on the statistics of the testing database.