

## A Corpus-based Chinese Syllable-to-Character System

CHIEN-PANG WANG AND TYNE LIANG\*

*Department of Computer and Information Science,*

*National Chiao Tung University,*

*1001 Ta Hsueh Rd., Hsinchu Taiwan 30050 R.O.C.*

*\*:responsible for all correspondences.*

**Keywords:** *Mandarin, Phoneme, Homonym, Tolerance.*

### Abstract

One of the popular input systems is based on Chinese phonetic symbols. Designing such kind of a syllable-to-character (STC) input system involves two major issues, namely, fault tolerance handling and homonym resolution. In this paper, the fault tolerance mechanism is constructed on the basis of a user-defined confusing set and a modified bucket indexing scheme is incorporated so as to satisfy real-time requirement. Meanwhile the homonym resolution is handled by binding force and heuristic selection rules. Both the system performance and tolerance ability are justified with real corpus in terms of searching speed and character conversion accuracy rate. Experimental results show that the proposed scheme can achieve 93.54% accuracy for zero-error syllable inputs and 80.13% for zero-tone syllable inputs. Furthermore both robustness and tolerance of the proposed system are proved for high input error rates.

## 1 Introduction

Among various kinds of Chinese input methods, the most popular one is based on phonetic symbols. This is because most of Chinese-speaking users are taught to use phonetic symbols in their elementary schools when they learn Chinese. However a syllable-to-character (STC) system is inherently associated with the serious homonym and similarly-pronounced phoneme problems. This is because a single syllable may correspond to several Chinese characters and there are indeed several Mandarin syllables which are sounded similarly. So it is not easy for users or acoustic recognizer to distinguish them when they are used. We call these syllables as confusing syllables. For example, syllable 尸 \ (shih4) and ㄥ \ (szu4) are sounded similarly in speaking and listening, and a user might treat 尸 \ (shih4) as ㄥ \ (szu4) at typing or pronouncing. Thus robust fault tolerance ability of a STC system has to be concerned so as to improve the phoneme-to-character conversion accuracy.

In recent years, various approaches have been proposed to construct a Chinese STC system either for speech input or keyboard input. For speech input, Chang [1994] used vector quantization to cluster words into classes when training Hidden Markov model so that words in the same class share the model's parameters. Contrast to the character N-gram based Markov model, a word N-gram based Markov model was proposed by Yang [1998]. Though Markov-based models are easy for implementation, they require large training corpus and large storage for large numbers of parameters. Furthermore, the parameters of Markov model are needed to be fixed, so they reflect the characters of training corpus only. Rather than using Markov model, Lin [1995] used mutual information to find the relation between base syllables and applied Heuristic Divide-and-Conquer Maximum Match (H-DCMM) Algorithm to detect prosodic-segment in a sentence. To train the robustness of prosodic-segment detection, a segmental K-means algorithm is also used.

As for syllable-based keyboard input, Gie [1990] used a hand-crafted dictionary for matching syllables of phrases and a set of impression rules for homonym selection. In Gie [1991], homonyms for new phrases are furtherly dealt by using a dictionary and occurrence frequencies. On the other hand, Lai [2000] used maximum likelihood ratio and good-tuning estimation to handle characters with multiple syllables. Lin [2002] combined N-gram model and selection rules for dealing with multiple PingIn codes. Unlike statistical approaches, context sensitive method proposed by Hsu [1995] was applied in a Chinese STC system called “Going.” The system relies heavily on semantic pattern matching which can reduce the huge amount of data processing required for homophonic character selection. The conversion accuracy rate is close to 96%. In [Tsai and Hsu 2002], a semantically-oriented approach was also presented by using both noun-verb event-frame word-pairs and statistical calculation. The experimental results show that their overall syllable-to-word accuracy can be 96.5%.

In this paper a corpus-based STC system to support high tolerance is presented and it can be used as a keyboard input method as well as a post-processor incorporated with an acoustic system. To support high tolerance ability, we used a bucket-based searching mechanism so that the searching time of confusing syllable is reduced. The presented homonym resolution is based on binding force information and selection rules. Various tests are implemented to justify the system performance. In zero-tolerance test, our character conversion accuracy is 93.54% out of 1052 characters. For zero-tone testing, the character conversion accuracy is 80.13%. In input syllables with 20% and 40% confusing set member replacement, the character conversion accuracy is 83.08% and 78.23% respectively. The feasibility and robustness of fault tolerance handling to a STC system are also proved by the experiments.

The outline of the paper is as follows. Section 2 introduces the preliminary

background of Chinese syllable structure. Section 3 describes the system architecture and section 4 presents the proposed searching mechanism. Section 5 explains our selection module and section 6 reports various experimental tests. Finally Section 7 gives the conclusion.

## 2 Mandarin syllable and Confusing set

### 2.1 Sets of syllables in Mandarin

According to [Wu 1998], a general Mandarin syllable structure contains four parts consonant, head of diphthong, vowel and tone. There are twenty-one consonants, sixteen vowels, and five tones. Since users usually pronounce head of diphthong and vowel simultaneously, so the syllable structure can be simplified to combine head of diphthong and vowel such as ㄨ and ㄨㄛ [Chen 1998].

Table 1 is the list of consonants, vowels, tones and the code number in our system. In this paper, we treat the syllable with tone=0 as tone=1. Because the amount of the syllables with tone=0 is quite few (19 out of 1302 Mandarin syllables), and their corresponding characters are few too (29 out of 14105 unique Mandarin characters).

Table 1: Consonants and vowels.

#### (a) Consonants

01	Nil	02	ㄅ	03	ㄆ	04	ㄇ	05	ㄏ
06	ㄎ	07	ㄉ	08	ㄊ	09	ㄋ	10	ㄌ
11	ㄍ	12	ㄆ	13	ㄇ	14	ㄏ	15	ㄎ
16	ㄉ	17	ㄊ	18	ㄋ	19	ㄌ	20	ㄍ
21	ㄆ	22	ㄇ						

#### (b) Vowels

01	Nil	02	ㄚ	03	ㄛ	04	ㄜ	05	ㄝ
06	ㄞ	07	ㄟ	08	ㄠ	09	ㄡ	10	ㄢ
11	ㄣ	12	ㄤ	13	ㄨ	14	ㄨㄛ	15	ㄨㄜ
16	ㄨㄝ	17	ㄨㄞ	18	ㄨㄟ	19	ㄨㄠ	20	ㄨㄡ
21	ㄨㄢ	22	ㄨㄣ	23	ㄨㄤ	24	ㄨㄨ	25	ㄨㄨㄛ
26	ㄨㄨㄜ	27	ㄨㄨㄝ	28	ㄨㄨㄞ	29	ㄨㄨㄟ	30	ㄨㄨㄠ
31	ㄨㄨㄡ	32	ㄨㄨㄢ	33	ㄨㄨㄣ	34	ㄨㄨㄤ	35	ㄨㄨㄨ
36	ㄨㄨㄨㄛ	37	ㄨㄨㄨㄜ	38	ㄨㄨㄨㄝ	39	ㄨㄨㄨㄞ		

### (c) Tones

	Nil	/	∨	∖
1	1	2	3	4

### 2.2 Confusing set

The confusing sets are the groups of syllables, which are recognized to be the same by the human or the acoustic recognizer. For example, ㄟㄢ (fei1) and ㄏㄨㄟ (hui1) are confusing syllables for many Chinese-speaking people in Taiwan.

Suppose Table 2 is the statistical results from an acoustic recognizer. Then the confusing sets of phonemes can be found by using the find-connected-components algorithm [Thomas 1998] in which phonemes are vertices of a graph and the confusing sets are those edges whose recognition probabilities are greater than a threshold. For example, two confusing sets of phonemes, {ㄨㄟ} and {ㄟ, ㄏ} are generated from Table 2 when their probabilities are greater than a given threshold at 25%.

Table 2: An example of acoustic data.

Phoneme	Result	Prob.	Result	Prob.	Result	Prob.
ㄨㄟ	ㄨㄟ	0.75	ㄟ	0.2	ㄏ	0.05
ㄟ	ㄟ	0.6	ㄏ	0.4		
ㄏ	ㄏ	0.7	ㄟ	0.3		

### 2.3 Bucket of confusing set

The confusing sets of syllable are obtained by using Cartesian product on two confusing sets of consonants and vowels, (an example shown in Table 3). Then a bucket  $B(\alpha \beta)$  will contain the grams from  $C(\alpha)$  of consonant confusing set and  $V(\beta)$  of vowel confusing set. Fig. 1 is an example of bucket of bigram syllable confusing set, and its corresponding bucket is  $B(08140607)$ .

Table 3: An Example of confusing sets.

**(a) Confusing sets of consonant**

C(01)	Nil, ㄇ, ㄍ, ㄈ	C(04)	ㄅ, ㄆ, ㄇ	C(07)	ㄐ, ㄑ
C(02)	ㄏ, ㄏ	C(05)	ㄌ, ㄎ	C(08)	ㄒ, ㄓ
C(03)	ㄉ, ㄊ	C(06)	ㄍ, ㄎ, ㄎ	C(09)	ㄑ, ㄒ

**(b) Confusing sets of vowel**

V(01)	Nil	V(06)	ㄩ, ㄩ, ㄩ	V(11)	ㄨ, ㄨ, ㄨ
V(02)	ㄜ, ㄜ	V(07)	ㄣ, ㄣ	V(12)	ㄣ, ㄣ, ㄣ
V(03)	ㄝ, ㄝ	V(08)	ㄤ, ㄤ	V(13)	ㄤ, ㄤ, ㄤ
V(04)	ㄞ, ㄞ	V(09)	ㄤ, ㄤ	V(14)	ㄤ, ㄤ, ㄤ
V(05)	ㄟ, ㄟ	V(10)	ㄤ, ㄤ	V(15)	ㄤ, ㄤ, ㄤ

C(08)	V(14)	×	C(06)	V(07)	=	抽獎, 衝向, 抽象, ...
ㄒ, ㄓ	ㄤ, ㄤ, ㄤ, ㄤ		ㄍ, ㄎ, ㄎ	ㄣ, ㄣ		

Fig. 1: a bucket example of  $B(08140607)$ .

**3 System Architecture**

Fig. 2 shows the system flowchart containing foreground process and background process. In the background process, we used news documents collected from the Chinatimes website (<http://news.chinatimes.com/>) in March 2001 and segmented these documents into grams. Next, the Mandarin syllables for each gram were generated by our syllable generation method. We also encoded the grams by its confusing set which is obtained from acoustic statistic data. Then grams with confusing set information are stored into gram database.

The foreground process consists of fault tolerance matching module and selection module. Fault tolerance matching module encodes the phonetic symbol sequence and searches the corresponding grams that have minimum error distance with phonetic symbol sequence in the corpus database. Then corresponding unigrams, bigrams, and trigrams will be searched and passed into selection module. Selection module is constructed on the basis of selection rules to decide the output gram. The binding force information is calculation is done with CKIP word database contains 78,410 Mandarin words and their corresponding syllables. Finally, the output gram

will replace the characters of character sequence from the tail.

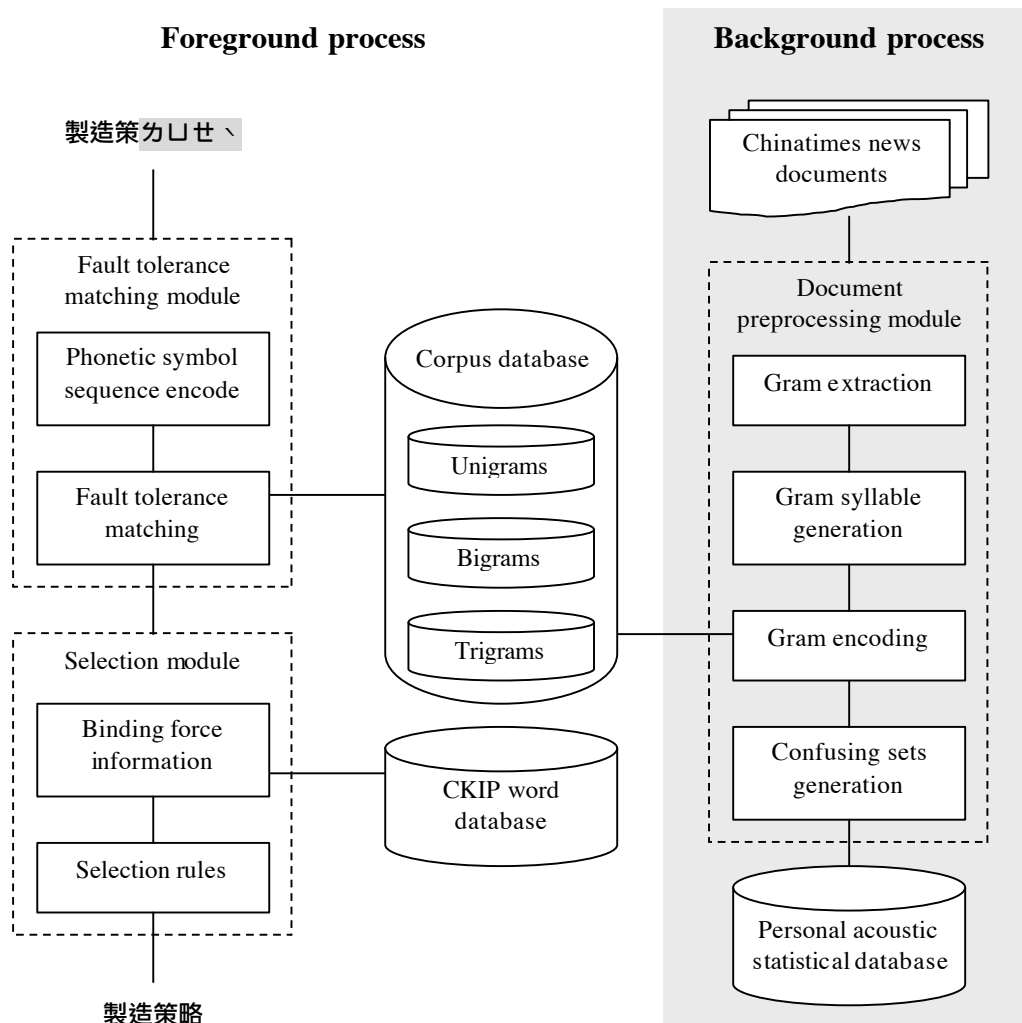


Fig. 2: The system architecture.

## 4 Fault Tolerance Matching Module

### 4.1 Base syllable distance

Let  $N_C$ ,  $N_V$  denote a confusing set number for consonant/vowel confusing set respectively. We define base syllable to be a syllable without tone. Then a bucket  $B(N_CN_V)$  will contain those grams having the syllable confusing set  $N_CN_V$ .

A base syllable distance is the number of different consonant or vowel confusing set pairs between two base syllables. Suppose a base syllable sequence

$SylSeq1=c_1v_1c_2v_2c_3v_3$  which belongs to  $B(N_{C1}N_{V1}N_{C2}N_{V2}N_{C3}N_{V3})$ , and another syllable sequence  $SylSeq2=c_1'v_1'c_2'v_2'c_3'v_3'$  which belongs to  $B(N_{C1}'N_{V1}'N_{C2}'N_{V2}'N_{C3}'N_{V3}')$ .  $SylSeq2$  has two base syllable distance from  $SylSeq1$  if there exists any two mismatch pairs of consonant or vowel confusing sets, like  $N_{C1} \neq N_{C1}'$  and  $N_{V2} \neq N_{V2}'$ . Similarly, there will be  $K$ -distance if there are  $K$  mismatch pairs between  $SylSeq1$  and  $SylSeq2$ .

## 4.2 Bucket index

To find the grams with minimum base syllable distance from a given gram, we start to find the bucket first which the grams belong to. Our searching is done with the string matching algorithm proposed by Du and Chang [1994]. We start from the buckets with zero syllable distance. If there is no such gram in these buckets, we increase base syllable distance by 1. The maximum distance is defined to be 2 in this paper. We use index structure to memorize these buckets for every base syllable distance.

Let  $[\gamma, \delta]_{(\phi, \omega)}$  denote a extension bucket index. Symbol  $\gamma$  and  $\delta$  are the buckets whose errors are at any position except  $\gamma$  and  $\delta$ ; symbol  $\phi$  is the base syllable distance and  $\phi \in \{1, 2\}$ ; symbol  $\omega$  represents bigram ( $\omega=2$ ) or trigram bucket index ( $\omega=3$ ). For example, extension bucket index  $[1,2]_{(1,3)}$  is a trigram index with one base syllable distance, and contains the buckets whose errors are at any position except the first and second ones. Therefore,  $[1,2]_{(1,3)}$  contains the following buckets:  $B(O_1O_2XO_4O_5O_6)$ ,  $B(O_1O_2O_3XO_5O_6)$ ,  $B(O_1O_2O_3O_4XO_6)$ , and  $B(O_1O_2O_3O_4O_5X)$  (we use X to indicate error occurrence and O correct one for notation simplification); similarly, extension bucket index  $[5,6]_{(1,3)}$  contains buckets:  $B(XO_2O_3O_4O_5O_6)$ ,  $B(O_1XO_3O_4O_5O_6)$ ,  $B(O_1O_2XO_4O_5O_6)$ , and  $B(O_1O_2O_3XO_5O_6)$ .

In fact extension bucket index  $[1,2]_{(1,3)}$  and  $[5,6]_{(1,3)}$  together will include all buckets with one base syllable distance. The combination of extension bucket index



set which contains all the buckets is called a covering extension bucket index. Similarly, extension bucket index  $[1,2]_{(1,3)}$  and  $[5,6]_{(1,3)}$  are the members of trigram covering extension bucket index with one base syllable distance. Thus, there exists more than one solution in finding covering extension bucket index. In fact, finding the covering extension bucket index is a NP-complete problem [Garey and Johnson 1979]. Since the length of syllable sequence is short and the number of errors is small, it is easy to find the covering extension bucket index. Thus, searching buckets can be done in real time.

## 5 Selection Module

The designed selection module is based on sliding window whose size is set to be five in the proposed system. Let  $C(S_{i-4})$ ,  $C(S_{i-3})$ ,  $C(S_{i-2})$ , and  $C(S_{i-1})$  be the characters in front of  $C(S_i)$  at inputting syllable  $S_i$ . Then the ranking scheme shown as equation (1) is used to rank monograms  $C(S_i)$ , bigrams  $C(S_{i-1})C(S_i)$  and trigrams  $C(S_{i-2})C(S_{i-1})C(S_i)$  which exist in the gram database and each type of the grams with the top values will be treated as our candidate outputs and will be placed at corresponding positions.

$$Rank(g) = \begin{cases} P(g) & \text{if } g \text{ is monogram or trigram} \\ P(g) \times BF(g) & \text{if } g \text{ is bigram} \end{cases} \quad (1)$$

In (Eq. 1)  $p(g)$  is the occurrence probability of  $g$  in the training corpus and the  $BF(g)$  is the binding force for two characters  $C_i, C_{i+1}$  composing bigram  $g$  [Sproat 1990] and it is calculated as following equation:

$$BF(C_i C_{i+1}) = \log_2 \frac{P(C_i C_{i+1})}{P(C_i)P(C_{i+1})} \quad (2)$$

Then selection rules applied to select the candidate grams are as follows:

1. For a trigram candidate  $C(S_{i-2})C(S_{i-1})C(S_i)$ 
  - 1.1. If either  $C(S_{i-4})C(S_{i-3})C(S_{i-2})$  or  $C(S_{i-3})C(S_{i-2})C(S_{i-1})$  exists in gram database, then if it has overlapping  $C(S_{i-1})$  or  $C(S_{i-2})C(S_{i-1})$  with  $C(S_{i-2})C(S_{i-1})C(S_i)$ , then output  $C(S_{i-2})C(S_{i-1})C(S_i)$ , otherwise abort  $C(S_{i-2})C(S_{i-1})C(S_i)$
  - 1.2. If neither  $C(S_{i-4})C(S_{i-3})C(S_{i-2})$  nor  $C(S_{i-3})C(S_{i-2})C(S_{i-1})$  is in trigram database, then
    - 1.2.1 if both  $BF(C(S_{i-3})C(S_{i-2}))$  and  $BF(C(S_{i-2})C(S_{i-1}))$  is less than a threshold, then output  $C(S_{i-2})C(S_{i-1})C(S_i)$
    - 1.2.2 if either  $BF(C(S_{i-3})C(S_{i-2}))$  or  $BF(C(S_{i-2})C(S_{i-1}))$  is greater than a threshold, and there exists overlapping  $C(S_{i-1})$  or  $C(S_{i-2})C(S_{i-1})$  with  $C(S_{i-2})C(S_{i-1})C(S_i)$ , then output  $C(S_{i-2})C(S_{i-1})C(S_i)$ .
    - 1.2.3 if either  $BF(C(S_{i-3})C(S_{i-2}))$  or  $BF(C(S_{i-2})C(S_{i-1}))$  is greater than a threshold but without any overlapping  $C(S_{i-1})$  or  $C(S_{i-2})C(S_{i-1})$  with  $C(S_{i-2})C(S_{i-1})C(S_i)$ , then abort  $C(S_{i-2})C(S_{i-1})C(S_i)$ .
2. If there is no  $C(S_{i-2})C(S_{i-1})C(S_i)$  in database or  $C(S_{i-2})C(S_{i-1})C(S_i)$  is aborted, then
  - 2.1. if  $C(S_{i-3})C(S_{i-2})C(S_{i-1})$  exists in database, then check :

if  $C(S_{i-3})C(S_{i-2})C(S_{i-1})$  has overlapping  $C(S_{i-1})$  with candidate  $C(S_{i-1})C(S_i)$ , then output  $C(S_{i-1})C(S_i)$ , otherwise output the candidate  $C(S_i)$
  - 2.2. if  $C(S_{i-3})C(S_{i-2})C(S_{i-1})$  is not in database but  $C(S_{i-2})C(S_{i-1})$  is, then check:

if  $C(S_{i-2})C(S_{i-1})$  has overlapping  $C(S_{i-1})$  with the candidate  $C(S_{i-1})C(S_i)$  then output  $C(S_{i-1})C(S_i)$ ,

else if  $BF(C(S_{i-2})C(S_{i-1})) < \text{threshold}$  then output candidate  $C(S_{i-1})C(S_i)$ ;

else if  $\text{threshold} < BF(C(S_{i-2})C(S_{i-1})) < BF(C(S_{i-1})C(S_i))$ , then output candidate  $C(S_{i-1})C(S_i)$ ;

else if  $BF(C(S_{i-1})C(S_i)) < BF(C(S_{i-2})C(S_{i-1}))$ , then output candidate  $C(S_i)$ .

## 6 Experimental results

The experiments were implemented to justify the system feasibility and tolerance ability. Our training data includes CKIP word database which contains 78,410 words from length 1 to length 9 and Chinatimes News on the website (<http://news.chinatimes.com/>) containing 6,582 articles in March 2001. The testing data are collected from Chinetimes News on the website containing 7,828 articles in April 2001. The system development and testing environment is Windows 98 on P II 450mHz PC with 128MG Ram.

One experiment is to measure the response time of searching a word in a database. A database without bucket indexing ‘no-bucket’ is compared with ‘bucket<sub>9x15</sub>’ which consists of nine consonant and fifteen vowel confusing sets as listed in Table 3 of Section 2. The searching time of the databases with bucket indexing mechanism is less than one second. Table 4 shows the best case of searching time and there B(50K) means 50K bigrams, T(11K) means 11K trigram and so on.

Table 4: Best case of searching time (seconds)

	B(50K)+T(11K)	B(100K)+T(210K)	B(200K)+T(410K)	B(400K)+T(1350K)
No bucket	0.2	0.77	1.69	15.2
Bucket <sub>9x15</sub>	0.02	0.03	0.03	0.04

Experiments are also implemented for various tolerance tests. There are 100 sentences randomly selected from the testing data and each sentence has 10.5 characters in average. We use two commercial STC systems for comparison, namely Microsoft IME 2002a (微軟新注音輸入法 XP), and Going 6.5 (自然注音輸入法). We compare the accuracy in various tolerance rates which is defined as Eq. 3. In this experiment, we disabled the system-defined confusing phonemes of MS 2002a, because its confusing mechanism and sets are quite different from ours. Table 5 shows the testing results with respect to different the accuracy among four systems.

$$Tolerance\ Rate = \frac{\sum \frac{Number\ of\ character's\ syllable\ replac\ ed\ by\ confusing\ sets\ in\ a\ Sentence}{Number\ of\ characters\ in\ a\ Sentence}}{Total\ Number\ of\ Sentence} \quad (3)$$

Table 5: the character accuracy of 100 testing sentences.

Tolerance rate	0%	20%	30%	40%
9x15	83.94%	83.08%	81.46%	78.23%
Going6.5	94.30%	67.97%	57.80%	45.34%
MS 2002a	94.87%	69.30%	56.18%	43.44%

On the other hand experiments to investigate the correlation between tolerance rate and positions were also implemented. The tolerance position is selected by testing users randomly. Both Table 6 and Table 7 show that the proposed STC system indeed supports robust fault tolerance ability.

Table 6: Character accuracy rate of bucket<sub>9x15</sub> using 30 training sentences.

	Tolerance at Consonant	Tolerance at Vowel	Tolerance at Any Position
Tolerance rate = 20%	91.77	92.89	94.33
Tolerance rate = 35%	89.3	85.76	86.6
Tolerance rate = 45%	86.42	86.27	86.22

Table 7: Character accuracy rate of bucket<sub>9x15</sub> using 30 testing sentences.

	Tolerance at Consonant	Tolerance at Vowel	Tolerance at Any Position
Tolerance rate = 20%	87.34	89.73	85.93
Tolerance rate = 30%	86.4	85.78	85.35
Tolerance rate = 40%	85.57	85.99	83.38

## 7 Conclusions

In this paper a high tolerant STC system useful for traditional Chinese input was presented. The proposed fault tolerance mechanism is constructed on the basis of a user-defined confusing set and a modified bucket indexing scheme is incorporated so as to satisfy real-time requirement. Meanwhile the homonym resolution is handled by

binding force and heuristic selection rules. The performance of the presented system is also justified and compared with various tests. However the drawbacks with the proposed system are its lack of semantic and syntactic checking at output selection. Hence errors like “珊瑚下單(蛋)”, “工作室(是)一種享受” will occur. So how to strengthen the selection module with more linguistic reasoning will be our next step to design an intelligent STC system.

## REFERENCE

- Chang T. Z. 1994. A Word-Class-Based Chinese Language Model and its Adaptation for Mandarin Speech Recognition, *Master Thesis, National Taiwan University*.
- Chen J. T. 1998. Neural Network-based Continuous Mandarin Speech Recognition System, *Master Thesis, National Chiao Tung University*.
- Chinese Knowledge Information Processing Group (CKIP) Corpus 3.0.  
<http://godel.iis.sinica.edu.tw/CKIP/>
- Du M. W. and Chang S. C. 1994. An Approach to Designing Very Fast Approximate String Matching Algorithms, *IEEE Transactions on Knowledge and Data Engineering*, 6, 4, 620-633.
- Garey M. R. and Johnson D. S. 1979. Computers and Intractability. A Guide to the Theory of NP-Completeness, *Freeman, San Francisco*.
- Gie C. X. 1990. A Phonetic Chinese Input System Based on Impression Principle, *Master Thesis, National Taiwan University*.
- Gie T. H. 1991. A Phonetic Input System for Chinese Characters Using A Word Dictionary and Statistics, *Master Thesis, National Taiwan University*.
- Hsu W. L. 1995. Chinese Parsing in a Phoneme-to-Character Conversion System based on Semantic Pattern Matching, *International Journal on Computer Processing of Chinese and Oriental Languages*, 40, 227-236.

- Lai S. C. 2000. The Preliminary Study of phonetic symbol-to-Chinese character Conversion, *Master Thesis, National Tsing Hua University*.
- Lee L. S., Tseng C.Y., Gu H. Y., Liu F. H., Chang C. H., Lin Y. H., Lee Y., Tu S. L., Hsieh S. H. and Chen C. H. 1993. Golden Mandarin (I) - A Real Time Mandarin Dictation Machine for Chinese Language with Vary Large Vocabulary, *IEEE Transactions on Speech and Audio Proceeding*, 1, 2.
- Lin S. W. 1995. Prosodic-Segment Based Chinese Language Processing for Continuous Mandarin Speech Recognition with very large Vocabulary, *Master Thesis, National Taiwan University*.
- Lin J. X. 2002. A Mandarin Input System Compatible With Multiple Pinyin Methods, *Master Thesis, National Chung Hsing University*.
- Tsai J. L. and Hsu W. L. 2002. Applying an NVEF Word-Pair Identifier to the Chinese Syllable-To-Word Conversion Problem, *The 19<sup>th</sup> International Conference on Computational Linguistics*.
- Sproat R. and Shih C. 1990. A Statistic Method for Finding Word Boundaries in Chinese Text, *Computer Process of Chinese and Oriental Languages*, 4, 336-349.
- Thomas H. C., Charles E. L. and Ronald L. R. 1998. Introduction To Algorithms, *McGraw-Hill Book Company, New York St. Louis San Francisco Montreal Toronto*, 440-443.
- X. X. Wu. 1998. A Bucket Indexing Scheme for Error Tolerant Chinese Phrase Matching. *Master Thesis, National Chiao-Tung University*.
- Yang K. C. 1998. Further Studies for Practical Chinese Language Modeling, *Master Thesis, National Taiwan University*.

# Interleaving Text and Punctuations for Bilingual Sub-sentential Alignment

**Wen-Chi Hsie, Kevin Yeh, Jason S. Chang**

Department of Computer Science  
National Tsing Hua University  
101, Kuangfu Road, Hsinchu, 300, Taiwan, ROC  
{g904307, jschang}@cs.nthu.edu.tw

**Thomas C. Chuang**

Department of Computer Science  
Van Nung Institute of Technology  
1 Van-Nung Road, Chung-Li, Taiwan, ROC  
tomchuang@cc.vit.edu.tw

## Abstract

We present a new approach to aligning bilingual English and Chinese text at sub-sentential level by interleaving alphabetic texts and punctuations matches. With sub-sentential alignment, we expect to improve the effectiveness of alignment at word, chunk and phrase levels and provide finer grained and more reusable translation memory.

## 1. Introduction

Recently, there are renewed interests in using bilingual corpus for building systems for statistical machine translation (Brown et al. 1988, 1991), including data-driven machine translation (Richardson et al. 2002), computer-assisted revision of translation (Jutras 2000) and cross-language information retrieval (Kwok 2001). It is therefore useful for the bilingual corpus to be aligned at the sentence level and even sub-sentence level with very high precision (Moore 2002; Chuang, You and Chang 2002, Kueng and Su 2002). Especially, for further analyses such as phrase alignment, word alignment (Ker and Chang 1997; Melamed 2000) and translation memory, high-precision alignment at sub-sentential levels would be very useful. Alignment at sub-sentential level has the potential of improving the effectiveness of alignment at word and phrase levels and providing finer grained and more reusable translation memory.

Much work has been reported in the literature of computational linguistics on how to align sentences, while very little is touched on alignment just below the sentence level. The most effective approach for sentence alignment is the length-based approach proposed by Brown et al. (1991) and by Gale and Church (1991). Both methods use normal distribution to model the ratio of lengths between the counterpart sentences measured in number of words or characters. Length-based approach for aligning parallel corpora has commonly been used and produces surprisingly good results for the language pair of French and English at success rates well over 96%. However, it does not perform as well for alignment of text in two distant languages such as Chinese and English.

Yeh (2003) proposed a punctuation-based approach for sentence alignment which produces even high accuracy rates than the length based approach. It was pointed out that the ways different languages use punctuations are more or less similar and the correspondence of punctuations across different languages can be obtained using a small set of training data. By soft matching punctuations of the two languages in ordered comparison, the probabilities of mutual translation for a pair of bilingual sentences can be estimated more effectively than lengths. This is not surprising since the average sentence contains many punctuations which carry more information than lengths. Yeh also examined the results of punctuation-based sentence alignment and observed:

“Although word alignment links do cross one and other a lot, they general seem not to cross the links between punctuations. It appears that we can obtain sub-sentential alignment at clause and phrase levels from the alignment of punctuation.”

This observation indicates that in bilingual corpus pieces of text delimited by punctuations behave much the same way as sentences with non-crossing alignment links. Therefore, it is reasonable to align pieces of text ending with a couple of punctuations, much the same way as sentence alignment. Building on their work, we develop a new approach to sub-sentential alignment by interleaving the matches of alphabetic texts and punctuations. In the following, we first give an example for bilingual sub-sentential alignment in Section 2. Then we introduce our probability model in Section 3. Next, we describe experimental setup and results in Section 4. We conclude in Section 5 with discussion and future work.

## 2. Example

Consider a pair of counterpart paragraphs in the official records of Hong Kong Legislative Council:

“My goal is simply this - to safeguard Hong Kong's way of life. This way of life not only produces impressive material and cultural benefits; it also incorporates values that we all cherish. Our prosperity and stability underpin our way of life. But, equally, Hong Kong's way of life is the foundation on which we must build our future stability and prosperity.”

我的目標很簡單，就是要保障香港的生活方式。這個生活方式，不單在物質和文化方面為我們帶來了重大的利益，而且更融合了大家都珍惜的價值觀。香港的安定繁榮是我們生活方式的支柱。同樣地，我們未來的安定繁榮，亦必須以香港的生活方式為基礎。(Source: Oct. 7, 1992, Governor Christopher Francis Patten's address to the HK LEGCO)

By sub-sentential alignment, we mean identifying the shortest possible pair of counterpart texts ending with punctuations. From the example above, the following is the intended results of sub-sentential alignment:



- My goal is simply this –  
我的目標很簡單，
- to safeguard Hong Kong's way of life.  
就是要保障香港的生活方式。
- This way of life not only produces impressive material and cultural benefits;  
這個生活方式，不單在物質和文化方面為我們帶來了重大的利益，
- it also incorporates values that we all cherish.  
而且更融合了大家都珍惜的價值觀。

Notice that longer pairs such as the following translation equivalent pair of sentences

My goal is simply this – to safeguard Hong Kong's way of life.  
我的目標很簡單，就是要保障香港的生活方式。

does not fit the bill, since a finer grained subdivision into two 1-1 matches, (My goal is simply this –, 我的目標很簡單，) and (to safeguard Hong Kong's way of life., 就是要保障香港的生活方式。) also preserve translation equivalence. Not unlike the situation in sentence alignment, there are many to one, one to many, and many to many matches. For instance, it is not possible to find a 1-1 match for “This way of life not only produces impressive material and cultural benefits;” since it only corresponds to “這個生活方式，” in part. Therefore, we have to combine the subsequent clause “不單在物質和文化方面為我們帶來了重大的利益，” for a 1-2 match.

### 3. Probability Model

In this section we describe our probability model. To do so, we will first introduce some necessary notation. Let  $E$  be an English fragment  $e_1, e_2, \dots, e_m$  and  $C$  be a Chinese paragraph  $c_1, c_2, \dots, c_n$ , which  $e_i$  and  $c_j$  is a text-fragment as described in Section 2. We define a **link**  $l(e_i, c_j)$  for  $e_i$  and  $c_j$  that are translation ( or part of a translation ) of one another. We define **null link**  $l(e_i, c_0)$  for  $e_i$  which does not correspond to a translation. The null link  $l(e_0, c_j)$  is defined similarly. An **alignment**  $A$  for two paragraphs  $E$  and  $C$  is a set of links such that every text-fragment in  $E$  and  $C$  participates in at least one link, and a text-block linked to  $e_0$  or  $c_0$  participates in no other links.

We define the alignment problem as finding the alignment  $A$  that maximizes  $P(A|E, C)$ . An alignment  $A$  consists of  $t$  links  $\{l_1, l_2, \dots, l_t\}$ , where each  $l_k = (e_{i_k}, c_{j_k})$  for some  $i_k$  and  $j_k$ . We will refer to consecutive subsets of  $A$  as  $l_i^j = \{l_i, l_{i+1}, \dots, l_j\}$ , Given this notation,  $P(A|E, C)$  can be decomposed as follows:

$$P(A | E, F) = P(l_1^t | E, F) = \prod_{k=1}^t P(l_k | E, C, l_1^{k-1})$$

For each condition probability, given any pair  $e_i$  and  $c_j$ , the link probabilities can be determined directly from combining the probability of length-based model with punctuation-based model. From the paper of Gale and Church in 1993 for length-based model, we know the match probability is  $Prob(\delta | match)$  and  $Prob(match)$  and  $Prob(\delta | match)$  can be estimated by

$$Prob(\delta | match) = 2(1 - Prob(|\delta|))$$

Where  $Prob(|\delta|)$  is the probability that random variable,  $z$ , with a standardized (mean zero, variance one) normal distribution, has magnitude at least as large as  $|\delta|$ . That is,

Where

$$Prob(\delta) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\delta} e^{-z^2/2} dz$$

We compute  $\delta$  directly from the length of two portions of text,  $l_1$  and  $l_2$ , and the two parameters,  $c$  and  $s^2$ . (Where  $c$  is the expected number of characters in  $L_2$  per character in  $L_1$ , and  $s^2$  is the variance of the number of characters in  $L_2$  per character in  $L_1$ .) That is,  $\delta = (l_2 - l_1 \times c) / \sqrt{l_1 s^2}$ . Then,  $Prob(|\delta|)$  is computed by integrating a standard normal distribution (with mean zero and variance 1).

Then, from Yeh (2003), for punctuation-based model, we know:

$$P_{\text{pun}}(e_i, c_j) = P(pe_i, pc_j)P(|pe_i|, |pc_j|) \text{ for some } l_k = (e_i, c_j)$$

where  $e_i$  and  $c_j$  is  $\lambda$ , one, or two punctuations,

$e_i, c_j$  = English and Chinese text-block

$pe_i$  = the ending English punctuations of  $e_i, i = 1, m$

$pc_j$  = the ending Chinese punctuations  $c_j, j = 1, n$ ,

$P(pc_i, pe_i)$  = probability of  $pc_i$  translates into  $pe_i$ ,

Thus, for each link  $l_k$  given  $E, C$  and  $l$ , we can compute the probability as follows:

$$P(l_k | E, C, l^{k-1}) = P(\delta | match)P(match) * P_{\text{pun}}(e_i, c_j), \text{ So}$$

$$P(A | E, F) = \prod_{k=1}^t P(\delta_k)P(m_k)P_{\text{pun}}(l_k)$$

#### 4. Experimental results

In order to assess the performance of our sub-sentential alignment model, we run the system on two test cases:

1. Official record of proceedings of Hong Kong Legislative Council at Oct. 7, 1992,
2. Harry Potter Book I Chapter one.

For probability of punctuation, we use a small set of hand aligned data which led to the following model parameters:

1. Punctuation translation probability (Table 1),
2. Sentence match type probability (Table 2).

**Table 1.** Punctuation Translation probability

English Pun.	Chinese Pun.	Match Type	Counts	Probability
,	,	1-1	541	0.809874
,	、	1-1	56	0.083832
,	。	1-1	41	0.061377
,	「	1-1	10	0.01497
,	:	1-1	5	0.007485
,	;	1-1	4	0.005988

**Table 2.** Match probability of clauses

Match Type	Probability
1-0	0.000197
0-1	0.000197
1-1	0.6513
2-2	0.0066
1-2	0.0526
2-1	0.1776
Other	0.0066

**Table 3.** Performance evaluation for the two test cases

Test cases	# of paragraphs	# of matches	# of correct matches	Precision (%)
Official record of proceedings of Hong Kong Legislative Council	10	188	174	93
Harry Potter Book I Chapter 1	110	634	540	85

Preliminary results shown in Table 3 indicate precision rates of 85% and 93% for the two test cases.

## 5. Discussion and future work

We propose a model interleaving length-based text alignment and punctuation alignment to carry out sub-sentential alignment. The method seems to work reasonably well with an average precision rate around

90% in the evaluation of a preliminary implementation. There is still a lot of room for improvement. We are currently working on identification of more punctuations useful for sub-sentential alignment, proper segmentation of text ending with punctuations, and better model for lengths of sub-sentential fragment. We are also looking into the issues of best weighting scheme of length and punctuation information. Finally, the cases where there is inversion of translated fragments are difficult to handle with length information alone. We are also preparing to work with additional lexical information to solve this kind of problem in the future.

### Acknowledgements

We acknowledge the support for this study through grants from Ministry of Education, Taiwan (MOE EX-91-E-FA06-4-4). Thanks are also due to Jim Chang for preparing the training data and evaluating the experimental results.

### References

Church, K and P. Hank, "Word Association Norms, Mutual Information, and Lexicography," *Computational Linguistics*, 16:1, 1990, pp. 22-29.

Sproat, Chinese Word Segmentation, *First International Conference on Language Resources & Evaluation: Proceedings*, 1998, pp. 417— 420.

Richard Sproat, Chilin Shih, 1990, A statistical method for finding word boundaries in Chinese text. *Computer Processing of Chinese & Oriental Languages*, 4(4): 336-351.

R. Sproat, Chilin Shih, William Gale, and Nancy Chang. 1996. A stochastic finite-state word-segmentation algorithm for chinese. *Computational Linguistics*, 22(3): 377-404.

### References

Brown, P. F., J. C. Lai and R. L. Mercer (1991), 'Aligning sentences in parallel corpora', in 29th Annual Meeting of the Association for Computational Linguistics, Berkeley, CA, USA. pp. 169-176.

Chen, Stanley F. (1993), Aligning Sentences in Bilingual Corpora Using Lexical Information. In *Proceedings*

Chuang, T., G.N. You, J.S. Chang (2002) Adaptive Bilingual Sentence Alignment, *Lecture Notes in Artificial Intelligence* 2499, 21-30.

Gale, William A. & Kenneth W. Church (1993), A program for aligning sentences in bilingual corpus. In *Computational Linguistics*, vol. 19, pp. 75-102.

Jutras, J-M 2000. An Automatic Reviser: The TransCheck System, In *Proc. of Applied Natural Language Processing*, 127-134.

Ker, Sue J. & Jason S. Chang (1997), A class-based approach to word alignment. In *Computational Linguistics*, 23:2, pp. 313-344.

Kueng, T.L. and Keh-Yih Su, 2002. A Robust Cross-Domain Bilingual Sentence Alignment Model, In *Proceedings of the 19th International Conference on Computational Linguistics*.

Kwok, KL. 2001. NTCIR-2 Chinese, Cross-Language Retrieval Experiments Using PIRCS. In *Proceedings of the Second NTCIR Workshop Meeting*, pp. (5) 14-20, National Institute of Informatics, Japan.

Melamed, I. Dan (1997), A portable algorithm for mapping bitext correspondence. In *The 35th Conference of the Association for Computational Linguistics (ACL 1997)*, Madrid, Spain.

Piao, Scott Songlin 2000 Sentence and word alignment between Chinese and English. Ph.D. thesis, Lancaster University.

## Appendix

Table A. all incorrect alignments of this experiment. Shaded parts indicate imprecision in alignment results. We calculated the precision rates by dividing the number of unshaded sentences (counting both English and Chinese sentences) by total number of sentences proposed. Since we did not exclude aligned pair using a threshold, the recall rate should be the same as the precision rate.

Sub-sentence alignment based on length and punctuation	
English text	Chinese Text
Now is the time to show how we mean to prepare for Hong Kong's future under that far-sighted concept, "one country, two systems".	現在也是時候表明我們打算怎樣按照「一國兩制」這個極具遠見的構思，為香港的未來作好準備。
- we shall maintain an economy which continues to thrive and prosper, generating the wealth required to provide the standards of public service that people rightly demand; Our prescription for prosperity is straightforward.	— 我們便可令經濟持續繁榮蓬勃，創造所需財富，使提供的公共服務，能達到市民要求的合理水平；我們締造繁榮的配方清楚簡單。我們相信，
We believe that businessmen not politicians or officials make the best commercial decisions.	最佳的商業決定是由商人，而不是由政治家或政府官員作出的。
We believe that government spending must follow not outpace economic growth.	我們相信，政府開支必須跟隨經濟增長，
We believe in competition within a sound, fair framework of regulation and law.	而不應超逾經濟增長。我們更相信，應在健全而公平的法規下進行競爭。
I am inviting distinguished members of the business community to join it.	並會邀請商界傑出人士加入。他的職責是，
Their mandate will be to advise me on:	就下開事項向我提供意見：

## Restoration of Case Information in All-Cap English Broadcast Transcription

**Yu-Ting Liang**

Department of Computer Science  
National Tsing Hua University  
101, Kuangfu Road, Hsinchu, 300, Taiwan,  
ROC

[u902518@cs.nthu.edu.tw](mailto:u902518@cs.nthu.edu.tw)

**Jian-Chen Wu**

Department of Computer Science  
National Tsing Hua University  
101, Kuangfu Road, Hsinchu, 300, Taiwan,  
ROC

[g904374@cs.nthu.edu.tw](mailto:g904374@cs.nthu.edu.tw)

The local broadcast ICRT (International Community Radio Taipei) in Taipei has their news scripts sent to their listeners in ALL CAPS, which makes the articles more difficult to read. Therefore, we think it may facilitate the readers if we transform the text into normal cases that we are familiar with. In this prototype system, we established a practical method of restoration of case information, using different techniques from NLP and statistics. The system can apply many different kinds of approach, however, in this prototype, we focus our analysis and test data on broadcast transcription.

Basically, our research involves:

- Establishing a very large database containing numerous vocabularies and uses as our training data.
- Obtaining text from ICRT news scripts sent by e-mails as our test data.
- Restoring the cases of the contents into cases that we are more acquainted with.
- Handling some exceptions.

**Establishing a very large database.** Our training data comes from VOA news, which consists of 9138 articles, 3 million words in total. For each article, we segment its contents into individual words and calculate their n-gram probabilities.

**Obtaining text from ICRT news scripts.** We perform similar piecing process on the news scripts. After obtaining each isolated word, we query its probabilities in unigram, bigram, and trigram probabilities, which have two, four, and eight values respectively from our training data.

**Restoring the cases of the contents.** After accomplishing Viterbi algorithm (Rabiner, 1989) to compute the highest probability and its P-model value (Lucian Vlad Lita, 2003), we acquire the best restoration of case for each word, and then we alter the texts. We have an example in Figure 1, an original text from one of the ICRT news scripts and the text after restoration.

---

Upper Case: WILLIAM SAMPSON SPENT 31 MONTHS IN PRISON IN SAUDI ARABIA, WHERE HE WAS SENTENCED TO DEATH OVER A SERIES OF BOMBINGS THAT KILLED ONE PERSON.

After Restoration: William Sampson spent 31 months in prison in Saudi Arabia, where he was sentenced to death over a series of bombings that killed one person.

---

Figure 1: An example of upper case text and restoration

**Handling exceptions.** Actually, the word ‘Sampson’ was not found in our training data, however, we assume unknown words as proper nouns and therefore we capitalize its first letter. Here we have another experiment in Figure 2.

---

Upper Case: THE U.S. MILITARY ISSUED A PUBLIC APOLOGY TO THE PEOPLE OF A SHIITE MUSLIM NEIGHBORHOOD IN BAGHDAD ON THURSDAY FOR AN INCIDENT IN WHICH A MAN WAS KILLED AND FOUR OTHERS WOUNDED AFTER AN AMERICAN BLACK HAWK HELICOPTER BLEW DOWN AN ISLAMIC BANNER WITH ITS ROTOR WASH. THAT APPEARS TO BE A MAJOR SHIFT IN THE MILITARY'S RELATIONS WITH THE NEWS MEDIA.

After Restoration: **the** U. S. military issued a public apology to the people of a Shiite Muslim neighborhood in Baghdad on Thursday for an incident in which a man was killed and four others wounded after an American Black Hawk helicopter blew down an Islamic banner with its rotor wash. **that** appears to be a major shift in the military's relations with the news media.

---

Figure 2: Another example of upper case text and restoration

Again, we found some adjustments have to be done, and the first letter of the first word in a sentence ought to be in upper case is one of them. Even so, we have to ask ourselves, “What is a sentence?” Is it something ends up with a period, an exclamation mark, or a question mark? Apparently, we can find a counter example with “U.S.”. Here we use heuristic sentence boundary detection algorithm to determine what a sentence is and capitalized the first words in a

sentence as shown in Figure 3.

---

After restoration: **The** U.S. military issued a public apology to the people of a Shiite Muslim neighborhood in Baghdad on Thursday for an incident in which a man was killed and four others wounded after an American Black Hawk helicopter blew down an Islamic banner with its rotor wash. **That** appears to be a major shift in the military's relations with the news media.

---

Figure 3: The previous example after sentence adjustment

Our demonstration model shows we can convert all-cap English news scripts quite well. There are some possible improvements and our future works are improving our performance, which can reduce the time we spend on transforming the text. Also, create a macro in Outlook so if the readers receive their e-mails from ICRT with Outlook, they may have the restoration done by running a macro. We are looking forward to finding the readers feeling this tool useful and somewhat convenient.

## **Acknowledgements**

We acknowledge the support of NSC under contract number: 92-2815-C-007 -004 -E . Many thanks are due to Dr. Jason S. Chang for his guidance in NLP and ICRT for their news scripts.

## **References**

- Lucian Vlad Lita, 2003. *tRuEcasIng*.
- Hai Leong Chieu, Hwee Tou Ng, 2002 Teaching a Weaker Classifier: Named Entity Recognition on Upper Case Text.
- Andrei Mikheev, 1999 A Knowledge-free Method for Capitalized Word Disambiguation.
- Alison Huettner, Pero Subasic, 2000 Fuzzy Typing for Document Management.
- Christopher D. Manning and Hinrich Schutze. Foundations of statistical natural language processing, 2000, pp. 123-136



# Using Punctuations and Lengths for Bilingual Sub-sentential Alignment

**Wen-Chi Hsien, Kevin Yeh, Jason S. Chang**  
Department of Computer Science  
National Tsing Hua University  
101, Kuangfu Road, Hsinchu, 300, Taiwan, ROC  
{g904307, jschang}@cs.nthu.edu.tw

**Thomas C. Chuang**  
Department of Computer Science  
Van Nung Institute of Technology  
1 Van-Nung Road, Chung-Li, Taiwan, ROC  
tomchuang@cc.vit.edu.tw

## Abstract

We present a new approach to aligning bilingual English and Chinese text at sub-sentential level by interleaving alphabetic texts and punctuations matches. With sub-sentential alignment, we expect to improve the effectiveness of alignment at word, chunk and phrase levels and provide finer grained and more reusable translation memory.

## 1. Introduction

Recently, there are renewed interests in using bilingual corpus for building systems for statistical machine translation (Brown et al. 1988, 1991), including data-driven machine translation (2002), computer-assisted revision of translation (Jutras 2000) and cross-language information retrieval (Kwok 2001). It is therefore useful for the bilingual corpus to be aligned at the sentence level and even sub-sentence level with very high precision (Moore 2002; Chuang, You and Chang 2002, Kueng and Su 2002). Especially, for further analyses such as phrase alignment, word alignment (Ker and Chang 1997; Melamed 2000) and translation memory, high precision and quality alignment at sentence or sub-sentential levels would be very useful. Furthermore, alignment at sub-sentential level has the potential of improving the effectiveness of alignment at word, chunk and phrase levels and providing finer grained and more reusable translation memory.

Much work has been reported in the literature of computational linguistics studying how to align sentences. One of the most effective approaches is length-based approach proposed by Brown et al. and by Gale and Church. Length-based approach for aligning parallel corpora has commonly been used and produces surprisingly good results for the language pair of French and English at success rates well over 96%. However, it does not perform as well for alignment of two distant languages such as Chinese-English. Furthermore, for sub-sentential alignment, length-based approach gets less effectiveness than running it in sentence level since sub-sentence has less information in length.

Punctuations based approach (Yeh, Chuang and Chang 2003 ) for sentence alignment produces high accuracy rates as same as length based approach and was independent of languages. Although the ways different languages around the world use punctuations vary, symbols such as commas and full stops are used in most languages to demarcate writing, while question and exclamation marks are used to show emphasis. However, for sub-sentential alignment, punctuation-based approach has the same problem as length-based approach — no enough information in sub-sentence since sub-sentence might be very short and just include one or two punctuations within it.

Yeh, Chuang and Chang (2003) examined the results of punctuation-based sentence alignment and observed:

“Although word alignment links do cross one and other a lot, they general seem not to cross the links between punctuations. It appears that we can obtain sub-sentential alignment at clause and phrase levels from the alignment of punctuation.”

Building on their work, we develop a new approach to sub-sentential alignment by interleaving the alignment of text and punctuations. In the following, we first give an example for bilingual sub-sentential alignment in Section 2. Then we introduce our probability model in Section 3. Next, we describe experimental setup and results in Section 4. We conclude in Section 5 with discussion and future work.

## 2. Example

Consider a pair of aligned sentences in a parallel corpus as below:

“My goal is simply this - to safeguard Hong Kong's way of life. This way of life not only produces impressive material and cultural benefits; it also incorporates values that we all cherish. Our prosperity and stability underpin our way of life. But, equally, Hong Kong's way of life is the foundation on which we must build our future stability and prosperity.”

我的目標很簡單，就是要保障香港的生活方式。這個生活方式，不單在物質和文化方面為我們帶來了重大的利益，而且更融合了大家都珍惜的價值觀。香港的安定繁榮是我們生活方式的支柱。同樣地，我們未來的安定繁榮，亦必須以香港的生活方式為基礎。

We can observe that although word alignment links might cross one and other a lot, there exist some text-blocks as follow that general seem not to cross the links between punctuations:

“My goal is simply this –“

“我的目標很簡單，”

“to safeguard Hong Kong's way of life.”

“就是要保障香港的生活方式。”

“This way of life not only produces impressive material and cultural benefits;”

“這個生活方式，不單在物質和文化方面為我們帶來了重大的利益，”

“it also incorporates values that we all cherish.”

“而且更融合了大家都珍惜的價值觀。”

...

That’s what we call sub-sentences here. From the examples above, we can define that a sub-sentence is a text-block that include at least one or more punctuations. That’s an unclear definition since a sentence and a paragraph also fit the definition too. However, what we want is to find out the shortest parallel text-block pairs that fit the definition. That’s why in the third pair of above examples, “這個生活方式，” is a Chinese text-block but we have to combine it with “不單在物質和文化方面爲我們帶來了重大的利益，”，because we can’t find any English text-block correspond to “這個生活方式，”，we have to combine the two Chinese above first, than we can find the corresponding one : “This way of life not only produces impressive material and cultural benefits;”.

### 3. Probability Model

In this section we describe our probability model. To do so, we will first introduce some necessary notation. Let  $E$  be an English paragraph  $e_1, e_2, \dots, e_m$  and  $C$  be a Chinese paragraph  $c_1, c_2, \dots, c_n$ , which  $e_i$  and  $c_j$  is a text-blocks as described in Section 2. We define a **link**  $l(e_i, c_j)$  to exist if  $e_i$  and  $c_j$  are translation ( or part of a translation ) of one another. We define **null link**  $l(e_i, c_0)$  to exist if  $e_i$  does not correspond to a translation of any  $c_j$ . The null link  $l(e_0, c_j)$  is defined similarly. An **alignment**  $A$  for two paragraphs  $E$  and  $C$  is a set of links such that every text-block in  $E$  and  $C$  participates in at least one link, and a text-block linked to  $e_0$  or  $c_0$  participates in no other links.

We define the alignment problem as finding the alignment  $A$  that maximizes  $P(A|E, C)$ . An alignment  $A$  consists of  $t$  links  $\{l_1, l_2, \dots, l_t\}$ , where each  $l_k = l(e_{i_k}, c_{j_k})$  for some  $i_k$  and  $j_k$ . We will refer to consecutive subsets of  $A$  as  $l_i^j = \{l_i, l_{i+1}, \dots, l_j\}$ , Given this notation,  $P(A|E, C)$  can be decomposed as follows:

$$P(A | E, C) = P(l_1^t | E, C) = \prod_{k=1}^t P(l_k | E, C, l_1^{k-1})$$

For each condition probability, given any pair  $e_i$  and  $c_j$ , the link probabilities can be determined directly from combining the probability of length-based model with punctuation-based model. From the paper of Gale and Church in 1993 for length-based model, we know the match probability is  $Prob(\delta | match)$   $Prob(match)$  and  $Prob(\delta | match)$  can be estimated by

$$Prob(\delta | match) = 2(1 - Prob(|\delta|))$$

Where  $Prob(|\delta|)$  is the probability that random variable,  $z$ , with a standardized (mean zero, variance one) normal distribution, has magnitude at least as large as  $|\delta|$ . That is,

Where

$$Prob(\delta) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\delta} e^{-z^2/2} dz$$

We compute  $\delta$  directly from the length of two portions of text,  $l_1$  and  $l_2$ , and the two parameters,  $c$  and  $s^2$ . (Where  $c$  is the expected number of characters in  $L_2$  per character in  $L_1$ , and  $s^2$  is the variance of the number of characters in  $L_2$  per character in  $L_1$ .) That is,  $\delta = (l_2 - l_1 \times c) / \sqrt{l_1 s^2}$ . Then,  $Prob(|\delta|)$  is computed by integrating a standard normal distribution (with mean zero and variance 1).

Then, from Yeh, Chuang and Chang (2003), for punctuation-based model, we know:

$$P(e_i, c_j) = P(pe_i, pc_j) p(|pe_i|, |pc_j|) \times \prod_{i=1}^{m-1} P(pe_i, null) p(|pe_i|, 0) \times \prod_{j=1}^{n-1} P(null, pc_j) (0, |pc_j|)$$

where  $e_i$  and  $c_i$  is  $\lambda$ , one, or two punctuations,

$e_i, c_j$  = English and Chinese text-block

$pe_1 pe_2 \dots pe_m = pE$ , the English punctuations,

$pc_1 pc_2 \dots pc_n = pC$ , the Chinese punctuations,

$|pe_i|$  and  $|pc_i|$  are the number of punctuations in

$pe_i$  and  $pc_i$  respectively,

$P(pc_i, pe_i)$  = probability of  $pc_i$  translates into  $pe_i$ ,

Thus, for each link  $l_k$  given  $E, C$  and  $l$ , we can computing the probability as following:

$$P(l_k | E, C, l^{k-1}) = P(\delta | match) P(match) * P(e_i, c_i), \text{ So}$$

$$P(A | E, F) = \prod_{k=1}^l P(\delta | match) P(match) P(e_{ik}, c_{jk})$$

#### 4. Experimental result

In order to assess the performance of our sub-sentential alignment model, we selected top ten bilingual articles from official record of proceedings of Hong Kong Legislative Council at Oct. 7, 1992 as our experimental data. For probability of punctuation, We use all the data such as punctuation translation probability (Table 1) and category frequency  $Prob(match)$  (Table 2) from Yeh, Chuang and Chang (2003) directly. For probability of length, we set  $c = 3.23$ , standard variance = 0.93 and match probability as Table 3:

**Table 1.** Punctuation Translation probability

English Pun.	Chinese Pun.	Match Type	Counts	Probability
--------------	--------------	------------	--------	-------------

,	,	1-1	541	0.809874
,	、	1-1	56	0.083832
,	。	1-1	41	0.061377
,	「	1-1	10	0.01497
,	：	1-1	5	0.007485
,	；	1-1	4	0.005988

**Table 2: P(match) Category Frequency Prob(match)**

Match type	1-1	1-0, 0-1	1-2	2-1	1-3	1-4	1-5
Probability	0.65	0.000197	0.0526	0.178	0.066	0.0013	0.00013

**Table 3. Match probability of sentences**

Match Type	Probability
1-0	0.000197
0-1	0.000197
1-1	0.6513
2-2	0.0066
1-2	0.0526
2-1	0.1776
1-3	0.0066
3-1	0.0658
1-4	0.00132
4-1	0.0132

After aligning by our model, we got 94 parallel records from the ten articles, and had precision rate at 92.55%. To calculate precision rate, we count English and Chinese sub-sentences isolated, so were the error records. For detail, refer to Appendix. Following table show the result:

Article	# of sub-sentence	errors	Prec(%)
Official record of proceedings of Hong Kong Legislative Council	188	14	92.55

## 5. Discussion and future work

We propose a model combining length-based approach with punctuation-based approach to do sub-sentential alignment and we got about 93% precision rates here. It was not bad but still had a lot of space to improve. We should change the sub-sentence match type probability first of all. We use the probability of sentence match type instead of sub-sentence match type in this experiment since we don't do sub-sentence training first. It causes a problem, because a sub-sentence has higher probability to include two

or three text-blocks within it than a sentence do. An inverted sentence causes the second problem here, no matter length-based or punctuation-based approach you used; they cannot solve this kind of problem. We might add lexical information in it to solve this kind of problem in the future.

## References

- Brown, P. F., J. C. Lai and R. L. Mercer (1991), 'Aligning sentences in parallel corpora', in 29th Annual Meeting of the Association for Computational Linguistics, Berkeley, CA, USA. pp. 169-176.
- Chen, Stanley F. (1993), Aligning Sentences in Bilingual Corpora Using Lexical Information. In Proceedings
- Chuang, T., G.N. You, J.S. Chang (2002) Adaptive Bilingual Sentence Alignment, Lecture Notes in Artificial Intelligence 2499, 21-30.
- Gale, William A. & Kenneth W. Church (1993), A program for aligning sentences in bilingual corpus. In Computational Linguistics, vol. 19, pp. 75-102.
- Jutras, J-M 2000. An Automatic Reviser: The TransCheck System, In Proc. of Applied Natural Language Processing, 127-134.
- Ker, Sue J. & Jason S. Chang (1997), A class-based approach to word alignment. In Computational Linguistics, 23:2, pp. 313-344.
- Kueng, T.L. and Keh-Yih Su, 2002. A Robust Cross-Domain Bilingual Sentence Alignment Model, In Proceedings of the 19th International Conference on Computational Linguistics.
- Kwok, KL. 2001. NTCIR-2 Chinese, Cross-Language Retrieval Experiments Using PIRCS. In Proceedings of the Second NTCIR Workshop Meeting, pp. (5) 14-20, National Institute of Informatics, Japan.
- Melamed, I. Dan (1997), A portable algorithm for mapping bitext correspondence. In The 35th Conference of the Association for Computational Linguistics (ACL 1997), Madrid, Spain.
- Piao, Scott Songlin 2000 Sentence and word alignment between Chinese and English. Ph.D. thesis, Lancaster University.
- Kevin C. Yeh, Thomas C. Chuang, Jason S. Chang (2003), Using Punctuations for Bilingual Sentence Alignment - Preparing Parallel Corpus for Distribution by the ACLCLP

## Appendix

Table A. all incorrect alignments of this experiment. Shaded parts indicate imprecision in alignment results. We calculated the precision rates by dividing the number of unshaded sentences (counting both English and Chinese sentences) by total number of sentences proposed. Since we did not exclude aligned pair using a threshold, the recall rate should be the same as the precision rate.

Sub-sentence alignment based on length and punctuation	
English text	Chinese Text
[Now] [is the time] [to show] [how we mean to prepare] [for Hong Kong's future] [under that far-sighted concept],	現在也是時候表明我們打算怎樣按照「一國兩制」這個極具遠見的構思，
["one country, two systems"].	為香港的未來作好準備。
[- we shall maintain] [an economy] [which] [continues to thrive] [and prosper,]	— 我們便可令經濟持續繁榮蓬勃，創造所需財富，
[generating the wealth] [required to provide] [the standards of public service] [that people rightly demand;]	使提供的公共服務，能達到市民要求的合理水平；
[Our prescription for prosperity] [is straightforward.]	我們締造繁榮的配方清楚簡單。我們相信，
[We believe that] [businessmen] [not politicians or officials] [make the best commercial decisions.]	最佳的商業決定是由商人，而不是由政治家或政府官員作出的。
[We believe that] [government spending] [must follow] [not outpace] [economic growth.]	我們相信，政府開支必須跟隨經濟增長，
[We believe in competition] [within] [a sound, fair framework of regulation and law.]	而不應超逾經濟增長。我們更相信，應在健全而公平的法規下進行競爭。
[I am inviting distinguished members] [of the business community] [to join it.]	並會邀請商界傑出人士加入。他的職責是，
[Their mandate] [will be] [to advise me] [on:]	就下開事項向我提供意見：

# **TotalRecall: A Bilingual Concordance in National Digital Learning Project - CANDLE**

**Jian-Cheng Wu, Wen-Chi Shei , Jason S. Chang**

Department of Computer Science

National Tsing Hua University

101, Kuangfu Road, Hsinchu, 300, Taiwan, ROC

{g904307, jschang }@cs.nthu.edu.tw

## **Abstract**

This paper describes a Web-based English-Chinese concordance system, TotalRecall, being developed in National Digital Learning Project – CANDLE, to promote translation reuse and encourage authentic and idiomatic use in second language learning. We exploited and structured existing high-quality translations from the bilingual Sinorama Magazine to build the concordance of authentic text and translation. Novel approaches were taken to provide high-precision bilingual alignment on the sentence, phrase and word levels. A browser-based user interface also developed for ease of access over the Internet. Users can search for word, phrase or expression in English or Chinese. The Web-based user interface facilitates the recording of the user actions to provide data for further research.

## **1 Introduction**

A concordance tool is particularly useful for studying a piece of literature when thinking in terms of a particular word, phrase or theme. It will show how often and where a word occurs, so can be helpful in building up some idea of how different themes recur within an article or a collection of articles. Concordances have been indispensable for lexicographers and increasingly considered instrumental for promoting learning effectiveness and motivation for language instructors and learners. A bilingual concordance tool is like a monolingual concordance, except that each sentence is associated with translation counterpart in a second language. It could be extremely useful for bilingual lexicographers, human translators and second language learners. Pierre Isabelle, in 1993, pointed out: “existing translations contain more solutions to more translation problems than any other existing resource.” It is particularly useful and convenient when the resource of existing translations is made available on the Internet. Web based bilingual concordances have proved to be very useful and popular. For example, the English-French concordance system, *TransSearch* (Macklovitch et al. 2000), provides a familiar interface for the users who only need to type in the expression in question, a list of citations will come up and it is easy to scroll down until one finds one that is useful.



In addition to the similar functionalities provided by *TransSearch*, **TotalRecall** comes with an additional feature making the *solution* more easily recognized: the user not only gets all the citations related to the expression in question, but also gets to see the translation counterpart highlighted.

**TotalRecall** extends the translation memory technology and provide an interactive tool intended for translators and non-native speakers trying to find ideas to properly express themselves. **TotalRecall** empowers the user by allowing her to take the initiative in submitting queries for searching authentic, contemporary use of English. These queries may be single words, phrases, or longer expressions, the system will search a substantial and relevant corpus and return bilingual citations that are helpful to human translators and second language learners.

## 2 Aligning the corpus

Central to **TotalRecall** is a bilingual corpus and a set of programs that provide the bilingual analyses to yield a *translation memory* database out of the bilingual corpus. Currently, we are working with a collection of Chinese-English articles from the Sinorama magazine. Two additional bilingual collections: Studio Classroom English lessons and Hansard of Hong Kong Legislative Council are now in the work. That would allow us to offer bilingual texts in both translation directions and with different levels of difficulty. Currently, the articles from Sinorama seems to be quite usefully by its own, covering a wide range of topics, reflecting the personalities, places, and events in Taiwan for the past three decade.

The concordance database is composed of bilingual sentence pairs. In addition, there are also tables to record additional information, including the source of each sentence pairs, metadata, and the information on phrase and word level alignment. With that additional information, **TotalRecall** provides various functions, including

1. Viewing of the full text of the source with a simple click.
2. Highlighted translation counterpart of the query words or phrases.
3. Ranking that is pedagogically useful for translation and language learning.

We are currently running an experimental prototype with Sinorama articles, dated mainly from 1990 to 2000. There are approximately 50,000 bilingual sentences and over 2 million words in total. We also plan to continuously updating the database with newer information from Sinorama magazine so that the concordance is kept up to date and relevant.

The bilingual texts that go into **TotalRecall** must be rearranged and structured. We describe the main steps below:

## **2.1 Sentence Alignment**


After parsing each article from files and put them into the database, we need to segment articles into sentences and align them into pairs of mutual translation. While the length-based approach (Church and Gale 1991) to sentence alignment produces surprisingly good results for the close language pair of French and English at success rates well over 96%, it does not fair as well for distant language pairs such as English and Chinese. Work on sentence alignment of English and Chinese texts (Wu 1994), indicates that the lengths of English and Chinese texts are not as highly correlated as in French-English task, leading to lower success rate (85-94%) for length-based aligners.

Simard, Foster, and Isabelle (1992) pointed out cognates in two close languages such as English and French can be used to measure the likelihood of mutual translation. However, for the English-Chinese pair, there are no orthographic, phonetic or semantic cognates readily recognizable by the computer. Therefore, the cognate-based approach is not applicable to the Chinese-English tasks.

At first, we used the length-based method for sentence alignment. The average precision of aligned sentence pairs is about 95%. We are now using a newer alignment method based on punctuation statistics. Although the average ratio of the punctuation counts in a text is low (less than 15%), punctuations provide valid additional evidence, helping to achieve high degree of alignment precision. It turns out that punctuations alone are telling evidences for sentence alignment, if we do more than hard matching of punctuations and take into consideration of intrinsic sequencing of punctuation in ordered comparison. Experiment results show that the punctuation-based approach outperforms the length-based approach with precision rates approaching 98%.

## **2.2 Phrase and Word Alignment**

After sentences and their translation counterparts are identified, we proceeded to carry out finer-grained alignment on the phrase and word levels. We employ part of speech patterns and statistical analyses to extract bilingual phrases/collocations from a parallel corpus. The preferred syntactic patterns are obtained from idioms and collocations in the machine readable English-Chinese version of Longman Dictionary of Contemporary of English.



Text Collection :

---

Query: (English)  (Chinese)   items/page  
 mono  bilingual mark  successive order by:

---

English Sentence	Chinese Sentence	Source
Hsueh notes that those two historical figures' <b>hard</b> work and creativity changed the way people live. This is also the spirit that pAsia wishes to project.	薛曉嵐表示，這些人的 <b>努力</b> 與創作，改變了人們的生活模式，這也正是資訊人想要傳達的精神。	Internet Pioneer Heidi Hsueh <a href="#">[110 citation]</a> <input type="button" value="text"/> <input type="button" value="para"/>
A: It's very <b>hard</b> to change a person's character. It's true that I'm someone with intense emotions, a woman who basically has a very full, happy life.	答：一個人的個性 <b>很難</b> 改變，我的確是一個感情充沛、基本上活得很充實、很快樂的女人。	Writing Blossoms on a Withered Tree-- Interview with Author Shih Shu-ching <a href="#">[105 citation]</a> <input type="button" value="text"/> <input type="button" value="para"/>
When he reaches age 65, after 30 years of disciplined saving and investment, Mr. B has NT\$27 million in the bank. Not bad. But when Mr. A reaches 65, after relying only on his 15 years of <b>hard</b> work as a young man and on cumulative returns thereafter, he has	到了六十五歲驗收成果，發現某乙 <b>辛苦</b> 投資三十年，連本帶利約二千七百萬，而某甲只靠年輕時的十五年投入，竟平白累積了一億二千五百萬，是某乙的五倍！	The Early Bird Gets the Penny Earned <a href="#">[38 citation]</a> <input type="button" value="text"/> <input type="button" value="para"/>

Figure 1. The results of searching for “hard” with citation ranking by counts of word and translation pairs.

Phrases matching the patterns are extract from aligned sentences in a parallel corpus. Those phrases are subsequently matched up via cross linguistic statistical association. Statistical association between the whole phrase as well as words in phrases are used jointly to link a collocation and its counterpart collocation in the other language. See Table 1 for an example of extracting bilingual collocations. The word and phrase level information is kept in relational database for use in processing queries, highlighting translation counterparts, and ranking citations. Sections 3 and 4 will give more details about that.

### 3 The Queries

The goal of the **TotalRecall** System is to allow a user to look for instances of specific words or expressions. For this purpose, the system opens up two text boxes for the user to enter queries in any one of the languages involved or both. We offer some special expressions for users to specify the following queries:

**Table 1** The result of Chinese collocation candidates extracted. The shaded collocation pairs are selected based on competition of whole phrase log likelihood ratio and word-based translation probability. Un-shaded items 7 and 8 are not selected because of conflict with previously chosen bilingual collocations, items 2 and 3.

No.	English collocations	Chinese collocations	LLR	Word Prob.
1.	iron rice bowl	鐵飯碗	103.3	0.0202
2.	performance review bonus	考績獎金	63.03	0.1374
3.	year-end bonus	年終獎金	59.21	0.0700
4.	civil service rice	公家飯	29.08	0.0378
5.	economic downturn	經濟景氣低迷	28.4	0.6961
6.	pay cut	減薪	28.4	0.0585
7.	year-end bonus	考績獎金	27.35	0.2037
8.	performance review bonus	年終獎金	26.31	0.0370
9.	starve to death	餓不死	26.31	0.5670

- Exact single word query - W. For instance, enter “work” to find citations that contain “work,” but not “worked”, “working”, “works.”
- Exact single lemma query – W+. For instance, enter “work+” to find citations that contain “work”, “worked”, “working”, “works.”
- Exact string query. For instance, enter “in the work” to find citations that contain the three words, “in,” “the,” “work” in a row, but not citations that contain the three words in any other way.
- Conjunctive and disjunctive query. For instance, enter “give+ advice+” to find citations that contain “give” and “advice.” It is also possible to specify the distance between “give” and “advice,” so they are from a VO construction. Similarly, enter “hard | difficult | tough” to find citations that involve difficulty to do, understand or bear something, using any of the three words.

Once a query is submitted, **TotalRecall** displays the results on Web pages. Each result appears as a pair of segments, usually one sentence each in English and Chinese, in side-by-side format. The words matching the query are highlighted, and a “context” hypertext link is included in each row. If this link is selected, a new page appears displaying the bilingual paragraph or article, containing query.

#### 4 Ranking

It is well known that the typical user has no patient to go beyond the first or second pages returned by a search engine. Therefore, ranking and putting the most useful information in the first one or two pages is of paramount importance for search engines. This is also true for a concordance.

Experiments with a focus group indicate that the following ranking strategies are important:

- Citations with a translation counterpart should be ranked first.
- Citations with a frequent translation counterpart appear before ones with less frequent translation
- Citations with same translation counterpart should be shown in clusters. The cluster can be called out entirely on demand.
- Ranking by nonlinguistic features should also be provided, including date, sentence length, query position in citations, relevancy as indicated via within document term frequency, etc.

With various ranking options available, the users can choose one that is most convenient and productive for the work at hand.

## 5 Conclusion

In this paper, we describe a bilingual concordance designed as a computer assisted translation and language learning tool. Currently, **TotalRecall** uses Sinorama Magazine corpus as the translation memory and will be continuously updated as new issues of the magazine becomes available. We have already put a beta version on line and experimented with a focus group of second language learners. The learners as well as their instructors seems to enjoy the novel features of **TotalRecall** including highlighting of query and corresponding translations, clustering and ranking of search results according translation and frequency.

**TotalRecall** enables the non-native speaker who is looking for a way to express an idea in English or Chinese. We are also adding on the basic functions to include a log of user activities, which will record the users' query behavior and their background. We could then analyze the data and find useful information for future research.

## Acknowledgement

We acknowledge the support for this study through grants from National Science Council and Ministry of Education, Taiwan (NSC 90-2411-H-007-033-MC and MOE EX-91-E-FA06-4-4) and a special grant for preparing the Sinorama Corpus for distribution by the Association for Computational Linguistics and Chinese Language Processing.

## References

- Chuang, T.C. and J.S. Chang (2002), Adaptive Sentence Alignment Based on Length and Lexical Information, ACL 2002, Companion Vol. P. 91-2.
- Gale, W. & K. W. Church, "A Program for Aligning Sentences in Bilingual Corpora" Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics, Berkeley, CA, 1991.
- Macklovitch, E., Simard, M., Langlais, P.: TransSearch: A Free Translation Memory on the World Wide Web. Proc. LREC 2000 III, 1201--1208 (2000).
- Nie, J.-Y., Simard, M., Isabelle, P. and Durand, R.(1999) Cross-Language Information Retrieval based on Parallel Texts and Automatic Mining of Parallel Texts in the Web. Proceedings of SIGIR '99, Berkeley, CA.
- Simard, M., G. Foster & P. Isabelle (1992), Using cognates to align sentences in bilingual corpora. In Proceedings of TMI92, Montreal, Canada, pp. 67-81.
- Wu, Dekai (1994), Aligning a parallel English-Chinese corpus statistically with lexical criteria. In The Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics, New Mexico, USA, pp. 80-87.
- Wu, J.C. and J.S. Chang (2003), Bilingual Collocation Extraction Based on Syntactic and Statistical Analyses, ms.
- Yeh, K.C., T.C. Chuang, J.S. Chang (2003), Using Punctuations for Bilingual Sentence Alignment- Preparing Parallel Corpus for Distribution by the ACLCLP, ms.

## Unsupervised Word Segmentation Without Dictionary

**Jason S. Chang**

Department of Computer Science  
National Tsing Hua University  
101, Kuangfu Road,  
Hsinchu, 300, Taiwan, ROC  
[jschang@cs.nthu.edu.tw](mailto:jschang@cs.nthu.edu.tw)

**Tracy Lin**

Department of Communication Engineering  
National Chiao Tung University  
1001, Ta Hsueh Road,  
Hsinchu, 300, Taiwan, ROC  
[tracylin@faculty.nctu.edu.tw](mailto:tracylin@faculty.nctu.edu.tw)

This prototype system demonstrates a novel method of word segmentation based on corpus statistics. Since the central technique we used is unsupervised training based on a large corpus, we refer to this approach as *unsupervised word segmentation*.

The unsupervised approach is general in scope and can be applied to both Mandarin Chinese and Taiwanese. In this prototype, we illustrate its use in word segmentation of Taiwanese Bible written in Hanzi and Romanized characters. Basically, it involves:

- Computing mutual information, MI, between Hanzi and Romanized characters  $A$  and  $B$ . If  $A$  and  $B$  have a relatively high MI, we lean toward treating  $AB$  as a word.
- Using a greedy method to form words of 2 to 4 characters in the input sentences.
- Building an N-gram model from the results of first-round word segmentation
- Segmenting words based on the N-gram model
- Iterating between the above two steps: building N-gram and word segmentation

**Computing mutual information.** Using mutual information is motivated by the observation of previous work by Hank and Church (1990) and Sproat and Shih (1990). If  $A$  and  $B$  have a relatively high MI that is over a certain threshold, we prefer to identify  $AB$  as a word over those having lower MI values. In the experiment with Taiwanese Bible, the system identified Hanzi and Romanized syllables. Out of those, we obtained pairs of consecutive single or double Hanzi characters and Romanized syllables. So those pairs are commonly known as character bigrams, trigrams, and fourgrams. We differed from the common N-gram calculation and treated those as pairs of character sequence in order to apply mutual information statistics. Table 1 shows some examples of the pairs and MI values. We have excluded pairs having MI 2.2 or lower.

Table 1. Example of consecutive pairs ( $C_1$ ,  $C_2$ ) and MI values for the text of Taiwanese Bible

$C_1$	$C_2$	Mutual Information
婦	仁人	568.6012
婦	仁	281.1248
果	子	34.9152
婦仁	人	34.6398
園內	樹	23.6275
仁	人	16.4376
內	樹裡	10.7914
園	內	10.6569
通	食	8.9151
樹	裡	4.2192
仁人	對	3.2395
阮	通食	2.8967

**Word Segmentation.** With the potential words and MI values indicating their likelihood, we proceeded to segment the text of a large corpus into words. For the Taiwanese Bible, we had to take care of the problem of text being written down in more than one writing system: we had mixed Hanzi and Romanized syllables as input. Using a greedy method, we gradually formed words of 2 to 4 characters (or Romanized syllables) in the input sentences. A word with high-MI constituent characters took precedence in forming words.

Table 2. Example of consecutive pairs and MI values for the text of Taiwanese Bible

Left Syllable String, $C_1$	Right Syllable String, $C_2$	Mutual Information $MI(C_1, C_2)$
2-Syllable pairs		
婦	仁	281.1248
果	子	34.9152
仁	人	16.4376
園	內	10.6569
通	食	8.9151
樹	裡	4.2192
仁人	對	3.2395
3-Syllable pairs		
婦	仁人	568.6012
婦仁	人	34.6398
園內	樹	23.6275
內	樹裡	10.7914
阮	通食	2.8967

When successive words were formed, they could not contradict with the words determined previously. For instance, given the input “婦仁人對蛇講：「園內樹裡的果子阮通食，” we looked



up the table storing MI statistics and obtained the information shown in Table 2. First, we formed words of two characters. Based on the information in Table 2, the system formed the words, 婦仁, 果子, 通食, 園內, 樹裡. Notice that 仁人 is not selected because of conflict with previous decision about the word 婦仁. Subsequently, we tried to extend the two-syllable words chosen. A word is extended to three or four syllables if the MI is increased and in the corpus over  $\tau$  % of instances the two-character words can be extended that way. Currently, we set  $\tau = 60$ .

Admittedly, there is limitation to what distributional regularity based on MI can be exploited for word segmentation and there were still many errors in the first-round word segmentation results. For instance, for the input, “我祈禱耶和華講：『主耶和華啊 ... ,” the system produced the segmentation of “我 / 祈禱 / 耶和華 / 講 / : / / 『 / 主耶 / 和華 / 啊 / .” The first instance of 耶和華 was segmented correctly, while the second instance of 耶和華 was over-segmented because of the frequent character 主 before it. That problem can be partially alleviated by an Expectation Maximization Algorithm of learning an N-gram model of word segmentation.

### Building an N-gram model.

Currently, we used the unigram model where the probability of each word was estimated based on the Good-Turing smoothing technique. First we tally the total number of words  $N$  and count  $R$  of each word  $W$ . Let  $N_r$  be the number of distinct words have count  $r$ . Also, let  $N_0$  be the number of distinct syllable strings that never appear as a word. Good-Turing smoothing stipulates that we calculate  $r'$  as an adjustment for  $r$  as follows:

$$r_0 = N_1 / N_0$$

$$r_i = (i+1) N_{i+1} / N_i$$

After the adjustment step, we obtained the probability for the unigram model as follows:

$$P(W) = r' / N \quad \text{where } r' \text{ is the smoothed count of } W$$

For instance, we had the counts after the first-round MI-based segmentation as showed in Table 3.

Table 3. Good-Turing estimates for unigrams: Adjusted frequencies and probabilities

$r$	$N_r$	$r^*$	$P_{GT}()$
-----	-------	-------	------------

0	972,444	<b>0.00417</b>	0.0000000074
1	<b>4,056</b>	<b>0.97436</b>	0.0000017360
2	<b>1,976</b>	<b>1.37854</b>	0.0000024562
3	<b>908</b>	2.94273	0.0000052431
4	668	3.69760	0.0000065881

Table 4. Probabilities used in word segmentation of “主耶和華”

Word	Raw Count	Probability
主耶	<b>268</b>	0.0004775030
和華	<b>433</b>	0.0007714881
主	<b>1,048</b>	0.0018672506
耶和華	<b>5,612</b>	0.0099990557
主耶和	<b>0</b>	0.0000000074
華	<b>404</b>	0.0007198180
耶和	<b>1</b>	0.0000017360

**Word Segmentation based on the N-gram model.** We proceeded to redo the word segmentation task on the same corpus with an aim of rectifying the errors occurring in the previous stage. This was done following the standard dynamic programming procedure of Viterbi Algorithm of finding segmentation  $S$  satisfying the following optimality condition:

$$S = \arg \max_{(W_1, W_n)} \prod_{i=1, n} P(W_i).$$

For the example of “我祈禱耶和華講：『主耶和華啊 ... 』” given earlier, the system is likely to produce correct segmentation “我 / 祈禱 / 耶和華 / 講 / : / 『 / 主 / 耶和華 / 啊 / ... .”

Table 5. Probabilities for various segmentations of “主耶和華”

Segmentation, $S$	$P(W_1)$	$P(W_2)$	$P(W_3)$	$P(S)$
主, 耶和華	0.0018672506	0.0099990557	-	0.0000186707
主耶, 和華	<b>0.0004775030</b>	<b>0.0007714881</b>	-	<b>0.0000003684</b>
主耶和, 華	<b>0.0000000074</b>	<b>0.0007198180</b>	-	<b>0.000000000053</b>
主, 耶和, 華	<b>0.0018672506</b>	<b>0.0000017360</b>	<b>0.0007198180</b>	<b>0.000000000023</b>

**Iterating between building N-gram and word segmentation.** The improved word segmentation obviously will bring about a better N-gram model for segmentation. Subsequently the improved N-gram will help to produce segmentation results of higher accuracy. The process of improvement usually converges quickly after a couple of iterations.

Our demonstration prototype sheds new lights on the extensively studied problem of word

segmentation. The prototype illustrates:

- It is possible to achieve high-precision word segmentation for a sufficiently large corpus without a dictionary, rivaling human annotation.
- The heuristic MI-based approach by Sproat can be extended effectively to handle words longer than two characters.
- A more theoretically sound approach based on N-gram model and unsupervised learning based on EM-like algorithm can bring about higher performance than the heuristic approach based on mutual information.
- Unsupervised, self-organized word segmentation can provide an objective view of word segmentation. This should be considered as a quantitative, corpus-dependent method when setting up a segmentation standard or benchmark for word segmentation.

High-precision segmentation of Hanzi text can be achieved by unsupervised training on a reasonably sized corpus. Unsupervised word segmentation represents an innovative way to acquire lexical units in a large corpus based on lexical distributional regularity. Word segmentation algorithm is standard Viterbi algorithm and is independent of N-gram trained on the corpus, making it easy to change domains. The approach is useful in an indefinite number of areas, and lends itself to customization for a particular user or task. For example, the results can be used to prepare a concordance, as the first steps in many natural language processing systems such as machine translation, information retrieval, or text-to-speech system. Finally, the model explored here can be a basis for self-organized word segmentation and alignment of bilingual Chinese-English corpus.

### **Acknowledgements**

We acknowledge the support for this study through grants from Ministry of Education, Taiwan (MOE EX-91-E-FA06-4-4).

### **References**

- Church, K and P. Hank, "Word Association Norms, Mutual Information, and Lexicography," *Computational Linguistics*, 16:1, 1990, pp. 22-29.
- Sproat, Chinese Word Segmentation, *First International Conference on Language Resources & Evaluation: Proceedings*, 1998, pp. 417— 420.
- Richard Sproat, Chilin Shih, 1990, A statistical method for finding word boundaries in Chinese text. *Computer Processing of Chinese & Oriental Languages*, 4(4): 336-351.
- R. Sproat, Chilin Shih, William Gale, and Nancy Chang. 1996. A stochastic finite-state word-segmentation algorithm for Chinese. *Computational Linguistics*, 22(3): 377-404.

# 盲胞有聲書語音查詢系統

林政源、謝明峰、張智星  
國立清華大學資訊工程學系  
{gavins, pacific, jang}@wayne.cs.nthu.edu.tw

摘要：

我們設計一套採用語音輸入和輸出的有聲書查詢系統，目的在讓視力障礙的盲胞能方便查詢並收聽清大盲友會有聲書。盲胞不但可以查詢資料庫的書籍，也可以下載有聲書直接收聽。這個系統採用了兩大語音處理的技術：語音辨識與語音合成。前者是利用 HMM-based 的原理，而後者是採用 concatenation-based PSOLA 的合成技術。在系統設計方面則運用了 Microsoft .NET 架構下的 Web Service 來進行所有功能的整合。最後，在系統辨識評比方面，我們也得到了不錯的成果。

一．系統介紹：

本系統是架構在 Microsoft .NET Framework 之下，利用 Web Service 的功能，來進行各種網路資料的傳輸。使用者利用簡單的按鍵來選擇使用書名、作者或出版社來查詢，再經過麥克風的語音輸入，系統會將語音檔傳到 Web Service 中，然後採用梅爾刻度式倒頻譜 (MFCC) [2]的方法進行語音特徵參數粹取。再將粹取後的參數和資料庫中訓練過的所有語句進行比對，找出分數最高的來當作辨認結果。利用辨認結果去查詢資料庫，取得該筆資料的所有相關資訊，例如書籍編

號、書名等等。

在輸出方面，我們採用文字顯示與語音合成兩種並行的設計。使用者可以從顯示器得知查詢結果，或者利用喇叭聆聽查詢結果。倘若使用者有興趣試聽找到的書籍，可以直接按鍵下載，系統將會從網路上抓取該本有聲書，直接撥放給使用者試聽。

以下是整個系統以語音輸入與文字和語音輸出的流程圖：

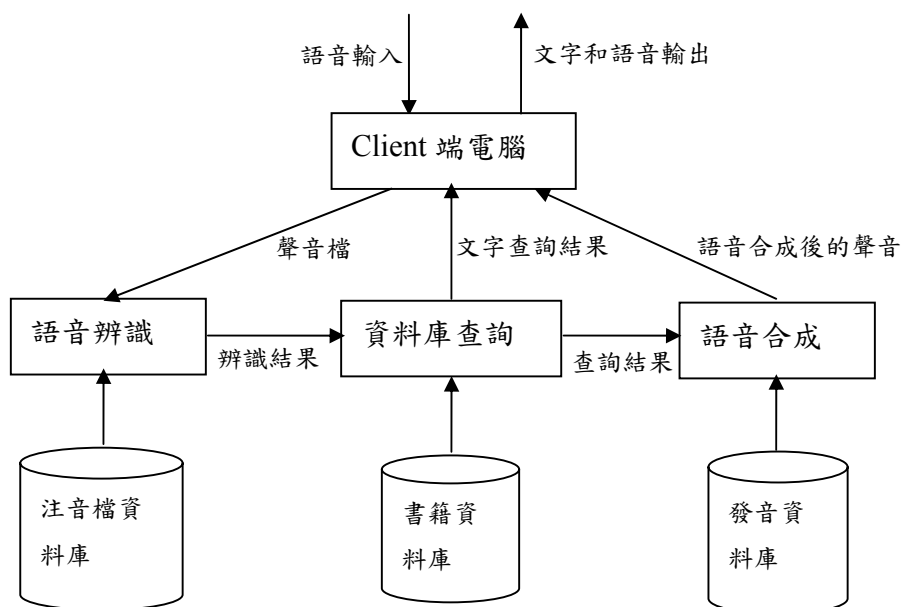


圖 1.系統架構圖

二．採用技術：

1. Microsoft .NET 平台：

.NET 平台提供了 Web Service 功能，在網路迅速普及，且寬頻蔚為風潮的現在，更能顯現其便利性。我們可將 Web Service 看成一個可使用的遠端函式。程式設計師只要連上網路，並了解函式的輸入參數和輸出結果，就可以直接使用該函式而不必花時間重新撰寫。充份用到了資源再利用的好

處。因此，在設計本系統的時候，將所需要的語音辨識、資料庫查詢、文字轉語音都以 Web Service 方式實作，如此可以減輕使用者端電腦的負擔，達到快速搜尋和方便使用的目的。

## 2. 語音辨識核心：

語音辨識的最大好處就是使用者可以不用依傳統文字輸入方式作查詢，而改以更人性化的語音輸入方式查詢，以期能拉近人與機器間的距離。而語音辨識主要分為二個步驟，MFCC (Mel-scale Frequency Cepstral Coefficient) 特徵參數粹取與 HMM (Hidden Markov Model) 比對辨識[2]。分述如下：

### I. 語音特徵參數粹取：

特徵參數粹取的目的是在於將一段語音檔的聲波(Waveform 形式)轉為另一個參數表示(通常資料量會明顯降低)，以便將來辨識之用。而這裡使用梅爾倒頻譜參數(MFCC)，其考慮到人耳對頻率的特性，較其他方法為佳。

### II. 比對辨識：

早期以 DTW (Dynamic Time Warping) [2]的技術來實作語音辨識，然而效果並不甚好。我們的系統則是採用隱藏式馬可夫模型(HMM, Hidden Markov Model) 為其辨識核心，其具有語音統計特性，經證實能夠有效模擬語音的細微變化，為近年來語音辨識最為廣泛的使用方法。

以上這二個步驟可利用 HTK (HMM Tool Kit) [3]加以完成。HTK 是一套功能

強大的語音辨識軟體，可以將大量的語音用 HMM 訓練之後，加以辨識。所以本系統採用 HTK 為辨識核心。並將所有的書名、作者、出版社從資料庫中粹取出來，進行標注音的動作。再將要被辨識的語音檔做特徵參數粹取，利用 Viterbi Search 演算法，將粹取出來的參數和之前標好的注音檔比對，找出一個最相似句子，當做語音辨識的結果。

### 3. 語音合成 ( TTS, Text to Speech ) :

為了能讓盲胞也能知道查詢的結果，整個系統需要能將文字轉成語音(TTS, Text to Speech)的功能，才能有聲音的輸出。首先先將文字輸入到 TTS 系統中，TTS 系統在收到文字後，根據原有在資料庫中的語音檔進行連音，調整長度、大小及聲調的動作。這裡採用的方法是基週同步疊加法，PSOLA (Pitch Synchronous Overlap and Add) [1]。另外處理中文時，必須考慮到聲調的轉換 (例如，李總統這三個字的聲調為：3 聲、2 聲、3 聲 )，所以我們必須另外建立一些常用的辭庫，來作注音修正。最後，將所得的語音檔播放出來即可。

### 三．操作介面說明：

本系統的操作說明如下：

1. 開始執行程式時，會以語音提示使用者使用語音查詢的方法，按數字鍵 1 進行書名查詢，按 2 進行作者查詢，按 3 進行出版社查詢

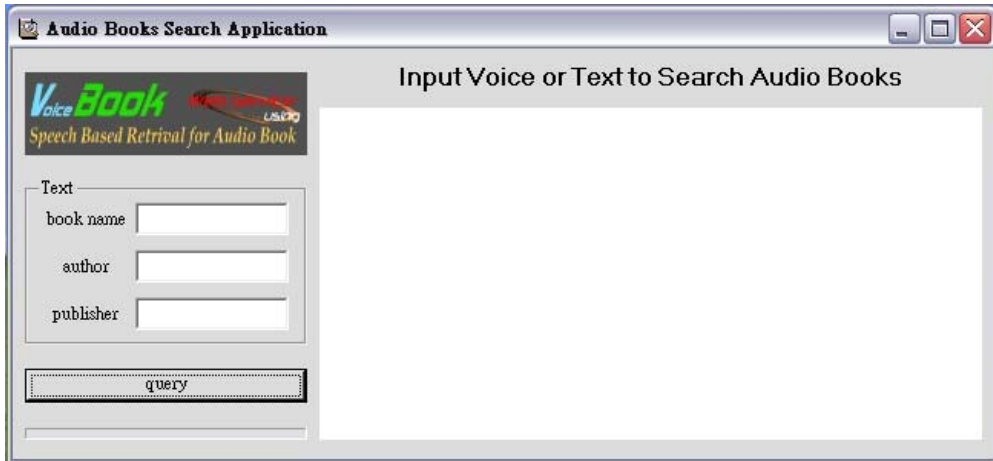


圖 2. 開始執行時的使用者介面

2. 在按下數字鍵之後，語音會提示在嗶聲後開始 3 秒錄音，而嗶聲響起，左下角的 progress bar 會開始進行，在此時對著麥克風說出要查詢的關鍵字。



圖 3. 錄音時, progress bar 正在移動

3. 錄完音之後，將會傳回辨識結果，不但將結果顯示出來，也利用語音合成的技術，將答案唸出來。若想聽該筆有聲書，可以按下 play 鈕或是按下指定的數字鍵，系統將會下載有聲書並播放出來。





圖 4. 所得的輸出結果，除了印出來外也會唸出來

4. 本系統也可以使用文字查詢，可直接在欄位上填入想查詢的資料

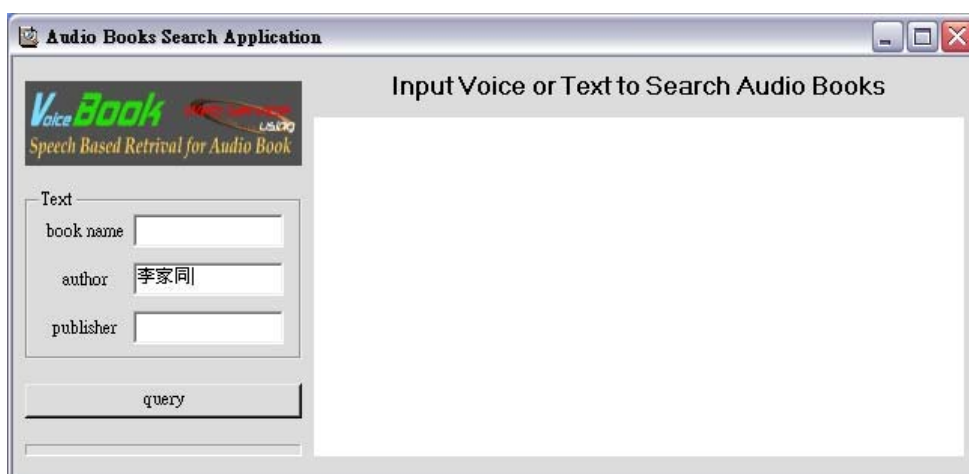


圖 5. 將關鍵字填入欲查詢的欄位中

#### 四·實驗成果：

本系統的目的地在於發展出語音和文字二種輸入和輸出方式，更優於一般資料庫只有文字輸入的方式。如此將使得盲胞和不會中文輸入法的一般民眾能方便的使用。因此，最重要的就是語音輸入的辨識率和系統操作的方便性。就辨識率而言，由於本語音辨識系統是採用最接近的句子當做辨識結果。被辨識系統資料的

多寡，平均每句的字數，都會影響正確率。下表是我們測試的結果：

資料庫	總共筆數	平均每筆字數	測試次數	正確次數	辨識率
書名	12147	7.24	113	95	80.53%
作者	5277	4.33	105	79	75.28%
出版社	1091	3.96	127	104	81.89%

從上表可得知，本系統的辨識率已達到大約 80% 的水準，已可達到方便使用的地步。而方便性方面，由於完全是以簡單的數字鍵和聲音來做輸入，所以對視障者和一般人來說，都能夠輕易地操作。經由實際全盲的人測試，印證了我們的想法。

#### 五．結論：

在本論文中，我們已經實際設計出一套完整的有聲書查詢系統，這是基於兩種語音技術下的整合系統，利用 .NET 架構的 Web Service 來結合系統所應用的技術，經過多人的測試，其結果堪稱理想且實用價值極高。未來可以更進一步提高辨識率或加速比對時間以求系統更加完善。

#### 六．參考書目：

1. Xuedong Huang, Alex Acero, and Hsiao-Wuen Hon. "Spoken Language Processing." Prentice Hall
2. Fundamentals of Speech Recognition, Prentice Hall.
3. The HTK book, 2000, copyright for Microsoft Corporation.

## 線上新聞語音檢索系統

陳江村 羅瑞麟 張智星  
國立清華大學 資訊工程系  
新竹市光復路二段 101 號

E-mail : {[jtchen](mailto:jtchen@wayne.cs.nthu.edu.tw), [roro](mailto:roro@wayne.cs.nthu.edu.tw), [jang](mailto:jang@wayne.cs.nthu.edu.tw)}@wayne.cs.nthu.edu.tw

TEL: (03)5715131-3582

摘要：

在此報告中，我們實作了一個結合隱藏式馬可夫模型(Hidden Markov Model, HMM)為基礎的 HTK(HMM Toolkit)和網頁資料檢索技術的線上新聞語音資料檢索系統。一般的網頁資料檢索(如 google)須使用者輸入相關文字，才得以文字比對方式進行檢索。在此我們則嘗試加入語音辨識的技術讓使用者更易進行檢索。本系統分成新聞前處理及語音查詢兩階段。在辨識內容固定，高準確度的辨識結果下，本系統特別適用於手機、PDA、嵌入式系統等小型、不易以手操作輸入的裝置。本系統亦經清大盲友會的盲人朋友試用，反應十分良好。

關鍵詞:語音辨識、資料檢索、HMM、Viterbi Search、新聞檢索。

### 1 前言

目前的網際網路中，[www.google.com](http://www.google.com)[6]是每個人都不可或缺的工具，其提供的準確性和資料的可用性一直為人稱道，堪稱為文字檢索的翹楚，也因此，網際網路上的資訊成了一個無所不包的資料庫。而在此同時，相關的多媒體檢索技術也相繼發表[1]，顯示了多媒體方面的檢索需求。而由於語音的便利性和可用性(相對於以內容為主的影像檢索)，語音方面的檢索方法已成了多媒資訊檢索的重要研究。

語音檢索的方法可分為兩種，一為語音文件檢索(Spoken Document Retrieval)，一為語音文字檢索(Speech Recognition and Retrieval)。前者不考慮到語音模型，直接以語音的特徵參數，在另一語音文件中進行比對，希望找出最接近的語音內容。這樣的檢索方式雖可跨越語言模型，但在長時間的語音文

件中，辨識率並不高，並不符合高準確性的要求[11]。而語音文字檢索，則是在特定語音模型下，先進行語音辨識，再以辨識出來的文字進行檢索[3]。由於目前文字比對技術成熟，故此法的關鍵在於語音辨識的好壞，辨識文字內容則可十分廣泛。

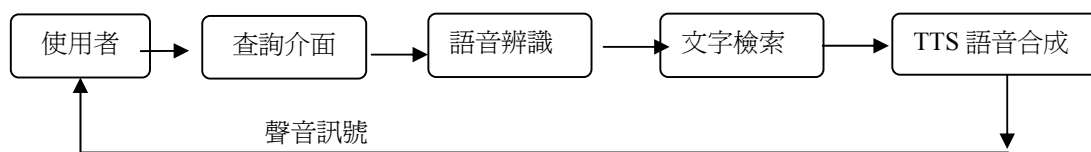
而所謂的語音辨識，主要是用來辨識出聲音的文字內容為何。一般來說，語音辨識的辨識成功率和辨識內容的範圍有很大關係。在大領域的文字辨識中，辨識結果往往出現相近但非正確的答案，這在目前仍是很難克服的問題。

目前此領域最有名的模型為 HMM。藉由特定語言語料的訓練，我們可以利用 HTK[7](Hidden Markov Model ToolKit)實作出某一特定領域的高準確度的語音辨識系統，比方說唐詩三百首的語音辨識，其準確率接近九成九[10]。

綜合以上觀點，我們實作了一個結合 HTK 和新聞網頁內容的檢索技術，希望能達到一個以語音為基礎的 News Google。

## 2 語音新聞檢索理論背景

本系統使用了語音辨識、文字比對和語音合成三種技術。流程圖如下：

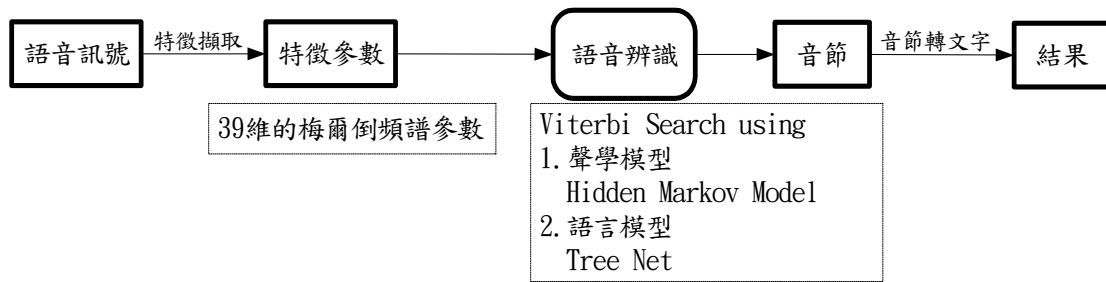


圖表 1 語音新聞檢索服務流程

文字比對部份，由我們只對新聞標題部份作比對，因此以下以語音辨識和合成作為介紹重點。

### 2.1 語音辨識部份

一般而言，語音辨識的演算法流程如下圖所示：

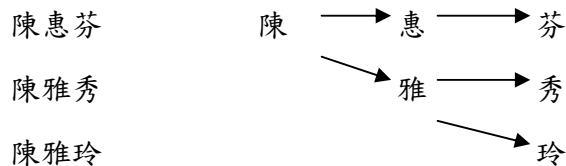


圖表 2 語音辨識的演算法

其中所採用的方法和理論如下：

### 1. 語言模型 Tree Net

把每個單音節視為一個節點，節點和節點間相連關係的樹狀結構。圖例如下：

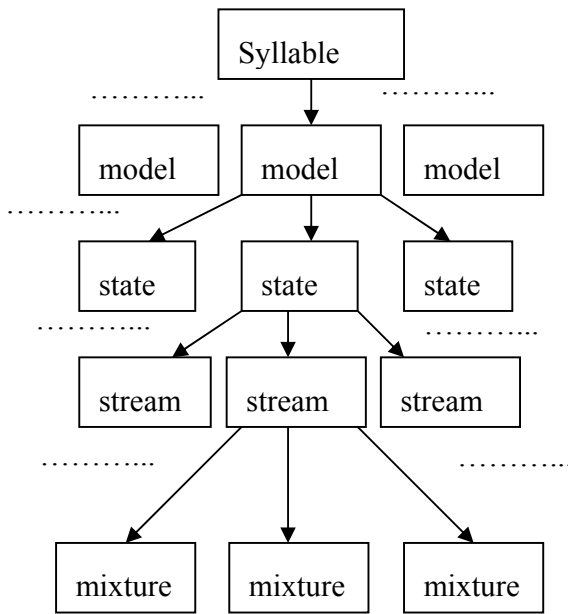


圖表 3 Tree Net 檔示意

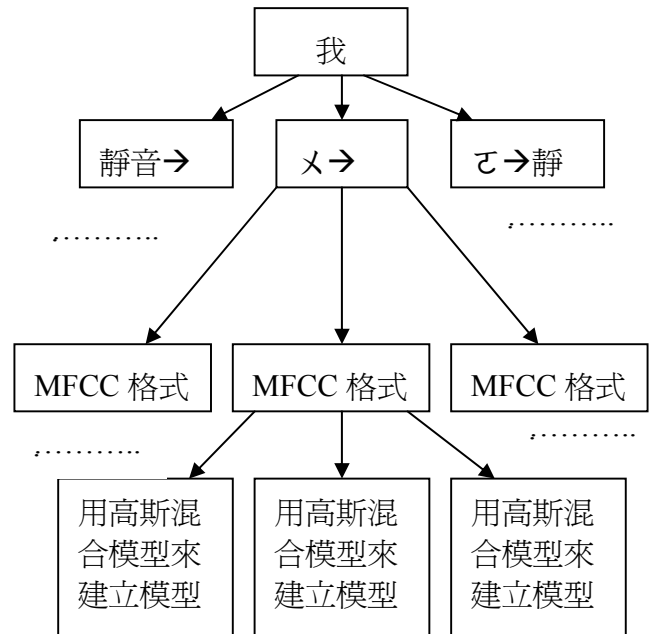
### 2. 聲學模型 Hidden Markov Model

隱藏式馬可夫模型基本上是一種雙重隨機過程，而之所以稱為隱藏式是因為其中有一組隨機過程是隱藏的，看不見的，在語音中就如同人類在發聲的過程中其發聲器官狀態變化是看不見的，好比喉嚨、舌頭與口腔的變化是不可能從可觀測的語音訊號序列看出來的。而另一組隨機過程稱為觀測序列 (observation sequence)，它是由狀態觀測機率 (state observation probability) 來描述在每個狀態下觀測到各種語音特徵參數的機率分佈。HMM 的狀態觀測機率函式  $b_j(o_t)$  是採用高斯混合密度函數或稱高斯混合模型 (Gaussian Mixture Model, GMM) 來計算連續機率密度，因此每一個聲音單元 (Model) 皆有一組 Continuous HMM 參數。

圖表 4 為 Model, State, Stream 和 Mixture 的階層示意圖，圖表 5 則以”我”此一 syllable 為例，示範 CHMM 的建立方式。



圖表 6 Model, State, Stream 和 Mixture 示意圖



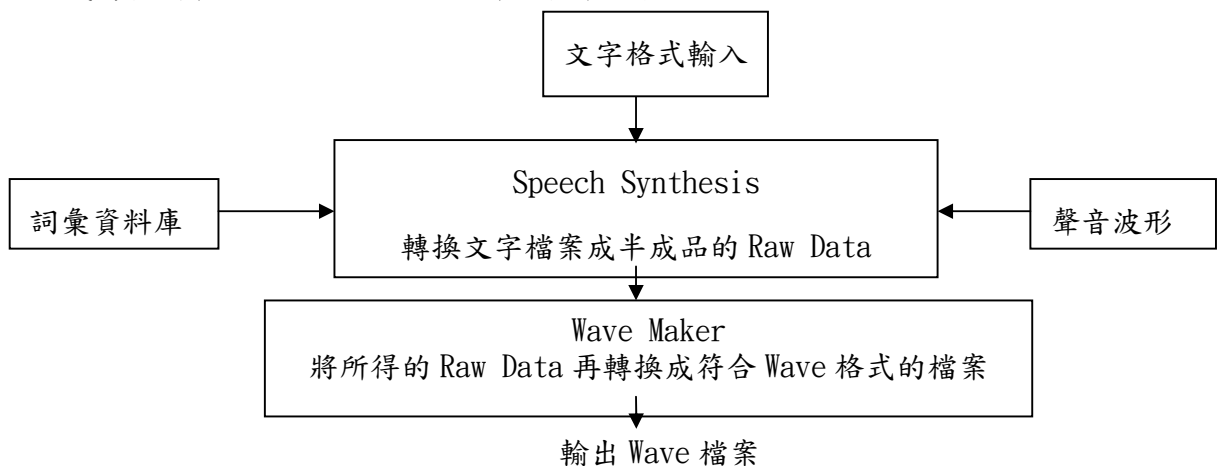
圖表 7 以 GMM 建立 syllable 的 CHMM 流程示意圖

### 3. 辨識方法

我們根據 Tree Net 的路徑進行以 Viterbi Search，以辨識出機率最高的路徑。其中我們也加上了 Beam Search(光束搜尋法)的作法以進行加速[2]。光速搜尋法在搜尋過程中會慢慢丟棄低機率的搜尋目標，使得愈後面的比對速度會愈加快，此法可有效減少搜尋時間且不會犧牲太多準確性[5]。

## 2.2 語音合成

在輸出方面我們使用和黃紹華老師合作的語音合成技術。此合成方式是連接式的合成為基礎(Concatenation-Based)，基本流程如下：

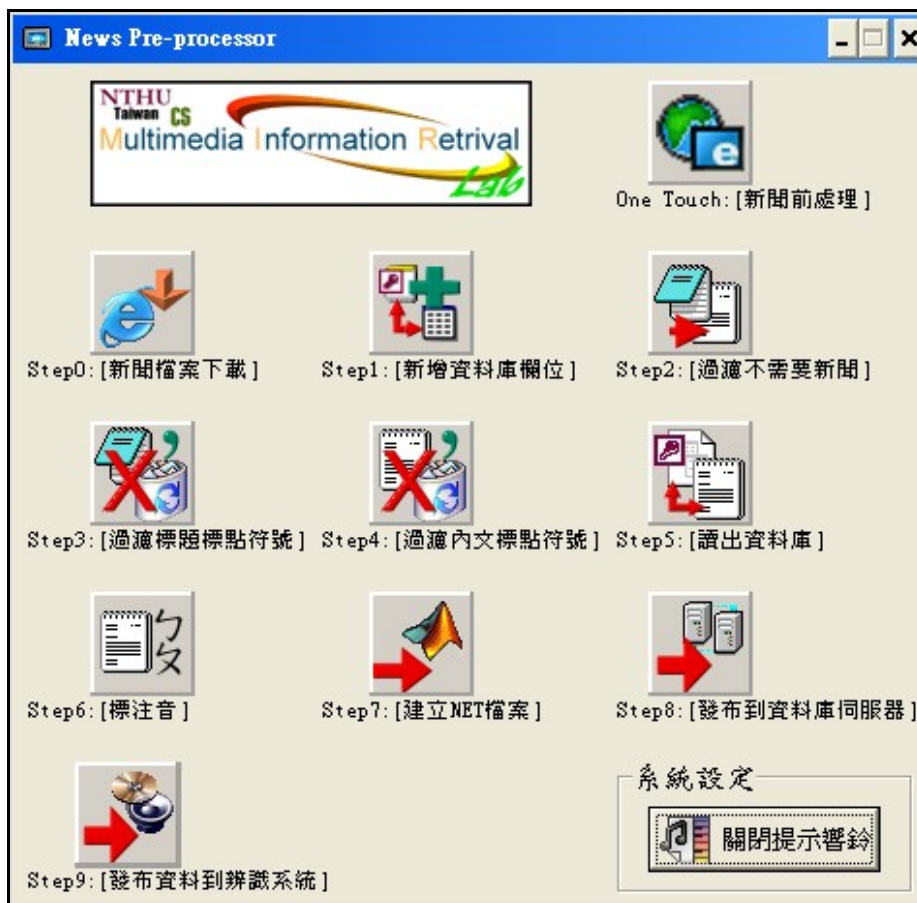


圖表 8 語音合成流程

### 3 語音新聞檢索系統架構

本系統分為兩大部份：新聞前處理及語音查詢新聞，分別介紹如下。

#### 1. 新聞前處理



圖表 9 新聞前處理介面

如圖表 10 所示，本程式分為十大步驟。【新聞前處理】按鈕則是按下後即可執行 Step0 ~ Step9。以下依序介紹各功能：

#### (1) 【新聞檔案下載】

以 PERL 程式，從網路上抓取當日的新聞，目前系統預設值為抓取中國時報、台灣新生報、中央社新聞、新浪網新聞等四家網站的新聞。

#### (2) 【新增資料庫欄位】

做完上一步驟抓取當日新聞完成後，在 Access 資料庫中新增兩個欄位，即 news\_title\_pure 及 news\_content\_pure，以便接下來的

處理。

(3) 【過濾不需要新聞】

過濾漁業氣象這類型的新聞，如果不需要也可以省略這步驟。

(4) 【過濾標題標點符號】

刪除標題標點符號，前處理系統才可以對純中文字進行標注音。將過濾完的文字存放於資料庫的 news\_title\_pure 欄位。

(5) 【過濾內文標點符號】

刪除內文標點符號，前處理系統才可以對純中文字進行標注音。將過濾完的文字存放於資料庫的 news\_content\_pure 欄位。

(6) 【讀出資料庫】

在進行標注音前將資料庫新增的兩欄位內的資料轉成文字檔。

(7) 【標注音】

針對從資料庫轉出來的文字檔進行標注音。

(8) 【建立 NET 檔案】

針對標好注音的檔案建立 Tree Net，以供語音辨識程式查詢。

(9) 【發布到資料庫伺服器】

更新資料庫伺服器的資料。

(10) 【發布資料到辨識系統】

更新辨識系統的辨識核心。

## 2. 語音查詢新聞

我們以 Borland C++ Builder 5.0 建構新聞語音查詢介面，如圖表 11。此介面分成標題查詢及內文查詢兩部份，顧名思義，標題查詢為找符合關鍵字的標題，而內文查詢則是只要內文有句子符合關鍵字即會顯示出來。以下介紹操作時大略的流程：

(1) 按下一個檢索按鈕，系統會以語音的方式提示使用者準備錄音，錄音時間為三秒鐘。

(2) 錄完音後辨識系統則開始辨識語音，而後將結果顯示在偏上方的白色區塊內。



(3)若使用者對查詢出來的新聞感興趣，則點選該條新聞後，偏下方的白色區塊即出現對應的新聞內容。

(4)使用者也可經由在下方白色區塊中按下滑鼠左鍵來聽取新聞的內容，該內容是以語音合成的方式即時產生的。



圖表 12 新聞語音查詢介面

#### 4 結論

在本篇報告中，我們介紹了一個「線上新聞語音資料檢索系統」。歸納結果，在此列出此系統的特性：

1. 語音輸入：不鍵盤等須其他工具，即可將查詢內容輸入。
2. 快速檢索：藉由 offline 的標題索引和即時的語音辨識、文字比對，提供新聞標題的快速檢索。
3. 語音輸出：使用 Text-To-Speech 的語音合成，將查詢所得新聞進行播報。
4. 定時更新：每日固定時間更新網頁上即時新聞。

新聞語音查詢系統能讓網際網路的使用者有更多的方便。本系統的語音查詢有相當不錯的辨識率，而語音合成的部份表現也不會太糟，相信若用於 PDA、手機等嵌入式的系統，會是個方便的工具。

線上新聞語音資料檢索系統雖然有許多優點，不過未來仍存在許多問題須要克服，例如解決斷詞的問題(文字辨識準確度)和分散式處理(大量使用者下的效率問題)等等，這些都是我們未來的工作。

## 5 參考資料

- [1] J. -S. Roger Jang, Jiang-Chun Chen, Ming-Yang Kao, "MIRACLE: A Music Information Retrieval System with Clustered Computing Engines", International Symposium on Music Information Retrieval (MUSIC IR 2001)
- [2] Jang, J. -S. Roger and Lin, Shiuan-Sung, "Optimization of Viterbi Beam Search in Speech Recognition", International Symposium on Chinese Spoken Language Processing, Taiwan, August 2002.
- [3] Lawrence Rabiner, B.H Juang, Fundamentals of speech recognition, Prentice Hall, 1993.
- [4] O' Shanughnessy, D., Speech Communication : human and machine, Addison-Wesley, 1987.
- [5] Rabiner, L. and Juang, B.-W., Fundamentals of Speech Recognition. Prentice Hall PTR, Upper Saddle River, New Jersey, 1993
- [6] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual Web search engine. In Ashman and Thistlewaite [2], pages 107-117. Brisbane, Australia. <http://citeseer.nj.nec.com/brin98anatomy.html>
- [7] Steven Young, The HTK Book version 3, Microsoft Corporation, 2000.
- [8] T.W. Parsons, Voice and Speech Processing, McGraw-Hill, 1986.
- [9] 中文文句翻語音之韻律訊息合成，交大電信博士論文，黃紹華。
- [10] 林玄松，“Viterbi 搜尋的最佳化以及多語系辨識”，清華大學碩士論文，民國九十年。
- [11] 謝宏坤，“語音說明中搜尋任意定義之關鍵詞的研究”，台灣科技大學碩士論文，民國 89 年