Using Word-Pair Identifier to Improve Chinese Input System

Jia-Lin Tsai

Tung Nan Institute of Technology, Department of Information Management Taipei 222, Taiwan, R.O.C.

tsaijl@mail.tnit.edu.tw

Abstract

This paper presents a word-pair (WP) identifier that can be used to resolve homonym/segmentation ambiguities and perform syllable-to-word (STW) conversion effectively for improving Chinese input systems. The experiment results show the following: (1) the WP identifier is able to achieve tonal (syllables with four tones) and toneless (syllables without four tones) STW accuracies of 98.5% and 90.7%, respectively, among the identified word-pairs; (2) while applying the WP identifier, together with the Microsoft input method editor 2003 and an optimized bigram model, the tonal and toneless STW improvements of the two input systems are 27.5%/18.9% and 22.1%/18.8%, respectively.

1 Introduction

More than 100 Chinese input methods have been developed in the past (Becker 1985, Huang 1985, Gu et al. 1991, Chung 1993, Kuo 1995, Fu et al. 1996, Lee et al. 1997, Hsu et al. 1999, Chen et al. 2000, Tsai and Hsu 2002, Gao et al. 2002, Lee 2003). Their underlying approaches can be classified into four types: (1) Optical character recognition (OCR) based (Chung 1993), (2) Online handwriting based (Lee et al. 1997), (3) Speech based (Fu et al. 1996, Chen et al. 2000), and (4) Keyboard based consists of phonetic and pinyin based (Chang et al. 1991, Hsu et al. 1993, Hsu 1994, Hsu et al. 1999, Kuo 1995, Lua and Gan 1992); arbitrary codes based [Fan et al. 1988]; and structure scheme based (Huang 1985).

Currently, the most popular method for Chinese input is phonetic and pinyin based, because Chinese people are taught to write the corresponding phonetic and pinyin syllables of each Chinese character and word in primary school. In Chinese, each Chinese character corresponds to at least one syllable; and each Chinese word can be a mono-syllabic word, such as "鼠 (mouse)", a bi-syllabic word, such as "袋鼠 (kangaroo)", or a multi-syllabic word, such as "米老鼠(Mickey mouse)." Although there are more than 13,000 distinct Chinese characters (of which 5,400 are commonly used), there are only about 1,300 distinct syllables. As per (Qiao et al. 1984), each Chinese syllable can be mapped from 3 to over 100 Chinese characters, with the average number of characters per syllable being 17. According to our computation, the minimum, maximum and average numbers f Chinese words per syllable-word in MOE-MANDARIN dictionary "教育部國語辭典" (one of most commonly-used Chinese dictionaries published by the Ministry of Education in Taiwan, its online dictionary is at (MOE)) are 1, 22 and 1.5, respectively. Since the size of problem space for syllable-to-word conversion is much less than that of syllable-to-character conversion, the most existing Chinese input systems (Hsu 1994, Hsu et al. 1999, Tsai and Hsu 2002, Gao et al. 2002, MSIME) are addressed on syllable-toword conversion, not syllable-to-character conversion. To the research field of Chinese speech recognition, the STW conversion is the main task of Chinese language processing in typical Chinese speech recognition systems (Fu et al. 1996, Lee et al. 1993, Chien et al. 1993, Su et al. 1992).

Conventionally, there are two approaches for syllable-to-word (STW) conversion: (1) the **lin-guistic approach** based on syntax parsing, se-

mantic template matching and contextual information (Hsu 1994, Fu et al. 1996, Hsu et al. 1999, Kuo 1995, Tsai and Hsu 2002); and (2) the statistical approach based on the n-gram models where n is usually 2 or 3 (Lin and Tsai 1987, Gu et al. 1991, Fu et al. 1996, Ho et al. 1997, Sproat 1990, Gao et al. 2002, Lee 2003). Although the linguistic approach requires considerable effort in designing effective syntax rules, semantic templates or contextual information, it is more user-friendly than the statistical approach on understanding why such a system makes a mistake (Hsu 1994, Tsai and Hsu 2002). On the other hand, the statistical language model (SLM) used in the statistical approach requires less effort and has been widely adopted in commercial Chinese input systems.

According to previous studies (Chung 1993, Fong and Chung 1994, Tsai and Hsu 2002, Gao *et al.* 2002, Lee 2003), *homophone selection* and *syllable-word segmentation* are two critical problems to the STW conversion in Chinese. Incorrect homophone selection and failed syllable-word segmentation will directly influence the STW conversion rate. For example, consider the syllable sequence "yi1 du4 ji4 yu2 zhong1 guo2 de5 niang4 jiu3 ji4 shu4" of the sentence "一度(once)覬覦(covet)中國(China)的(of)釀酒 (making-wine) 技術 (technique)." As per the MOE-MANDARIN dictionary, the two possible syllable-word segmentations (in pinyin) are:

(F)"yi1/du4ji4/yu2/zhong1guo2/de5/niang4ji u3/ji4shu4"; and

(B)"yi1/du4/ji4yu2/zhong1guo2/de5/niang4ji u3/ji4shu4."

(We use the forward (F) and the backward (B) longest syllable-word first strategies (Chen *et al.* 1986, Tsai and Hsu 2002), and "/" to indicate a syllable-word boundary).

Among the above syllable-word segmentations, there is an ambiguous syllable-word section: /du4ji4/yu2/ (/{妒忌}/{于,圩,余,於,盂,俞,娛,魚, 愉,渝,腴,萸,隅,畬,愚,榆,瑜,虞,逾,漁,諛,餘, 輿}/); and /du4/ji4yu2/ (/{妒,杜,肚,度,渡,鍍, 蠧}/{覬覦,鯽魚}/), respectively. For the ambiguous syllable-word section, the set of wordpairs comprised of two multi-syllabic Chinese words (including bi-syllabic words in the following) and their corresponding word-pair frequencies found in the UDN2001 corpus are: {-度-覬覦(1), 一度-中國(1), 一度-技術(4), 覬覦-

中國(1), 覬覦-技術(1), 中國-技術(26), 釀酒-技 術(19)}. The UDN2001 corpus (Tsai and Hsu 2002) is a collection of 4,539,624 Chinese sentences extracted from whole 2001 articles on the United Daily News Website (UDN) in Taiwan. For this case, if the word-pair "中國(China)-技 術(technique)" with the maximum frequency 26 is used to be the key word-pair, the set of cooccurrence word-pairs with the key word-pair found in the UDN2001 will be {一度-覬覦, -度-中國,一度-技術, 覬覦-中國, 覬覦-技術 }. Then, by the key word-pair "中國-技術" and its co-occurrence word-pair set {一度-覬覦, 一度-中國,一度-技術, 覬覦-中國, 覬覦-技術}, the mentioned ambiguous syllable-word section (/du4ji4/yu2/ and /du4/ji4yu2/) and the homophone selection of syllable-word /ji4 shu4/ (/{技 術(technique),計數(count)}/) of this case can be resolved, simultaneously. Thus, the Chinese words "一度(once)", "覬覦(covet)", "中國 (China)" and "技術(technique)" in the syllable sequence "yi1 du4 ji4 yu2 zhong1 guo2 de5 niang4 jiu3 ji4 shu4" can then be correctly identified. If we use the Microsoft Input Method Editor 2003 for Traditional Chinese (MSIME) to translate the syllables, it will be converted into "一度(once)繼(continue)於(to)中國(China)的 (of)釀酒(making-wine)技術(technique)." As per (Gao et al. 2002), MSIME is a trigam-like Chinese input system. The two error converted words "繼(continue)" and "於(to)" are widely recognized that unseen event (覬覦-中國) and over-weighting (於-中國) the two major problems of SLM systems (Fu et al. 1996, Gao et al. 2002).

The objective of this study is to illustrate the effectiveness of word-pairs for resolving the STW conversion for improving the Chinese input systems. We also conduct STW experiments to show the tonal and toneless STW accuracies of a commercial input product and a bigram model can be improved by our word-pair identifier without a tuning process. Here, the "tonal" is to indicate the syllables input with four tones, such as "niang4(釀) jiu3(酒) ji4(技) shu4(術)" and the "toneless" is to indicate the syllables input without four tones, such as "niang4(釀) jiu3(酒) jii4($\frac{1}{2}$) shu4($\frac{1}{2}$) jiu($\frac{1}{2}$) jii($\frac{1}{2}$) shu($\frac{1}{2}$)."

The remainder of this paper is arranged as follows. In Section 2, we present a method for auto-generating word-pair (AUTO-WP) database from Chinese sentences. Then, we develop a word-pair identifier with the WP database to effectively resolve homonym and segmentation ambiguities of STW conversion on the WPrelated portion in Chinese syllables. In Section 3, we present our STW experiment results. Finally, in Section 4, we give our conclusions and suggest some future research directions.

2 Development of Word-Pair Identifier

The system dictionary of our word-pair identifier is comprised of 155,746 Chinese words taken from the MOE-MANDARIN dictionary (MOE) and 29,408 unknown words auto-found in UDN2001 corpus by a Chinese word autoconfirmation (CWAC) system (Tsai *et al.* 2003). The system dictionary provides the knowledge of words and their corresponding pinyin syllable-words. The pinyin syllable-words were translated by phoneme-to-pinyin mappings, such as " $U - \chi$ "-to-"ji4."

2.1 Generating the Word-Pair Database

The steps of our AUTO-WP to auto-discovery word-pairs from a given Chinese sentence are as below:

- Step 1. Segmentation: Generate the word segmentation for a given Chinese sentence by backward maximum matching (BMM) techniques (Chen et al. 1986) with the system dictionary. Take the Chinese sentence "將軍用零件帶來(bring the military component parts here)" as an example. Its BMM word-segmentation is "將(get)/軍用(military)/零件(component parts)/ 帶 來 (bring)" and its forward maximum matching (FMM) wordsegmentation is "將軍(a general)/用(use)/ 零件 (component parts)/帶來 (bring)." According to our previous work (Tsai et al. 2004), the word segmentation precision of BMM is about 1% greater than that of FMM.
- *Step 2.* **Initial WP set**: Extract all the combinations of word-pairs from the word segmentations of Step 1 to be the initial

WP set. For the above case, there are six combinations of word-pairs extracted: {"將/軍用", "將/零件", "將/帶來", "軍 用/零件", "軍用/帶來", "零件/帶來"}.

Step 3. Final WP set: Select out the wordpairs comprised of two multi-syllabic Chinese words to be the finial WP set. For the final WP set, if the word-pair is not found in the WP database, insert it into the WP database and set its frequency to 1; otherwise, increase its frequency by 1. In the above case, the final WP set includes three word-pairs: {"軍用 /零件", "軍用/帶來", "零件/帶來"}.

By applying our AUTO-WP to the UDN2001 corpus (the training corpus), totally 25,439,679 word-pairs were generated. From the generated WP database, the frequencies of word-pairs "軍 用/零件", "軍用/帶來" and "零件/帶來" are 1, 1 and 2, respectively. The frequency of a wordpair is the number of sentences that contain the word-pair with the same word-pair order in the training corpus.

2.2 Word-Pair Identifier

The algorithm of our WP identifier for a given Chinese syllables is as follows:

- Step 1. Input tonal or toneless syllables.
- *Step 2*. Generate all possible word-pairs comprised of two multi-syllabic Chinese words for the input syllables to be the input of Step 3.
- Step 3. Select out the word-pairs that match a word-pair in the WP database to be the initial WP set, firstly. Then, from the initial WP set, select the word-pair with maximum frequency as the key word-pair. Finally, find the co-occurrence wordpairs with the key word-pair in the training corpus to be the final WP set. If there are two or more word-pairs with the same maximum frequency, one of them is randomly selected as the key word-pair.
- *Step 4*. Arrange all word-pairs of the final WP set into a WP-sentence. If no word-pairs can be identified in the input syllables, a NULL WP-sentence is produced.

Table 1 is a step by step example to show the details of applying our WP identifier on the Chinese syllables "yi1 ge5 wen2 ming2 de5 shuai1 wei2 guo4 cheng2(一個 [a] 文明 [civilization]的[of]衰微[decay]過程[process])." For this case, we have a WP-sentence "一個文 明 de5shuai1wei2 過程." As we have mentioned in Section 1, we found this WP-sentence can also be used to correct the MSIME converted errors in its output "一個[a]閳名[famous] 的[of]衰微[decay]過程[process]."

Table 1. An illustration of a WP-sentence generation for the Chinese syllables "yil ge5 wen2 ming2 de5 shuail wei2 guo4 cheng2(一個[a]文明[civilization] 的[of]衰微[decay]過程[process])"

Step #	Results
Step.1	yi1 ge5 wen2 ming2 de5 shuai1 wei2 guo4 cheng2
	(一個文明的衰微過程)
Step.2	The found word-pair / word-pair frequency:
	一個(yi1 ge5)-文明(wen2 ming2)/9
	一個(yi1 ge5)-聞名(wen2 ming2)/1
	一個(yi1 ge5)-衰微(shuai1 wei2) / 0
	一個(yi1 ge5)-過程(guo4 cheng2) / 65
	文明(wen2 ming2)- 衰微(shuai1 wei2) / 0
	文明(wen2 ming2)-過程(guo4 cheng2)/3
	衰微(shuai1 wei2) -過程(guo4 cheng2) / 0
Step.3	The key word-pair:
	一個(yi1 ge5)-過程(wen2 ming2)
	The co-occurrence word-pairs:
	一個(yi1 ge5)-文明(wen2 ming2)
	文明(wen2 ming2)-過程(guo4 cheng2)
Step.4	WP-sentence:
	一個文明 de5 shuai1 wei2 過程

3 The STW Experiments

To evaluate the STW performance of our WP identifier, we define the STW accuracy, identified character ratio (ICR) and STW improvement, by the following equations:

STW accuracy = # of correct characters / # of total characters. (1)

Identified character ratio (ICR) = # of characters of identified WP / # of total characters in testing sentences. (2)

STW improvement (i.e. STW error reduction rate) = (accuracy of STW system with WP –

accuracy of STW system)) / (1 – accuracy of STW system). (3)

3.1 Generation of the Word-Pair Database

To conduct the STW experiments, firstly, use the inverse translator of phoneme-to-character (PTC) provided in GOING system to convert testing sentences into their corresponding syllables. Then, all the error PTC translations of GOING were corrected by post human-editing. Then, apply our WP identifier to convert these testing syllables back to their WP-sentences. Finally, calculate its STW accuracy and identified character ratio by Equations (1) and (2). Note that all test sentences are composed of a string of Chinese characters in this study.

The training/testing corpus, closed/open test sets and the testing WP database used in the STW experiments are described as below:

- (1) Training corpus: We used the UDN2001 corpus mentioned in Section 1 as our training corpus. All knowledge of word frequencies, word-pairs, word-pair frequencies was autogenerated and computed by this corpus.
- (2) Testing corpus: The UDN2002 corpus was selected as our testing corpus. It is a collection of 3,321,504 Chinese sentences that were extracted from whole 2002 articles on the United Daily News Website (UDN).
- (3) Closed test set: 10,000 sentences were randomly selected from the UDN2001 corpus as the closed test set. The {minimum, maximum, and mean} of characters per sentence for the closed test set were {4, 37, and 12}.
- (4) Open test set: 10,000 sentences were randomly selected from the UDN2002 corpus as the open test set. At this point, we checked that the selected open test sentences were not in the closed test set as well. The {minimum, maximum, and mean} of characters per sentence for the open test set were {4, 43, and 13.7}.
- (5) Testing WP database: By applying our AUTO-WP on the UDN2001 corpus, we created 25,439,679 word-pairs as the testing WP database.

We conducted the STW experiment in a progressive manner. The results and analysis of the experiment are described in Sub-sections 3.2 and 3.3.

3.2 STW Experiment of the WP Identifier

The purpose of this experiment is to demonstrate the tonal and toneless STW accuracies among the identified word-pairs by using the WP identifier with the testing WP database.

From Table 2, the average tonal and toneless STW accuracies of the WP identifier for the closed and open test sets are 98.5% and 90.7%, respectively. Between the closed and the open test sets, the differences of the tonal and tone-less STW accuracies of the WP identifier are 0.5% and 1.4%, respectively. These results strongly support that the WP identifier can be used to effectively perform Chinese STW conversion on the WP-related portion.

 Table 2. The results of the tonal and toneless STW

 experiment for the WP identifier on the identified

 word-pairs

	Closed	Open	Average (ICR)
Tonal	98.7%	98.2%	98.5% (47%)
Toneless	91.4%	90.0%	90.7% (39%)

3.3 A Commercial IME System and A Bigram Model with WP Identifier

We selected Microsoft Input Method Editor 2003 for Traditional Chinese (MSIME) as our experimental commercial Chinese input system. In addition, an optimized bigram model called BiGram was developed. The BiGram STW system is a bigram-based model developing by SRILM (Stolcke 2002) with Good-Turing back-off smoothing (Manning and Schuetze, 1999), as well as forward and backward longest syllable-word first strategies (Chen *et al.* 1986, Tsai *et al.* 2004). The training corpus and system dictionary of the BiGram system are same with that of the WP identifier. All the bigram probabilities were calculated by the UDN2001 corpus.

Table 3a compares the results of MSIME and MSIME with the WP identifier on the closed and open test sentences. Table 3b compares the results of BiGram and BiGram with the WP identifier on the closed and open test sentences. In this experiment, the STW output of the MSIME with the WP identifier, or the BiGram with the WP identifier, was collected by directly replacing the identified word-pairs (WP-sentences) from the corresponding STW output of MSIME or BiGram.

 Table 3a. The results of the tonal and toneless STW

 experiment for the MSIME and the MSIME with the

 WP identifier

	MSIME	MSIME+WP ^a	Improvement
Tonal	94.9%	96.3%	27.5%
Toneless	86.9%	89.8%	22.1%
a	• •	1 1 1	11 1 1000 00

^a STW accuracies of the words identified by the MSIME with the WP identifier

Table 3b. The results of the tonal and toneless STWexperiment for the BiGram and the BiGram with theWP identifier

	BiGram	BiGram+WP ^a	Improvement
Tonal	96.3%	97.0%	18.9%
Toneless	8 86.2%	88.8%	18.8%
a communi		1 1 1 1 1 1 1 1 1 1	I DIG

^a STW accuracies of the words identified by the BiGram with the WP identifier

From Table 3a, the tonal and toneless STW improvements of MSIME by using the WP identifier are 27.5% and 22.1%, respectively. Meanwhile, from Table 3b, the tonal and toneless STW improvements of BiGram by using the WP identifier are 18.9% and 18.8%, respectively. (Note that we also developed a TriGram STW system with the same source and techniques of BiGram. However, the differences between the tonal and toneless STW accuracies of BiGram and TriGram are only about 0.2%)

To sum up the results of this experiment, we conclude that the WP identifier can achieve a better STW accuracy than that of the MSIME and BiGram systems on the WP-related portion. The results of Tables 3a and 3b indicate that the WP identifier can effectively improve the tonal and toneless STW accuracies of MSIME and BiGram without tuning processing. Appendix A presents two cases of STW results that were obtained from the experiment.

3.4 Error Analysis of the STW Conversion

We examine the Top 300 cases in the *tonal* and *toneless* STW conversion errors, respectively, from the open testing results of BiGram with the WP identifier. As per our analysis, the problems of STW conversion errors can be classified into three major types:

(1) Unknown word problem: For any Chinese NLP system, unknown word extraction is one of the most difficult problems and a critical issue (Tsai *et al.* 2003). When an error is caused only by the lack of words in the system dictionary, we call it unknown

word problem.

- (2) *Inadequate syllable segmentation problem*: When an error is caused by syllable-word overlapping (or say ambiguous syllableword segmentation), instead of an unknown word problem, we call it *inadequate syllable segmentation*.
- (3) *Homophones problem*: These are the remaining STW conversion errors.

 Table 4. The coverage of three problems caused the tonal and toneless STW conversion errors

Problems	Coverage (%)		
	Tonal	Toneless	
Unknown Word	12%	11%	
Inadequate Syllable	36%	51%	
Segmentation			
Homophone	53%	39%	
a	1 1 1 1 1 1 1 1 1	1 1 1	

^a STW accuracies of the words identified by the BiGram with the WP identifier

Table 4 is the coverage of the three problems. From Table 4, we have two observations:

- (1) The coverage of unknown word problem for tonal and toneless STW systems is similar. Since the unknown word problem is not specifically a STW problem, it can be easily taken care of through manual editing or semi-automatic learning during input. In practice, therefore, the tonal and toneless STW accuracies could be raised to 98% and 91%, respectively. Although some of unknown words have been incorporated in the system dictionary by a CWCA system (Tsai *et al.* 2004), they could still face the problems: *inadequate syllable segmentation* and *failed homophone disambiguation*.
- (2) The major problem caused error conversions in tonal and toneless STW systems is different. To improve tonal STW systems, the major targets should be the cases of failed homophone selection (53% coverage). For toneless STW systems, on the other hand, the cases of inadequate syllable segmentation (51% coverage) should be the focus for improvement.

To sum up the above two observations, the bottlenecks of the STW conversion lie in the second and third problems. To resolve these issues, we believe one simple and effective approach is to extend the size of WP database, because our experiment results show that the WP identifier can achieve better tonal and toneless STW accuracies than those of MSIME and BiGram on the WP-related portion.

4 Conclusion and Future Directions

In this paper, we have applied a WP identifier to support the Chinese language processing on the STW conversion and obtained a high STW accuracy on the identified word-pairs. All of the WP data can be generated fully automatically by applying the AUTO-WP on the system and user corpus. We are encouraged by the fact that WP knowledge can achieve tonal and toneless STW accuracies of 98.5% and 90.7%, respectively, for the WP-related portion on the testing syllables. The WP identifier can be easily integrated into existing Chinese input systems by identifying word-pairs in a post-processing step. Our experimental results show that, by applying the WP identifier together with MSIME (a trigram-like model) and BiGram (an optimized bigram model), the tonal and toneless STW improvements of the two Chinese input systems are 27.5%/22.1% and 18.9%/18.8%, respectively. For adaptation STW approach, we have tried to apply the AUTO-WP to extract the word-pairs from the 10,000 open testing sentences into the testing WP database, the tonal and toneless STW accuracies of the MSIME with the adaptation WP identifier and the Bi-Gram with the adaptation WP identifier will become 97.0%/97.2% and 91.1%/90.0%, respectively.

Currently, our approach is quite basic when more than one WP occurs in the same sentence. Although there is room for improvement, we believe it would not produce a noticeable effect as far as the STW accuracy is concerned. However, this issue will become important as we want to apply the WP knowledge to speech recognition. According to our computations, the collection of testing WP knowledge can cover approximately 50% and 40% of the characters in the UDN2001 and UDN2002 corpus, respectively.

We will continue to expand our collection of WP knowledge to cover more characters in the UDN2001 and UDN2002 corpus with Web corpus (search engine results) for improving our STW system. In other directions, we will try to improve our WP-based STW conversion with other statistical language models, such as HMM, and extend it to other areas of NLP, especially word segmentation and speech recognition.

Acknowledgement

We thank the Mandarin Promotion Council of the Ministry of Education in Taiwan for providing us the MOE-MANDARIN dictionary.

References

- Becker, J.D. 1985. Typing Chinese, Japanese, and Korean, *IEEE Computer* 18(1):27-34.
- Chang, J.S., S.D. Chern and C.D. Chen. 1991. Conversion of Phonemic-Input to Chinese Text Through Constraint Satisfaction, *Proceedings* of *ICCPOL'91*, 30-36.
- Chen, B., H.M. Wang and L.S. Lee. 2000. Retrieval of broadcast news speech in Mandarin Chinese collected in Taiwan using syllable-level statistical characteristics, *Proceedings of the 2000 International Conference on Acoustics Speech and Signal Processing*.
- Chen, C.G., Chen, K.J. and Lee, L.S. 1986. A model for Lexical Analysis and Parsing of Chinese Sentences, *Proceedings of 1986 International Conference on Chinese Computing*, 33-40.
- Chien, L.F., Chen, K.J. and Lee, L.S. 1993. A Best-First Language Processing Model Integrating the Unification Grammar and Markov Language Model for Speech Recognition Applications, *IEEE Transactions on Speech and Audio Processing*, 1(2):221-240.
- Chung, K.H. 1993. Conversion of Chinese Phonetic Symbols to Characters, M. Phil. thesis, Department of Computer Science, Hong Kong University of Science and Technology.
- Fong, L.A. and K.H. Chung. 1994. Word Segmentation for Chinese Phonetic Symbols, *Proceedings of International Computer Symposium*, 911-916.
- Fu, S.W.K, C.H. Lee and Orville L.C. 1996. A Survey on Chinese Speech Recognition, Communications of COLIPS, 6(1):1-17.
- Gao, J, Goodman, J., Li, M. and Lee K.F. 2002. Toward a Unified Approach to Statistical Language Modeling for Chinese, ACM Transactions on Asian Language Information Processing, 1(1):3-33.
- Gu, H.Y., C.Y. Tseng and L.S. Lee. 1991. Markov modeling of mandarin Chinese for decoding the phonetic sequence into Chinese characters, *Computer Speech and Language* 5(4):363-377.
- Ho, T.H., K.C. Yang, J.S. Lin and L.S. Lee. 1997. Integrating long-distance language modeling to

phonetic-to-text conversion, *Proceedings of ROCLING X International Conference on Computational Linguistics*, 287-299.

- Hsu, W.L. and K.J. Chen. 1993. The Semantic Analysis in GOING - An Intelligent Chinese Input System, *Proceedings of the Second Joint Conference of Computational Linguistics*, Shiamen, 1993, 338-343.
- Hsu, W.L. 1994. Chinese parsing in a phoneme-tocharacter conversion system based on semantic pattern matching, *Computer Processing of Chinese and Oriental Languages* 8(2):227-236.
- Hsu, W.L. and Chen, Y.S. 1999. On Phoneme-to-Character Conversion Systems in Chinese Processing, *Journal of Chinese Institute of Engineers*, 5:573-579.
- Huang, J.K. 1985. The Input and Output of Chinese and Japanese Characters, *IEEE Computer* 18(1):18-24.
- Kuo, J.J. 1995. Phonetic-input-to-character conversion system for Chinese using syntactic connection table and semantic distance, *Computer Processing and Oriental Languages*, 10(2):195-210.
- 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 Speech Dictation Machine for Chinese Language with Very Large Vocabulary, *IEEE Transaction on Speech and Audio Processing*, 1(2).
- Lee, C.W., Z. Chen and R.H. Cheng. 1997. A perturbation technique for handling handwriting variations faced in stroke-based Chinese character classification, *Computer Processing of Oriental Languages*, 10(3):259-280.
- Lee, Y.S. 2003. Task adaptation in Stochastic Language Model for Chinese Homophone Disambiguation, ACM Transactions on Asian Language Information Processing, 2(1):49-62.
- Lin, M.Y. and W.H. Tasi. 1987. Removing the ambiguity of phonetic Chinese input by the relaxation technique, *Computer Processing and Oriental Languages*, 3(1):1-24.
- Lua, K.T. and K.W. Gan. 1992. A Touch-Typing Pinyin Input System, *Computer Processing of Chinese and Oriental Languages*, 6:85-94.
- Manning, C. D. and Schuetze, H. 1999. Fundations of Statistical Natural Language Processing, MIT Press: 191-220.
- Microsoft Research Center in Beijing, "http://research.microsoft.com/aboutmsr/labs/be ijing/"
- MOE, MOE-MANDARIN online dictionary,

"http://140.111.34.46/dict/?open"

UDN, On-Line United Daily News, "http://udnnews.com/NEWS/"

- Qiao, J., Y. Qiao and S. Qiao. 1984. Six-Digit Coding Method, *Commun. ACM* 33(5):248-267.
- Sproat, R. 1990. An Application of Statistical Optimization with Dynamic Programming to Phonemic-Input-to-Character Conversion for Chinese, *Proceedings of ROCLING III*, 379-390.
- Stolcke A. 2002. SRILM An Extensible Language Modeling Toolkit, Proc. Intl. Conf. Spoken Language Processing, Denver.
- Su, K.Y., Chiang, T.H. and Lin, Y.C. 1992. A Unified Framework to Incorporate Speech and Language Information in Spoken Language Processing, *ICASSP-92*, 185-188.
- Tsai, J.L. and W.L. Hsu. 2002. Applying an NVEF Word-Pair Identifier to the Chinese Syllable-to-Word Conversion Problem, *Proceedings of 19th COLING 2002*, 1016-1022.
- Tsai, J,L, Sung, C.L. and Hsu, W.L. 2003. Chinese Word Auto-Confirmation Agent, *Proceedings* of ROCLING XV, 175-192.
- Tsai, J.L, Hsieh, G. and Hsu, W.L. 2004. Auto-Generation of NVEF knowledge in Chinese, *Computational Linguistics and Chinese Lan*guage Processing, 9(1):41-64.

Appendix A. Two STW results used in this study (The frequencies and English words in parentheses are included for explanatory purposes only)

Case I.

Tonal STW results for the Chinese tonal syllable input "ji2fu4qi2min2zu2te4se4" of the Chinese sentence "極富 (abundance)其(it)民族(folk)特色(characteristic)"

Methods	STW results
	民族/特色(13) (Key WP)
	極富/特色(11) (Co-occurrence WP)
WP-sentence	極富 qi2 民族特色
MSIME	及復其民族特色
MSIME+WP	極富 其 民族特色
BiGram	極富期民族特色
BiGram+WP	極富 期 民族特色

Toneless STW results for the Chinese toneless syllable input "jifuqiminzutese" of the Chinese sentence "極富 (abundance)其(it)民族(folk)特色(characteristic)"

Methods	STW results
	民族/特色(13) (Key WP)
	極富/特色(11) (Co-occurrence WP)
WP-sentence	極富 qi 民族特色
MSIME	及夫妻民族特色
MSIME+WP	極富 妻 民族特色
BiGram	及夫妻民族特色
BiGram+WP	極富 妻民族特色

Case II.

Tonal STW results for the Chinese tonal syllable input "cong2qian2shui3diao4yu2chong1lang4yang2fan2chu1hai 3you2yong3" of the Chinese sentence "從(from)潛水(dive) 釣魚(fishing)衝浪(surfing)揚帆(driving sail)出海(outward bound)游泳(swim)"

Methods	STW results
	出海/游泳(2) (Key WP)
	潛水/釣魚(1) (Co-occurrence WP)
	潛水/游泳(1) (Co-occurrence WP)
	釣魚/游泳(1) (Co-occurrence WP)
	揚帆/出海(1) (Co-occurrence WP)
WP-sentence	
C01	ng2 潛水釣魚 chong1 lang4 揚帆出海遊泳
MSIME	從前水釣魚衝浪揚帆出海游泳
MSIME+WP	從 潛水釣魚 衝浪揚帆出海遊泳
BiGram	從前水釣魚衝浪揚帆出海游泳
BiGram+WP	從 潛水釣魚 衝浪揚帆出海遊泳

Tonal STW results for the Chinese tonal syllable input "congqianshuidiaoyuchonglangyangfanchuhaiyouyong" of the Chinese sentence "從(from)潛水(dive)釣魚(fishing)衝 浪(surfing)揚帆(driving sail)出海(outward bound)游泳 (swim)"

Methods	STW results
	出海/游泳(2) (Key WP)
	潛水/釣魚(1) (Co-occurrence WP)
	潛水/游泳(1) (Co-occurrence WP)
	釣魚/游泳(1) (Co-occurrence WP)
	揚帆/出海(1) (Co-occurrence WP)
WP-sentence	
cor	ng2 潛水釣魚 chong1 lang4 揚帆出海遊泳
MSIME	從前水釣魚衝浪揚帆出海游泳
MSIME+WP	從 潛水釣魚 衝浪 <u>揚帆出海遊泳</u>
BiGram	從前水釣魚衝浪揚帆出海游泳
BiGram+WP	從 潛水釣魚 衝浪 揚帆出海遊泳