# Using Word Support Model to Improve Chinese Input System

**Jia-Lin Tsai**
Tung Nan Institute of Technology, Department of Information Management
Taipei 222, Taiwan
tsaijl@mail.tnit.edu.tw

## Abstract

This paper presents a word support model (WSM). The WSM can effectively perform homophone selection and syllable-word segmentation to improve Chinese input systems. The experimental results show that: (1) the WSM is able to achieve tonal (syllables input with four tones) and toneless (syllables input without four tones) syllable-to-word (STW) accuracies of 99% and 92%, respectively, among the converted words; and (2) while applying the WSM as an adaptation processing, together with the Microsoft Input Method Editor 2003 (MSIME) and an optimized bigram model, the average tonal and toneless STW improvements are 37% and 35%, respectively.

## 1 Introduction

According to (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; Tsai, 2005), the approaches of Chinese input methods (i.e. Chinese input systems) can be classified into two types: (1) ***keyboard based approach***: including 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); and (2) ***non-keyboard based approach***: including optical character recognition (OCR) (Chung, 1993), online handwriting (Lee et al., 1997) and speech recognition (Fu et al., 1996; Chen et al., 2000). Currently, the most popular Chinese input system is phonetic and pinyin based approach, because Chinese people are taught to write phonetic and pinyin syllables of each Chinese character in primary school.

In Chinese, 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)." The corresponding phonetic and pinyin syllables of each Chinese word is called syllable-words, such as "dai4 shu3" is the pinyin syllable-word of "袋鼠(kangaroo)." According to our computation, the {minimum, maximum, average} words per each distinct mono-syllable-word and poly-syllable-word (including bi-syllable-word and multi-syllable-word) in the CKIP dictionary (Chinese Knowledge Information Processing Group, 1995) are {1, 28, 2.8} and {1, 7, 1.1}, respectively. The CKIP dictionary is one of most commonly-used Chinese dictionaries in the research field of Chinese natural language processing (NLP). Since the size of problem space for syllable-to-word (STW) conversion is much less than that of syllable-to-character (STC) conversion, the most pinyin-based Chinese input systems (Hsu, 1994; Hsu et al., 1999; Tsai and Hsu, 2002; Gao et al., 2002; Microsoft Research Center in Beijing; Tsai, 2005) are addressed on STW conversion. On the other hand, 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).

As per (Chung, 1993; Fong and Chung, 1994; Tsai and Hsu, 2002; Gao et al., 2002; Lee, 2003; Tsai, 2005), ***homophone selection*** and ***syllable-word segmentation*** are two critical problems in developing a Chinese input system. Incorrect homophone selection and syllable-word seg-

*Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 842–849,
Sydney, July 2006. ©2006 Association for Computational Linguistics

mentation will directly influence the STW conversion accuracy. Conventionally, there are two approaches to resolve the two critical problems: (1) ***linguistic approach***: based on syntax parsing, semantic template matching and contextual information (Hsu, 1994; Fu et al., 1996; Hsu et al., 1999; Kuo, 1995; Tsai and Hsu, 2002); and (2) ***statistical approach***: based on the n-gram models where n is usually 2, i.e. bigram model (Lin and Tsai, 1987; Gu et al., 1991; Fu et al., 1996; Ho et al., 1997; Sproat, 1990; Gao et al., 2002; Lee 2003). From the studies (Hsu 1994; Tsai and Hsu, 2002; Gao et al., 2002; Kee, 2003; Tsai, 2005), the linguistic approach requires considerable effort in designing effective syntax rules, semantic templates or contextual information, thus, it is more user-friendly than the statistical approach on understanding why such a system makes a mistake. The statistical language model (SLM) used in the statistical approach requires less effort and has been widely adopted in commercial Chinese input systems.

In our previous work (Tsai, 2005), a word-pair (WP) identifier was proposed and shown a simple and effective way to improve Chinese input systems by providing tonal and toneless STW accuracies of 98.5% and 90.7% on the identified poly-syllabic words, respectively. In (Tsai, 2005), we have shown that the WP identifier can be used to reduce the over weighting and corpus sparseness problems of bigram models and achieve better STW accuracy to improve Chinese input systems. As per our computation, poly-syllabic words cover about 70% characters of Chinese sentences. Since the identified character ratio of the WP identifier (Tsai, 2005) is about 55%, there are still about 15% improving room left.

The objective of this study is to illustrate a word support model (WSM) that is able to improve our WP-identifier by achieving better identified character ratio and STW accuracy on the identified poly-syllabic words with the same word-pair database. We conduct STW experiments to show the tonal and toneless STW accuracies of a commercial input product (Microsoft Input Method Editor 2003, MSIME), and an optimized bigram model, BiGram (Tsai, 2005), can both be improved by our WSM and achieve better STW improvements than that of these systems with the WP identifier.

The remainder of this paper is arranged as follows. In Section 2, we present an auto word-pair (AUTO-WP) generation used to generate the WP database. Then, we develop a word support model with the WP database to perform STW conversion on identifying words from the Chinese syllables. In Section 3, we report and analyze our STW experimental results. Finally, in Section 4, we give our conclusions and suggest some future research directions.

## 2 Development of Word Support Model

The system dictionary of our WSM is comprised of 82,531 Chinese words taken from the CKIP dictionary and 15,946 unknown words auto-found in the UDN2001 corpus by a Chinese Word Auto-Confirmation (CWAC) system (Tsai et al., 2003). The UDN2001 corpus is a collection of 4,539624 Chinese sentences extracted from whole 2001 UDN (United Daily News, 2001) Website in Taiwan (Tsai and Hsu, 2002). 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 “ㄩˊ”-to-“ju2.”

### 2.1 Auto-Generation of WP Database

Following (Tsai, 2005), the three steps of auto-generating word-pairs (AUTO-WP) for a given Chinese sentence are as below: (the details of AUTO-WP can be found in (Tsai, 2005))

***Step 1*. *Get forward and backward word segmentations***: Generate two types of word segmentations for a given Chinese sentence by forward maximum matching (FMM) and backward maximum matching (BMM) techniques (Chen et al., 1986; Tsai et al., 2004) with the system dictionary.

***Step 2*. *Get initial WP set***: Extract all the combinations of word-pairs from the FMM and the BMM segmentations of Step 1 to be the initial WP set.

***Step 3*. *Get finial WP set***: Select out the word-pairs comprised of two poly-syllabic words from the initial WP set into the final 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.

## 2.2 Word Support Model

The four steps of our WSM applied to identify words 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 poly-syllabic words for the input syllables to be the WP set of Step 3.

***Step 3***. Select out the word-pairs that match a word-pair in the WP database to be the WP set. Then, compute the **word support degree** (**WS degree**) for each distinct word of the WP set. The WS degree is defined to be the total number of the word found in the WP set. Finally, arrange the words and their corresponding WS degrees into the WSM set. If the number of words with the same syllable-word and WS degree is greater than one, one of them is randomly selected into the WSM set.

***Step 4***. Replace words of the WSM set in descending order of WS degree with the input syllables into a WSM-sentence. If no words can be identified in the input syllables, a NULL WSM-sentence is produced.

Table 1 is a step by step example to show the four steps of applying our WSM on the Chinese syllables “sui1 ran2 fu3 shi2 jin4 shi4 sui4 yue4 xi1 xu1(雖然俯拾盡是歲月唏噓).” For this input syllables, we have a WSM-sentence “雖然俯拾盡是歲月唏噓.” For the same syllables, outputs of the MSIME, the BiGram and the WP identifier are “雖然腐蝕進士歲月唏噓,” “雖然俯拾盡是歲月唏噓” and “雖然 fu3 shi2 近視 sui4 yue4 xi1 xu1.”

# 3 STW Experiments

To evaluate the STW performance of our WSM, 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) (i.e. STW error reduction rate) = (accuracy of STW system with WP – accuracy of STW system)) / (1 – accuracy of STW system). (3)

| Step # | Results |
|---|---|
| Step.1 | sui1 ran2 fu3 shi2 jin4 shi4 sui4 yue4 xi1 xu1<br>(雖 然 俯 拾 盡 是 歲 月 唏 噓) |
| Step.2 | WP set (word-pair / word-pair frequency) =<br>{雖然-近視/6 (key WP for WP identifier),<br>俯拾-盡是/4, 雖然-歲月/4, 雖然-盡是/3,<br>俯拾-唏噓/2, 雖然-俯拾/2, 俯拾-歲月/2,<br>盡是-唏噓/2, 盡是-歲月/2, 雖然-唏噓/2,<br>歲月-唏噓/2} |
| Step.3 | WSM set (word / WS degree) =<br>{雖然/5, 俯拾/4, 盡是/4, 歲月/4, 唏噓/4,<br>近視/1}<br>Replaced word set =<br>雖然(sui1 ran2), 俯拾(fu3 shi2),<br>盡是(jin4 shi4), 歲月(sui4 yue4),<br>唏噓(xi1 xu1) |
| Step.4 | WSM-sentence:<br>雖然俯拾盡是歲月唏噓 |

**Table 1.** An illustration of a WSM-sentence for the Chinese syllables “sui1 ran2 fu3 shi2 jin4 shi4 sui4 yue4 xi1 xu1(雖然俯拾盡是歲月唏噓).”

## 3.1 Background

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. All the error PTC translations of GOING PTC were corrected by post human-editing. Then, apply our WSM to convert the testing input syllables back to their WSM-sentences. Finally, calculate its STW accuracy and ICR 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 system/user WP database used in the following STW experiments are described as below:

**(1) Training corpus**: We used the UDN2001 corpus as our training corpus, which is a collection of 4,539624 Chinese sentences extracted from whole 2001 UDN (United Daily News, 2001) Website in Taiwan (Tsai and Hsu, 2002).

**(2) Testing corpus**: The Academia Sinica Balanced (AS) corpus (Chinese Knowledge Information Processing Group, 1996) was selected as our testing corpus. The AS corpus is one of most famous traditional Chinese corpus used in the Chinese NLP research field (Thomas, 2005).

**(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 are {4, 37, and 12}.

**(4) Open test set**: 10,000 sentences were randomly selected from the AS 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 are {4, 40, and 11}.

**(5) System WP database**: By applying the AUTO-WP on the UDN2001 corpus, we created 25,439,679 word-pairs to be the system WP database.

**(6) User WP database**: By applying our AUTO-WP on the AS corpus, we created 1,765,728 word-pairs to be the user WP database.

We conducted the STW experiment in a progressive manner. The results and analysis of the experiments are described in Subsections 3.2 and 3.3.

## 3.2 STW Experiment Results of the WSM

The purpose of this experiment is to demonstrate the tonal and toneless STW accuracies among the identified words by using the WSM with the system WP database. The comparative system is the WP identifier (Tsai, 2005). Table 2 is the experimental results. The WP database and system dictionary of the WP identifier is same with that of the WSM.

From Table 2, it shows the average tonal and toneless STW accuracies and ICRs of the WSM are all greater than that of the WP identifier. These results indicate that the WSM is a better way than the WP identifier to identify polysyllabic words for the Chinese syllables.

| | Closed | Open | Average (ICR) |
|---|---|---|---|
| Tonal (WP) | 99.1% | 97.7% | 98.5% (57.8%) |
| Tonal (WSM) | 99.3% | 97.9% | 98.7% (71.3%) |
| Toneless (WP) | 94.0% | 87.5% | 91.3% (54.6%) |
| Toneless (WSM) | 94.4% | 88.1% | 91.6% (71.0%) |

**Table 2**. The comparative results of tonal and toneless STW experiments for the WP identifier and the WSM.

## 3.3 STW Experiment Results of Chinese Input Systems with the WSM

We selected Microsoft Input Method Editor 2003 for Traditional Chinese (MSIME) as our experimental commercial Chinese input system. In addition, following (Tsai, 2005), 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 system dictionary of the BiGram is same with that of the WP identifier and the WSM.

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

| | Ms | Ms+WP (I)[a] | Ms+WSM (I)[b] |
|---|---|---|---|
| Tonal | 94.5% | 95.5% (18.9%) | 95.9% (25.6%) |
| Toneless | 85.9% | 87.4% (10.1%) | 88.3% (16.6%) |

[a] STW accuracies and improvements of the words identified by the MSIME (Ms) with the WP identifier

[b] STW accuracies and improvements of the words identified by the MSIME (Ms) with the WSM

**Table 3a**. The results of tonal and toneless STW experiments for the MSIME, the MSIME with the WP identifier and with the WSM.

| | Bi | Bi+WP (I)[a] | Bi+WSM (I)[b] |
|---|---|---|---|
| Tonal | 96.0% | 96.4% (8.6%) | 96.7% (17.1%) |
| Toneless | 83.9% | 85.8% (11.9%) | 87.5% (22.0%) |

[a] STW accuracies and improvements of the words identified by the BiGram (Bi) with the WP identifier

[b] STW accuracies and improvements of the words identified by the BiGram (Bi) with the WSM

**Table 3b**. The results of tonal and toneless STW experiments for the BiGram, the BiGram with the WP identifier and with the WSM.

From Table 3a, the tonal and toneless STW improvements of the MSIME by using the WP identifier and the WSM are (18.9%, 10.1%) and (25.6%, 16.6%), respectively. From Table 3b, the tonal and toneless STW improvements of the BiGram by using the WP identifier and the WSM are (8.6%, 11.9%) and (17.1%, 22.0%), respectively. (Note that, as per (Tsai, 2005), the differences between the tonal and toneless STW accuracies of the BiGram and the TriGram are less than 0.3%).

Table 3c is the results of the MSIME and the BiGram by using the WSM as an adaptation processing with both system and user WP database. From Table 3c, we get the average tonal and toneless STW improvements of the MSIME and the BiGram by using the WSM as an adaptation processing are 37.2% and 34.6%, respectively.

| | Ms+WSM (ICR, I)[a] | Bi+WSM (ICR, I)[b] |
|---|---|---|
| Tonal | 96.8% (71.4%, 41.7%) | 97.3% (71.4%, 32.6%) |
| Toneless | 90.6% (74.6%, 33.2%) | 97.3% (74.9%, 36.0%) |

[a] STW accuracies, ICRs and improvements of the words identified by the MSIME (Ms) with the WSM

[b] STW accuracies, ICRs and improvements of the words identified by the BiGram (Bi) with the WSM

**Table 3c.** The results of tonal and toneless STW experiments for the MSIME and the BiGram using the WSM as an adaptation processing.

To sum up the above experiment results, we conclude that the WSM can achieve a better STW accuracy than that of the MSIME, the BiGram and the WP identifier on the identified-words portion. (Appendix A presents two cases of STW results that were obtained from this study).

## 3.4 Error Analysis

We examine the Top 300 STW conversions in the *tonal* and *toneless* from the open testing results of the BiGram with the WP identifier and the WSM, respectively. As per our analysis, the STW errors are caused by three problems, they are:

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

(2) ***Inadequate Syllable-Word Segmentation (ISWS) problem***: When an error is caused by ambiguous syllable-word segmentation (including overlapping and combination ambiguities), we call it inadequate syllable-word segmentation problem.

(3) ***Homophone selection problem***: The remaining STW conversion error is homophone selection problem.

| Problem | Coverage | |
|---|---|---|
| | Tonal | Toneless |
| | WP, WSM | WP, WSM |
| UW | 3%, 4% | 3%, 4% |
| ISWS | 32%, 32% | 58%, 56% |
| HS | 65%, 64% | 39%, 40% |
| # of error characters | 170, 153 | 506, 454 |
| # of error characters of mono-syllabic words | 100, 94 | 159, 210 |
| # of error characters of poly-syllabic words | 70, 59 | 347, 244 |

**Table 4**. The analysis results of the STW errors from the Top 300 tonal and toneless STW conversions of the BiGram with the WP identifier and the WSM.

Table 4 is the analysis results of the three STW error types. From Table 4, we have three observations:

(1) ***The coverage of unknown word problem for tonal and toneless STW conversions is similar***. In most Chinese input systems, unknown word extraction is not specifically a STW problem, therefore, it is usually taken care of through online and offline manual editing processing (Hsu et al, 1999). The results of Table 4 show that the most STW errors should be caused by ISWS and HS

problems, not UW problem. This observation is similarly with that of our previous work (Tsai, 2005).

(2) ***The major problem of error conversions in tonal and toneless STW systems is different***. This observation is similarly with that of (Tsai, 2005). From Table 4, the major improving targets of tonal STW performance are the HS errors because more than 50% tonal STW errors caused by HS problem. On the other hand, since the ISWS errors cover more than 50% toneless STW errors, the major targets of improving toneless STW performance are the ISWS errors.

(3) ***The total number of error characters of the BiGram with the WSM in tonal and toneless STW conversions are both less than that of the BiGram with the WP identifier***. This observation should answer the question "Why the STW performance of Chinese input systems (MSIME and BiGram) with the WSM is better than that of these systems with the WP-identifier?"

To sum up the above three observations and all the STW experimental results, we conclude that the WSM is able to achieve better STW improvements than that of the WP identifier is because: (1) the identified character ratio of the WSM is 15% greater than that of the WP identifier with the same WP database and dictionary, and meantime (2) the WSM not only can maintain the ratio of the three STW error types but also can reduce the total number of error characters of converted words than that of the WP identifier.

## 4 Conclusions and Future Directions

In this paper, we present a word support model (WSM) to improve the WP identifier (Tsai, 2005) and support the Chinese Language Processing on the STW conversion problem. All of the WP data can be generated fully automatically by applying the AUTO-WP on the given corpus. We are encouraged by the fact that the WSM with WP knowledge is able to achieve state-of-the-art tonal and toneless STW accuracies of 99% and 92%, respectively, for the identified poly-syllabic words. The WSM can be easily integrated into existing Chinese input systems by identifying words as a post processing. Our experimental results show that, by applying the WSM as an adaptation processing together with the MSIME (a trigram-like model) and the BiGram (an optimized bigram model), the average tonal and toneless STW improvements of the two Chinese input systems are 37% and 35%, respectively.

Currently, our WSM with the mixed WP database comprised of UDN2001 and AS WP database is able to achieve more than 98% identified character ratios of poly-syllabic words in tonal and toneless STW conversions among the UDN2001 and the AS corpus. Although there is room for improvement, we believe it would not produce a noticeable effect as far as the STW accuracy of poly-syllabic words is concerned.

We will continue to improve our WSM to cover more characters of the UDN2001 and the AS corpus by those word-pairs comprised of at least one mono-syllabic word, such as "我們(we)-是(are)". In other directions, we will extend it to other Chinese NLP research topics, especially word segmentation, main verb identification and Subject-Verb-Object (SVO) autoconstruction.

## 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 Charact*ers, M. Phil. thesis, Department of Computer Science, Hong Kong

University of Science and Technology.

Chinese Knowledge Information Processing Group. 1995. *Technical Report no. 95-02, the content and illustration of Sinica corpus of Academia Sinica*. Institute of Information Science, Academia Sinica.

Chinese Knowledge Information Processing Group. 1996. *A study of Chinese Word Boundaries and Segmentation Standard for Information processing* (in Chinese). Technical Report, Taiwan, Taipei, Academia Sinica.

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-to-character 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 Process*sing, 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 Language*s, 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/beijing/"

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.

Thomas E. 2005. The Second International Chinese Word Segmentation Bakeoff, In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, Oct. Jeju, Koera, 123-133.

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 Language Processing*, 9(1):41-64.

Tsai, J.L. 2005. Using Word-Pair Identifier to Improve Chinese Input System, *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing, IJCNLP2005*, 9-16.

United Daily News. 2001. On-Line United Daily News, http://udnnews.com/NEWS/

## Appendix A. Two cases of the STW results used in this study.

### Case I.

(a) Tonal STW results for the Chinese tonal syllables "guan1 yu2 liang4 xing2 suo3 sheng1 zhi1 shi4 shi2" of the Chinese sentence "關於量刑所生之事實"

| Methods | STW results |
|---|---|
| WP set | 關於-知識/4 (key WP),<br>關於-量刑/3, 量刑-事實/1,<br>關於-事實/1 |
| WSM Set | 關於(guan1 yu2)/3, 量刑(liang4 xing2)/2,<br>事實(shi4 shi2)/2, 知識(zhi1 shi4)/1 |
| WP-sentence | 關於 liang4 xing2 suo3 sheng1 知識 shi2 |
| WSM-sentence | 關於量刑 suo3 sheng1 zhi1 事實 |
| MSIME | 關於量行所生之事實 |
| MSIME+WP | **關於**量行所生**知識**實 |
| MSIME+WSM | **關於量刑**所生之**事實** |
| BiGram | 關於量刑所生之事時 |
| BiGram+WP | **關於**量刑所生**知識**時 |
| BiGram+WSM | **關於量刑**所生之**事實** |

(b) Toneless STW results for the Chinese toneless syllables "guan yu liang xing suo sheng zhi shi shi" of the Chinese sentence "關於量刑所生之事實"

| Methods | STW results |
|---|---|
| WP set | 關於/實施/4 (key WP),<br>關於/知識/4, 關於/量刑/3,<br>兩性/知識/2, 兩性/實施/2,<br>關於/失事/2, 量刑/事實/1,<br>關於/兩性/1, 關與/實施/1,<br>生殖/實施/1, 關於/事實/1,<br>關於/史實/1 |
| WSM Set | 關於(guan yu)/7, 實施(shi shi)/4,<br>兩性(liang xing)/3, 量刑(liang xing)/2,<br>知識(zhi shi)/2, 事實(shi shi)/2,<br>失事(shi shi)/1, 關與(guan yu)/1,<br>生殖(shengzhi)/1 |
| WP-sentence | 關於 liang xing suo sheng zhi 實施 |
| WSM-sentence | 關於兩性 suo 生殖實施 |
| MSIME | 關於兩性所生之事實 |
| MSIME+WP | **關於**兩性所生之**實施** |
| MSIME+WSM | **關於兩性**所**生殖實施** |
| BiGram | 貫譽良興所升値施事 |
| BiGram+WP | **關於**良興所升値**實施** |
| BiGram+WSM | **關於兩性**所**生殖實施** |

### Case II.

(a) Tonal STW results for the Chinese tonal syllables "you2 yu2 xian3 he4 de5 jia1 shi4" of the Chinese sentence "由於顯赫的家世"

| Methods | STW results |
|---|---|
| WP set | 由於/家事/6 (key WP),<br>顯赫/家世/2, 由於/家世/2<br>由於/家飾/1, 由於/顯赫/1 |
| WSM set | 由於(you2 yu2)/4, 顯赫(xian 3he4)/2,<br>家世(jia1 shi4)/2, 家事(jia1 shi4)/1 |
| WP-sentence | 由於 xian2 he4 de5 家事 |
| WSM-sentence | 由於顯赫 de 家世 |
| MSIME | 由於顯赫的家事 |
| MSIME+WP | **由於**顯赫的**家事** |
| MSIME+SWM | **由於顯赫**的**家世** |
| BiGram | 由於顯赫的家事 |
| BiGram+WP | **由於**顯赫的**家事** |
| BiGram+SWM | **由於顯赫**的**家世** |

(b) Toneless STW results for the Chinese toneless syllables "you yu xian he de jia shi" of the Chinese sentence "由於顯赫的家世"

| Methods | STW results |
|---|---|
| WP set | 由於-駕駛/14 (key WP),<br>由於-假釋/6, 由於-家事/6<br>顯赫/家世/2, 由於/家世/2<br>由於/家飾/1, 由於/顯赫/1 |
| WSM set | 由於(you yu)/6, 顯赫(xian he)/2,<br>家世(jia shi)/2, 駕駛(jia shi)/1 |
| WP-sentence | 由於 xian he de 駕駛 |
| WSM-sentence | 由於顯赫 de 家世 |
| MSIME | 由於顯赫的架勢 |
| MSIME+WP | **由於**顯赫的**駕駛** |
| MSIME+SWM | **由於顯赫**的**家世** |
| BiGram | 由於現喝的假實 |
| BiGram+WP | **由於**現喝的**駕駛** |
| BiGram+SWM | **由於顯赫**的**家世** |