Applying a Mix Word-Pair Identifier to the Chinese Syllable-to-Word Conversion Problem

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 describes a mix word-pair mix-WP) identifier to resolve homonym/segmentation ambiguities as well as perform STW conversion effectively for Chinese input. The mix-WP identifier includes a specific word-pair (SWP) identifier and a common word-pair (CWP) identifier. It is designed as a supporting processing with Chinese input systems. Our experiments show that by applying the mix-WP identifier, together with the Microsoft input method editor 2003 (MSIME) and an optimized bigram model (BiGram), the tonal and toneless STW performance of the two input systems can be improved.

1 Introduction

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. 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 existing Chinese input systems (Hsu 1994, Hsu *et al.* 1999, Tsai and Hsu 2002, Gao *et al.* 2002, MSIME) are addressed on STW conversion.

Conventionally, there are two approaches for STW conversion: (1) the **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) 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 (Gao *et al.* 2002, Lee 2003).

According to (Fong and Chung 1994, Tsai and Hsu 2002), ***homophone selectio**n* 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. The goal of this study is to illustrate the effectiveness of specific word-pairs and common word-pairs for resolving homonym/segmentation ambiguities to perform STW conversion in Chinese. In this paper, we use *tonal* to indicate the syllables with four tones, such as "ji4(技)shu4(術)" and *toneless* to indicate the syllables without four tones, such as "ji(技) shu(術)."

The remainder of this paper is arranged as follows. In Section 2, we firstly propose the method for auto-generating the specific word-pairs and the common word-pairs from given Chinese sentences. Then, we develop a mix word-pair (mix-WP) identifier includes a specific word-pair identifier and a common word-pair identifier. The mix-WP identifier is based on pre-collected datasets of specific and common word-pairs. 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 Mix-WP Identifier

In this study, a mix word-pair identifier includes a specific word-pair (SWP) identifier and a common word-pair (CWP) identifier. The system dictionary of the mix-WP identifier is comprised of the CKIP lexicon (CKIP, 1995) and those unknown words found automatically from the UDN 2001 corpus by a Chinese word auto-confirmation (CWAC) system (Tsai *et al*. 2003). The pinyin syllable-words were translated by phoneme-to-pinyin mappings, such as “ㄐㄧˋ”-to-“ji4.”

2.1 Development of SWP Identifier

2.1.1 Auto-Generate SWP Data.

The steps of auto-generating specific word-pair (AUTO-SWP) for a given Chinese sentence:

Step 1. Generate the segmentation for the given Chinese sentence with a backward maximum matching (BMM) technique. As pre (Tsai *et al*. 2004), the performance of BMM is better than that of forward maximum matching.

Step 2. Extract the BEGIN, END and BOUND word-pairs from the BMM segmentation of ***Step 1*** by following processes, respectively:

(1) BEGIN word-pair. When the word number of segmentation is greater than 1, the first two words will be comprised as a BEGIN word-pair. For the segmentation “音樂會(concert)現場(locale)湧入(enter)許多(many)觀眾(audience members),” the “音樂會-現場” will be generated as a BEGIN word-pair.

(2) END word-pair. When the word number of segmentation is greater than 2, the last two words will be comprised as an END word-pair. For the segmentation “全部(whole)工程(construction)預定(prearrange)年底(end of year)完成(complete),” the “年底-完成” will be generated as an END word-pair.

(3) BOUND word-pair. When the word number of segmentation is greater than 2, the first word and the last word will be comprised as a BOUND word-pair. For the segmentation “物價(price)大抵(ordinarily)維持(maintain)平穩(stable),” the “物價-平穩” will be generated as a BOUND word-pair.

Step 3. If the generated SWP was not found in its corresponding datasets, insert the generated SWP into the BEGIN, END and BOUND word-pair datasets, respectively.

2.1.2 SWP Identifier.

In Figure 1, the SWP data is a collection of auto-generated BEGIN, END and BOUND SWP datasets. If a SWP identifier only uses one of the BEGIN, END or BOUND SWP dataset, it will naturally become a BEGIN(BN), END(ED) or BOUND(BD) SWP identifier. The algorithm of our SWP identifier is as follows:

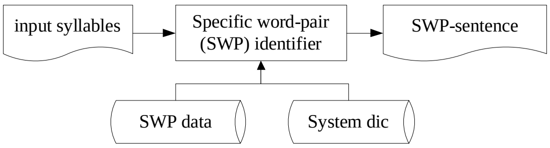

Fig. 1. *A system overview of the SWP identifier*

Step 1. Input tonal or toneless syllables.

Step 2. Generate all found BN, ED and BD SWP in the input syllables to be the initial SWP set. if the initial SWP set, if the found SWP number of the word-syllable pair of a BN, ED or BD SWP is greater than one in the BN, ED or BD datasets, respectively, the SWP will be dropped from the initial SWP set.

Step 3. Use the longest syllabic word-pair first (LS-WPF) strategy (Tsai and Hsu. 2002) to select the BN, ED and BD word-pair from the initial SWP set into the final SWP set.

Step 4. Replace corresponding syllable-word pair of the input syllables with the word-pairs of the final SWP set to be a SWP-sentence. As per our experiment, the performance of the three SWP identifiers is BD < BN < ED. Thus, the identifying sequence of our SWP identifier is from BD, BN to ED.

Table 1 is a step by step example that illustrates the four steps of our SWP identifier for the Chinese syllables “shu3 dou1 shu3 bu4 qing1 (數[count]都[always]數不清[innumerable]).” Note that when we used the Microsoft Input Method Editor 2003 for Traditional Chinese, a trigram-like input system (MSIME), to convert the same syllables, the output was “屬(belong)都(always)鼠(mouse)不(not)清(clear).”

Table 1. An illustration of a SWP-sentence for the Chinese syllables “shu3 dou1 shu3 bu4 qing1(數[count]都[always]數不清[innumerable])”

Step #	Results
Step.1	shu3 dou1 shu3 bu4 qing1 (數 都 數 不 清)
Step.2	The specific word-pairs found: 數(shu3)-都(dou1)/BEGIN pair 都(dou1)-數不清(shu3)/END pair
Step.3	The selected specific word-pairs: 數(shu3)-都(dou1)/BEGIN pair 都(dou1)-數不清(shu3)/END pair
Step.4	SWP-sentence: 數 都 shu3 bu4 qing1 (“shu3 dou1” replace with the BEGIN pair of Step 3) 數 都 數 不 清 (“shu3 bu4 qing1” replace with the END pair of Step 3)

2.2 Development of CWP Identifier

2.2.1 Auto-Generate CWP Data

The steps of auto-generating common word-pair (AUTO-CWP) for a given Chinese sentence:

Step 1. Generate the word segmentation for the given Chinese sentence by BMM technique.

Step 2. Extract all the combinations of word-pairs from the BMM segmentation of ***Step 1*** to be the initial CWP set. For the segmentation “我/不會/開車,” three CWP will be extracted, i.e. “我-不會”, “我-開車” and “不會-開車.”

Step 3. Select the word-pairs comprised of two multi-syllabic Chinese words (such as “不會(can not)”) to be the finial CWP set. For the final CWP set, if the word-pair is not found in the CWP database, insert it into the CWP database and set its frequency to 1; otherwise, increase its frequency by 1. In the above case, the final CWP set includes one word-pair, i.e. “不會-開車.”

2.2.2 CWP Identifier

The system overview of the CWP identifier is same with that of the SWP identifier as shown in Fig. 1. The algorithm of our CWP identifier 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 CWP database to be the initial CWP set, firstly. Then, from the initial CWP set, select the word-pair with maximum frequency as the key word-pair. Finally, find the co-occurrence word-pairs with the key word-pair in the training corpus to be the final CWP 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 CWP set into a CWP-sentence. If no word-pairs can be identified in the input syllables, a NULL CWP-sentence is produced.

If applying the CWP identifier on the syllables “yi1 ge5 wen2 ming2 de5 shuai1 wei2 guo4 cheng2(一個[a]文明[civilization]的[of]衰微[decay]過程[process]),” the generated WP-sentence will be “一個文明 de5shuai1wei2 過程.” For the same syllables, the MSIME will convert them into “一個[a]聞名[famous]的[of]衰微[decay]過程[process].” The detailed analysis and demonstration of our CWP identifier can be found in (Tsai 2005). Appendix A presents a case of the CWP identified results.

3 The STW Experiments

To evaluate the STW performance of our mix-WP identifier, the STW accuracy, the identified character ratio (ICR) and the STW improvement were used (Tsai 2005).

3.1 Experimental Data

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 were corrected by post human-editing. We, then, apply our SWP, CWP and mix-WP identifier to convert the syllable sequence back to words and calculate its STW accuracy and identified character ratio. All test sentences are composed of a string of Chinese characters.

In following experiments, the training and testing corpus, closed/open test sets and the collection of the testing SWP and CWP data were:

Training corpus: The UDN 2001 corpus was selected as our training corpus. It is a collection of 4,539,624 Chinese sentences extracted from whole 2001 articles on the United Daily News Website (UDN) in Taiwan.

Testing corpus: The UDN 2002 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 (UDN).

Closed testing set: 10,000 sentences were randomly selected from the UDN 2001 corpus as the closed testing set.

Open testing set: 10,000 sentences were randomly selected from the UDN 2002 corpus as the open testing set. At this point, we checked that the selected open testing sentences were not in the closed testing set as well.

Testing SWP data: By applying our AUTO-SWP on the UDN 2001 corpus, we created 1,754,055 BN, 1,594,036 ED and 2,502,241 BD specific word-pairs.

Testing CWP data: By applying our AUTO-CWP on the UDN 2001 corpus, we created 25,439,679 common word-pairs.

In this study, we conducted the STW experiment in a progressive manner. The experimental results of the SWP, CWP and mix-WP identifiers are described in Sub-sections 3.2, 3.3.and 3.4, respectively.

3.2 Experiment of SWP Identifier

This experiment is to demonstrate the tonal and toneless STW accuracies by using the SWP identifier with the testing BN, ED, BD and ALL datasets, respectively. Note that the symbol ALL stands for a mixed collection of all BN, ED and BD word-pairs generated from the UDN 2001 corpus.

Table 2. The results of tonal/toneless STW experiments for the SWP identifier with BN, ED, BD and ALL specific word-pairs

Data	Closed	Open	Average (ICR)
BN	99.7 / 97.7	99.1 / 96.1	99.4(11.6)/96.7(9.2)
ED	99.9 / 99.6	99.3 / 97.3	99.6(14.3)/98.4(12.1)
BD	99.6 / 98.0	99.2 / 95.9	99.3(17.7)/96.3(13.4)
ALL [a]	99.7 / 98.3	99.2 / 96.3	99.4(30.7)/97.1(22.8)

[a] The performance of SWP identifier with three SWP data and the word-pair replacing sequence of the SWP is from BD, BN to ED

Table 2 shows the average tonal and toneless STW accuracies of the SWP identifier with ALL SWP data for the closed and open test sets are 99.4% and 97.1%, respectively. Meanwhile, between the closed and open test sets, the differences of tonal and toneless STW accuracies of the SWP identifier are 0.5% and 2%, respectively.

3.3 Experiment of CWP Identifier

This experiment is to demonstrate the tonal and toneless STW accuracies among the identified word-pairs by using the CWP identifier with the testing CWP data.

Table 3. The results of the tonal and toneless STW experiment for the CWP identifier

	Closed	Open	Average (ICR)
Tonal	99.1	98.4	98.8 (61.9)
Toneless	94.1	90.9	92.6 (58.6)

Table 3 shows the average tonal and toneless STW accuracies of the CWP identifier for closed and open test sets are 98.8% and 92.6%, respectively. Meanwhile, between the closed and open test sets, the differences of tonal and toneless STW accuracies of the CWP identifier are 0.7% and 3.2%, respectively.

3.4 Experiment of Mix-WP Identifier

This experiment is to demonstrate the tonal and toneless STW accuracies among the identified word-pairs by using the mix-WP identifier with all testing WP data. From Tables 2 and 3, the STW performance of the SWP identifier is better than that of the CWP identifier. Therefore, our mix-WP identifier uses the CWP identifier to identify CWP first and the SWP identifier to identifier SWP last for a given syllables.

Table 4. The results of tonal and toneless STW experiments for the mix-WP identifier

	Closed	Open	Average (ICR)
Tonal	99.2	98.4	98.8 (67.6)
Toneless	94.9	91.8	93.5 (64.6)

Table 4 shows the average tonal and toneless STW accuracies of the mix-WP identifier for closed and open test sets are 98.8% and 93.5%, respectively. Meanwhile, between the closed and open test sets, the differences of tonal/toneless STW accuracies of the mix-WP identifier are 0.8% and 3.1%, respectively. The average identified character ratio (ICR) of the tonal and the toneless syllables are 67.6% and 64.6%, respectively. To sum up the results of Tables 2 to 4, we conclude that the mix-WP (SWP and CWP) data can be used to effectively convert Chinese STW on the mix-WP-related portion (including the SWP-related portion and the CWP-related portion, respectively).

3.5 Commercial IME System and 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 (Tsai *et al.* 2004). 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 LS-WPF strategies (Chen *et al.* 1986, Tsai *et al.* 2004). The training corpus and the system dictionary of this BiGram system are same with that of the mix-WP identifier. In this experiment, the STW output of the MSIME with the mix-WP identifier, or the BiGram with the mix-WP identifier, was collected by directly replacing the identified word-pairs from the corresponding STW output of MSIME or BiGram.

Table 5. The results of tonal and toneless STW experiment for the MSIME and the MSIME with the mix-WP identifier

	MSIME	MSIME+WP [a]	Improvement
Tonal	94.7%	96.3%	29.3%
Toneless	86.4%	89.4%	22.5%

[a] STW accuracies of the words identified by the MSIME with the mix-WP identifier

Table 6. The results of the tonal and toneless STW experiment for the BiGram and the BiGram with the mix-WP identifier

	BiGram	BiGram+WP [a]	Improvement
Tonal	96.4%	96.9%	12.8%
Toneless	85.2%	88.1%	19.6%

[a] STW accuracies of the words identified by the BiGram with the mix-WP identifier

From Table 5, the tonal and toneless STW improvements of the MSIME by using the mix-WP identifier are 29.2% and 22.5%, respectively. On the other hand, from Table 6, the tonal and toneless STW improvements of the BiGram by using the mix-WP identifier are 12.8% and 19.6%, respectively. To sum up the results of this experiment, we conclude that the mix-WP identifier can achieve better WP-portion STW accuracy than that of the MSIME and BiGram Chinese input systems.

4 Conclusion and Future Directions

In this paper, we have applied a mix-WP identifier to the Chinese STW conversion and obtained a high STW accuracy on the identified word-pairs with ICR of more than 60%. All of the testing mix-WP data was auto-generated by using the AUTO-SWP and the AUTO-CWP on the training corpus. We are encouraged by the fact that mix-WP knowledge can achieve tonal and toneless STW accuracies of 98.8% and 93.5%, respectively, for the mix-WP-related portion of the testing syllables. The mix-WP identifier can be easily integrated into existing Chinese input systems or Chinese language processing of typical speech recognition systems by identifying word-pairs in a post-processing step. Our experimental results show that, by applying the mix-WP identifier together with the MSIME and the BiGram input systems, the tonal and toneless STW improvements are 29%/23% and 13%/20%, respectively. To the adaptive approach, we also tried to use the AUTO-SWP and the AUTO-CWP to auto-extract new SWP and CWP from the open test sentences into the mix-WP data, firstly. Then, we found the overall tonal and toneless STW accuracies of the MSIME and the BiGram for closed/open syllables become 96.5%/90% and 97.1%/89%, respectively.

Currently, our approach is quite basic when more than one SWP or CWP 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 apply the mix-WP knowledge to speech recognition. According to our computations, the collection of our mix-WP knowledge can cover approximately 70% and 60% of the characters in the UDN 2001 and 2002 corpus, respectively.

We will continue to expand our collection of mix-WP knowledge with Web corpus. In other directions, we will try to improve our WP-based STW conversion with other types of WP data, such as NEVF and MWP (Tsai *et al.* 2002 and 2004), and statistical language models, such as HMM, and extend it to other areas of NLP, especially word segmentation and the mix-WP identifier from the word lattice of Chinese speech recognition systems.

References

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.

CKIP. Technical Report no. 95-02. 1995. The content and illustration of Sinica corpus of Academia Sinica. Institute of Information Science, 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.

GOING, "http://www.iqchina.com/"

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. 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 Y.S., Chen. 1999. On Phoneme-to-Character Conversion Systems in Chinese Processing, *Journal of Chinese Institute of Engineers*, 5:573-579.

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, 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.

Manning, C. D. and Schuetze, H. 1999. *Fundations of Statistical Natural Language Processing*, MIT Press: 191-220.

MSIME, Microsoft Research Center in Beijing, "http://research.microsoft.com/aboutmsr/labs/beijing"

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*.

Tsai, J.L. 2005. Using Word-Pair Identifier to Improve Chinese Input System, *Proceedings of 4th SIGHAN workshop on Chinese Language Processing*, Korea.

Tsai, J.L, G., Hsieh and W.L., Hsu. 2004. Auto-Generation of NVEF knowledge in Chinese, *Computational Linguistics and Chinese Language Processing*, 9(1):41-64.

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, C.L., T.J., Jiang and W.L., Hsu. 2004. Applying Meaningful Word-Pairs on Syllabel-to-Word Conversion Problem in Chinese, *Proceedings of ROCLING XVI*, Taiwan, 79-88.

Tsai, J,L, C.L., Sung and W.L., Hsu. 2003. Chinese Word Auto-Confirmation Agent, *Proceedings of ROCLING XV*, Taiwan, 175-192.

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

Appendix A.

Input syllables "ji2fu4qi2min2zu2te4se4" of the Chinese sentence "極富(abundance)其(it)民族(folk)特色(characteristic)"

Tonal STW results

Methods	STW results
WP-sentence	民族/特色(13) (Key WP) 極富/特色(11) (Co-occurrence WP) 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

Methods	STW results
WP-sentence	民族/特色(13) (Key WP) 極富/特色(11) (Co-occurrence WP) 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**