Applying Meaningful Word-Pair Identifier to the Chinese Syllable-to-Word Conversion Problem

Jia-Lin Tsai, Tien-Jien Chiang and Wen-Lian Hsu
Institute of Information Science, Academia Sinica,
Nankang, Taipei, Taiwan, R.O.C.
{tsaijl,tmjiang, hsu}@iis.sinica.edu.tw

Abstract

Syllable-to-word (STW) conversion is a frequently used Chinese input method that is fundamental to syllable/speech understanding. The two major problems with STW conversion are the segmentation of syllable input and the ambiguities caused by homonyms. This paper describes a meaningful word-pair (MWP) identifier that can be used to resolve homonym/segmentation ambiguities and perform STW conversion effectively for Chinese language texts. It is designed as a support system with Chinese input systems. In this paper, five types of meaningful word-pairs are investigated, namely: noun-verb (NV), noun-noun (NN), verb-verb (VV), adjective-noun (AN) and adverb-verb (DV). The pre-collected datasets of meaningful word-pairs are based on our previous work *auto-generation of NVEF knowledge in Chinese* [Tsai *et al*. 2003a and 2004], where NVEF stands for noun-verb event frame.

The main purpose of this study is to illustrate that a hybrid approach of combining statistical language modeling (SLM) with contextual information, such as meaningful word-pairs, is effective for syllable-to-word conversion and is important for syllable/speech understanding. Our experiments show the following: (1) the MWP identifier achieves ***tonal*** (syllables with four tones) and ***toneless*** (syllables without four tones) STW accuracies of 98.69% and 90.7%, respectively, among the identified word-pairs for the test syllables; (2) by STW error analysis, we find that the major critical problem of tonal STW systems is the failure of homonym disambiguation (52%), while that of toneless STW systems is inadequate syllable segmentation (48%); (3) by applying the MWP identifier, together with an optimized bigram model and the Microsoft input method editor (MSIME 2003), the tonal/toneless STW accuracies of the two STW systems can be improved from 96.27%/85.47% to 96.75%/87.74% and from 95.05%/86.94% to 96.30%/89.97%, respectively.

Keywords: syllable-to-word, contextual information, word pair, top-down identifier, bigram

1. Introduction

More than 100 Chinese input methods have been developed in the past [Becker *et al*. 1985, Huang 1985, Lin *et al.* 1987, Sproat 1990, Gu *et al.* 1991, Kuo 1995, Fu *et al.* 1996, Ho *et al.* 1997, Hsu *et al*. 1999, Tsai *et al*. 2002a , Gao *et al*. 2002, Lee 2003].

Their underlying approaches can be classified into four types:

(a) Optical character recognition (OCR) based [Chung *et al.* 1993],

(b) On-line handwriting based [Lee *et al.* 1997],

(c) Speech based [Fu *et al*. 1996, Chen *et al.* 2000], and

(d) Keyboard based, such as syllabic-input-to-character [Chang *et al*. 1991, Hsu *et al*. 1993, Hsu 1994, Hsu *et al*. 1999, Kuo 1995, Lua *et al*. 1992]; arbitrary codes based [Fan *et al.* 1988]; and structure scheme based [Huang 1985]. The major goal of these syllable input systems is to achieve high STW accuracy, but syllable understanding is rarely considered [Hsu *et al*. 1999].

Currently, the most popular method for Chinese input is syllable based (or phonetic/pinyin based), because Chinese people are taught to write the corresponding phonetic/pinyin syllable of each Chinese character in primary school. Basically, each Chinese character corresponds to at least one syllable. Although there are more than 13,000 distinct Chinese characters (of which 5,400 are commonly used), there are only 1,300 distinct syllables. The homonym (homophone) problem is, therefore, quite severe when using a Chinese phonetic input method [Chung *et al.* 1993]. 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. Therefore, *homonym disambiguation* is a critical problem that requires the development of an effective syllable-to-word (STW) conversion system for Chinese. A comparable problem for STW conversion in English is word-sense disambiguation (WSD).

There are two conventional approaches for STW conversion: 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 *et al*. 2002a]; and the *statistical approach* based on the *n*-gram model where *n* is usually 2 or 3 [Lin *et al.* 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 (i.e. it is easier to understand why such a system makes a mistake). On the other hand, the statistical language model (SLM) used in the statistical approach requires less effort and has been widely adopted in commercial systems. However, the power of the statistical approach depends on the training corpus because the SLM usually pays little attention to syllable understanding. Following the work of [Gu *et al.* 1991, Hsu 1994, Kuo 1995, Fu *et al*. 1996, Ho *et al.* 1997, Tsai *et al*. 2002a, Gao *et al.* 2002], a better

approach to STW conversion is to integrate both linguistic knowledge (such as contextual information) and statistical approaches (such as an n-gram model). We believe that our research proves the efficacy of such an integrated approach.

According to previous studies [Chung *et al.* 1993, Fong *et al.* 1994 Tsai *et al.* 2002a, Gao *et al.* 2002, Lee 2003], besides homonyms, correct *syllable-word segmentation* is another crucial problem of STW conversion. Incorrect syllable-word segmentation directly influences the conversion rate of STW. 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)." According to the CKIP lexicon [CKIP 1995], the two possible syllable-word segmentations are:

(F) "yi1/du4/ji4yu2/zhong1guo2/de5/niang4jiu3/ji4shu4"; and

(B) "yi1/du4ji4/yu2/zhong1guo2/de5/niang4jiu3/ji4shu4."

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

Among the above syllable-word segmentations, there is an ambiguous syllable-word section: /du4/ji4yu2/ (/{妒,杜,肚,度,渡,鍍,蠹}/{覬覦,鯽魚,計於,繼於}/); and /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}/), respectively. In this case, if the system has the contextual information that the pairs "技術(technique)-覬覦(covet)" and "一度(once)-覬覦(covet)" are, respectively, meaningful noun-verb (NV) and adverb-verb (DV) word-pairs, then the ambiguous syllable-word section can be effectively resolved and the word-pairs "技術(technique)-覬覦(covet)" and "一度(once)-覬覦(covet)" of this syllable sequence can be correctly identified.

For the above case, if we look at the Sinica corpus [CKIP 1995], the bigram frequencies of "覬覦(covet)-中國(China)" and "於(at)-中國(China)" are 0 and 24, respectively. Therefore, by using a bigram model trained with the Sinica corpus, the forward syllable-word segmentation would conclude that the following word segmentation /於/中國/, will be incorrect. In fact, if we use Microsoft Input Method Editor 2003 for Traditional Chinese (a tri-gram like STW product), the syllables of the above example will be converted to "一度(once)繼(continue)於(to)中國(China)的(of)釀酒(making-wine)技術(technique)." It is widely recognized that unseen event ("覬覦-中國") and over-weighting ("於-中國") are two major problems of SLM systems [Fu *et al.* 1996, Gao *et al.* 2002]. Practical SLM is either a bigram or a trigram model. As the above case shows, the meaningful word-pairs (or contextual information) "技術(technique)-覬覦(covet)" and "一度(once)-覬覦(covet)" can be used to overcome both the unseen event and over-weighting problems of SLM-based STW systems. In [Tsai *et al.* 2002a], we showed that the knowledge of noun-verb event frame (NVEF) sense-pairs and their corresponding NVEF word-pairs (NVEF knowledge) are useful for effectively resolving word sense ambiguity with an

accuracy of 93.7%). In [Tsai *et al*. 2002b], we showed that a NVEF word-pair identifier with pre-collected NVEF knowledge can be used to obtain a tonal (syllables with four tones) STW accuracy of more than 99% for the NVEF related portion in Chinese.

The objective of this study is to illustrate the effectivness of meaningful noun-verb (NV), noun-noun (NN), verb-verb (VV), adjective-noun (AN) and adverb-verb (DV) word-pairs for solving Chinese STW conversion problems. We conduct STW experiments to show that the ***tonal*** and ***toneless*** STW accuracies of conventional SLM models and the commercial input products can be improved by using a meaningful word-pair identifier without a tuning process. In this paper, we use ***tonal*** to indicate the syllables input with four tones, such as "niang4(釀) jiu3(酒) ji4(技)　shu4(術)," and ***toneless*** to indicate the syllables input without four tones, such as "niang(釀) jiu(酒) ji(技) shu(術)."

The remainder of this paper is arranged as follows. In Section 2, we propose the method for *auto-generating the meaningful word-pairs* in Chinese based on [Tsai *et al*. 2003a and 2004], and a *meaningful word-pair identifier* to resolve homonym/segmentation ambiguities of STW conversion in Chinese. The meaningful word-pair identifier is based on pre-collected datasets of meaningful word-pairs. In Section 3, we present our STW experiment results and analysis. Finally, in Section 4, we give our conclusions and suggest some future research directions.

2. Development of the Meaningful Word-Pair Identifier

To develop the meaningful word-pair (MWP) identifier, we selected Hownet [Dong 1999] as our system's dictionary because it provides knowledge of Chinese words, word senses and part-of-speeches (POS). The Hownet dictionary used in this study contains 58,541 Chinese words, among which there are 33,264 nouns, 16,723 verbs, 8,872 adjectives and 882 adverbs.

In this system's dictionary, the syllable-word for each word is obtained by using the inverse phoneme-to-character system presented in [Hsu 1994], while the word frequencies are computed according to a fixed-size *United Daily News (UDN)* 2001 corpus. The latter is a collection of 4,539,624 Chinese sentences extracted from articles on the *United Daily News* Website [On-Line United Daily News] from January 17, 2001 to December 30, 2001. Table 1 shows the statistics of the number of articles per article class in this *UDN* 2001 corpus.

Table 1. The number of articles per article class in the training corpus

article class	大陸 China	地方 Local	社會 Society	股市 Stock	政治 Politics	科技 Science	旅遊 Travel
# of articles	90	26,843	136	19,699	133	5,870	6,183
article class	消費 Consumption	財經 Financial	國際 World	運動 Sport	影視 Entertainment	醫藥 Health	藝文 Arts
# of articles	12498	23,563	7,404	12,404	18,674	5,653	9,989

2.1 Generating the Meaningful Word-Pair

In [Tsai *et al*. 2004], we propose an AUTO-NVEF system to auto-generate NVEF knowledge in Chinese. It extracts NVEF knowledge from Chinese sentences by four major processes: (1) segmentation checking; (2) Initial Part-of-Speech (IPOS) sequence generation; (3) NV knowledge generation; and (4) NVEF knowledge auto-confirmation. The details of the four processes can be found in [Tsai *et al*. 2004]. Take the Chinese sentence "音樂會(concert)/現場(locale)/湧入(enter)/許多(many)/觀眾(audience members)" as an example. For this sentence, AUTO-NVEF will generate two collections of NVEF knowledge: 現場(locale)-湧入(enter) and 觀眾(audience members)-湧入(enter). In [Tsai *et al*. 2004], we reported that AUTO-NVEF achieved 98.52% accuracy for news and 96.41% for specific text types, which included research reports, classical literature and modern literature. In addition, it automatically discovered over 400,000 NVEF word-pairs in the *UDN* 2001 corpus.

Using AUTO-NVEF as the base, we extended the system into a meaningful word-pair (MWP) generation called AUTO-MWP. The steps of AUTO-MWP are:

.

Step 1. Use AUTO-NVEF to generate NVEF word-pairs for the given Chinese sentence. AUTO-NVEF adopts a *forward=backward* maximum matching technique to perform word segmentation and a *bigram-like* model to perform POS tagging [Tsai *et al*. 2004]. If no NVEF word-pairs are generated, go to Step 3.

Step 2. According to the generated NVEF word-pairs and the word-segmented sentence with POS tagging from Step 1, the auto-generation methods of meaningful NN, VV, AN and DV word-pairs are:

(1) *Generation of NN word-pair*. When the number of generated NVEF word-pairs is greater than 1, this sub-process will be triggered. If the nouns of two generated NVEF word-pairs share the same verb, the two nouns will be designated as a meaningful NN word-pair. Take the generated NVEF word-pairs of 現場(locale)-湧入(enter) and 觀眾(audience members)-湧入(enter) for the sentence "音樂會(concert)現場(locale)湧入(enter)許多(many)觀眾(audience members)" as examples. The noun 現場(locale) and the noun 觀眾(audience members) are designated as one NN word-pair because the two nouns share the same verb 湧入(enter) in this sentence.

(2) *Generation of VV word-pair*. When the number of generated NVEF word-pairs is greater than 1, this sub-process will be triggered. If the verbs of two generated NVEF word-pairs share the same noun, the two verbs will be designated as a meaningful VV word-pair. Take the generated NVEF word-pairs 年底(the end of year)-預定(prearrange) and 年底(the end of

year)-完成(complete) for the sentence "全部(whole)工程(construction)預定(prearrange)年底(the end of year)完成(complete)" as examples. The verb 預定(prearrange) and the verb 完成(complete) are designated as one VV word-pair because the two verbs share the same noun 年底(the end of year).

(3) *Generation of AN word-pair*. For each noun of a generated NVEF word-pair, if the word immediately to its left is an adjective, the noun and the adjective are designated as one AN word-pair. Take the generated NVEF word-pair 觀眾(audience members)-湧入(enter) for the word-segmented and POS-tagged sentence "音樂會(N)現場(N)湧入(V)許多(ADJ)觀眾(N)" as an example. Since the word immediately to the left of 觀眾(audience members) is an adjective 許多(many), the adjective 許多(many) and the noun 觀眾(audience members) are designated as one AN word-pair.

(4) *Generation of DV word-pair*. For each verb of a generated NVEF word-pair, if the word immediately to its left is an adverb, the verb and the adverb are designated as one DV word-pair. Take the generated NVEF word-pair 物價(price)-維持(maintain) for the word-segmented and POS-tagged sentence "物價(N)大抵(ADV)維持(V)平穩(ADJ)" as an example. Since the word immediately to the left of 維持(maintain) is an adverb 大抵(ordinarily), the adverb 大抵(ordinarily) and the verb 維持(maintain) are designated as one DV word-pair.

Step 3. Stop.

Table 2. The number of generated NV, NN, VV, AN and DV word-pairs obtained by applying AUTO-MWP to the *UDN* 2001 corpus

NV	NN	VV	AN	DV	Total
430,698	533,780	220,022	138,055	111,879	1,434,434

Table 3. Fifteen randomly selected examples of meaningful NV, NN, VV, AN and DV word-pairs and their corresponding frequencies from the generated MWP datasets for the *UDN* 2001 corpus

NV	NN	VV	AN	DV
大學-附設/118	全國-比賽/83	開放-投資/541	對外-交通/206	即將-舉行/188
人選-決定/35	偶像-明星/103	接受-訪問/1483	資深-釣友/103	幾乎-都是/390
路線-是/96	市府-人員/107	擔心-造成/124	最高-榮譽/129	再度-成爲/144

Table 2 shows the number of generated NV, NN, VV, AN and DV word-pairs obtained by applying AUTO-MWP to the *UDN* 2001 corpus. The frequencies of all the generated meaningful word-pairs were computed by the *UDN* 2001 corpus. Note that the frequency of a meaning-

ful word-pair is the number of sentences that contain the word-pair *with the same word-pair order* in the *UDN* 2001 corpus. Table 3 shows fifteen randomly selected NV, NN, VV, AN and DV word-pairs and their corresponding frequencies in the generated MWP datasets for the *UDN* 2001 corpus.

2.2 Meaningful Word-Pair Identifier

We developed a NVEF word-pair identifier [Tsai *et al.* 2002a] for Chinese syllable-to-word (STW) and achieved a tonal STW accuracy of more than 99% on the NVEF related portion. This NVEF word-pair identifier is based on the techniques of *longest syllabic NVEF-word-pair first* (LS-NVWF), *exclusion-word-list* (EWL) checking and pre-collected NVEF knowledge. By modifying the algorithm of this identifier in [Tsai *et al*. 2002a], we obtain our meaningful word-pair (MWP) identifier, (Figure 1). As shown in the figure, if the MWP identifier only uses one of the meaningful NV, NN, VV, AN or DV word-pair datasets, it will naturally become an MNV, MNN, MVV, MAN or MDV word-pair identifier.

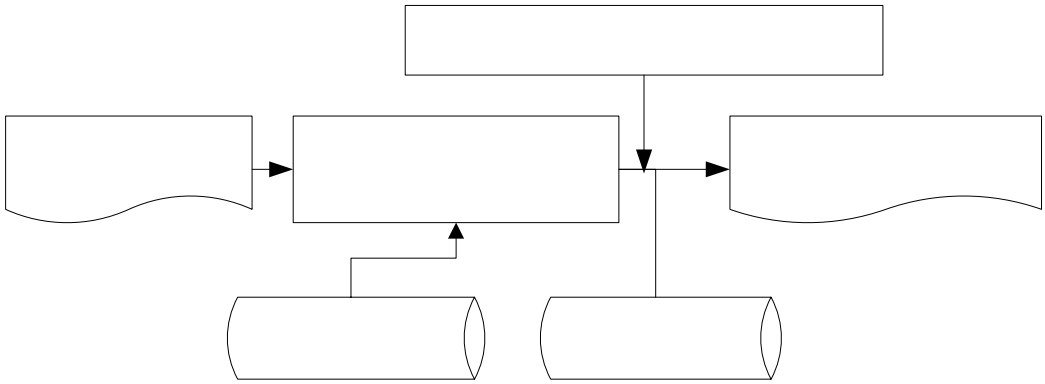

Figure 1. A system overview of the meaningful word-pair (MWP) identifier

The algorithm of the MWP identifier is as follows:

Step 1. Input tonal (with four tones) or toneless (without four tones) syllables.

Step 2. Generate all possible word-pairs found in the input syllables. Exclude certain word-pairs based on EWL checking. Appendix A lists all of the exclusion words used in this study. Note that our meaningful word-pairs include monosyllabic nouns/adjectives/adverbs and monosyllabic verbs, except "是(be)" and "有(has/have)" that are dropped in this Step.

Step 3. Word-pairs that match a meaningful word-pair in the generated MWP data are used as the initial MWP set for the input syllables.. From the initial MWP set, select a key word-pair and its co-occurring word-pairs to be the final MWP set. Conflicts are resolved using the *longest syllabic word-pair first* (LS-WPF) strategy. If there

are two or more word-pairs with the same condition, the system triggers the following processes.

(1) The word-pair with the greatest frequency (the number of sentences that contain the word-pair *with the same word-pair* order in the *UDN* 2001 corpus) is selected as the key word-pair. If there are two or more word-pairs with the same frequency, one of them is randomly selected as the key word-pair.

(2) The word-pairs that co-occur with the key word-pair in the *UDN* 2001 corpus are selected.

(3) The key and co-occurred word-pairs are then combined as the final MWP set for Step 4.

Step 4. Arrange all word-pairs of the final MWP set into a *MWP-sentence* as shown in Table 3. If no word-pairs can be identified from the input syllables, a null MWP-sentence is produced.

Table 3. An illustration of an MWP-sentence for 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])." (The English words in parentheses are included for explanatory purposes only.)

Process	Results	Pair freq.
Step.1	yi1 ge5 wen2 ming2 de5 shuai1 wei2 guo4 cheng2 (一 個 文 明 的 衰 微 過 程)	
Step.2	The meaningful word-pairs found: 文明(wen2 ming2)-過程(guo4 cheng2)/NN pair 文明(wen2 ming2)-衰微(shuai1 wei2)/NV pair	 3 1
Step.3	The key meaningful word-pair: 文明(wen2 ming2)-過程(guo4 cheng2)/NN pair The co-occurred word-pair: 文明(wen2 ming2)-衰微(shuai1 wei2)/NV pair	
Step.4	MWP-sentence: yi1 ge5 文 明 de5 衰 微 過 程	

Table 3 is a step by step example that illustrates the four processes of our MWP identifier for 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])." When we used MSIME 2003 to convert the same syllables, the output was "一個(one)聞名(famous)的(of)衰微(decay)過程(process)." Obviously, the over-weighted bigram "聞名-的(wen2 ming2-de5)" causes an STW error in MISIME 2003, which uses a statistical language model (SLM) with a trigram-like Chinese input product [Microsoft Research Center in Beijing]. If we use the MWP-sentence shown in Step 4 to directly replace the corresponding characters of the MSIME 2003 output in this example, the

error converted word “聞名(famous)”, caused by the over-weighting of MSIME 2003, becomes the correct word “文明(civilization).”

3. The STW experiment

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

STW accuracy =

# of correct characters / # of total characters. (1)

Identified character ratio (ICR) =

# of characters of identified MWPs / # of total characters in testing sentences. (2)

3.1 Closed Test Set and Open Test Set

We use the inverse translator of the phoneme-to-character system in [Hsu *et al.* 1994] to convert a test sentence into a syllable sequence. We then apply our MWP identifier to convert this syllable sequence back to characters and calculate its STW accuracy and identified character ratio by Equations (1) and (2). All test sentences are composed of a string of Chinese characters.

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

Training corpus: We used the *UDN* 2001 corpus mentioned in Section 2 as our training corpus. All knowledge of word frequencies, meaningful word-pairs, MWP frequencies was auto-generated and computed by this corpus.

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 articles on the *United Daily News* Website [On-Line United Daily News] from January 1, 2002 to December 30, 2002.

Closed test set: 10,000 sentences were randomly selected from the *UDN* 2001 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}.

Open test set: 10,000 sentences were randomly selected from the *UDN* 2002 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}.

Meaningful word-pair data: By applying our AUTO-MWP on the *UDN* 2001 corpus, we created 430,698 NV, 533,780 NN, 220,022 VV, 138,055 AN and 111,879 DV word-pairs as the

MWP testing data.

In this study, we conducted the STW experiment in a progressive manner. The results and analysis of the experiment are described in Sub-sections 3.2, 3.3 and 3.4. Appendix B presents three STW results that were obtained from the experiment.

3.2 STW Experiment for the MWP Identifier

The purpose of this experiment is to demonstrate the tonal and toneless STW accuracies by using the MWP identifier with NV, NN, VV, AN, DV and (NV+NN+VV+AN+DV) word-pairs.

Table 4a. The results of the tonal STW experiment for the MWP identifier with NV, NN, VV, AN, DV and (NV+NN+VV+AN+DV) word-pairs.

	Closed	Open	Average (identified character ratio)
NV	99.08%	98.70%	98.90% (21.69%)
NN	98.54%	98.30%	98.43% (34.56%)
VV	98.25%	97.255	97.81% (14.64%)
AN	97.41%	96.83%	97.14% (10.07%)
DV	98.07%	97.45%	97.80% (9.46%)
NV+NN+VV+AN+DV	98.69%	98.20%	98.46% (46.67%)

Table 4b. The results of the toneless STW experiment for the MWP identifier with NV, NN, VV, AN, DV and (NV+NN+VV+AN+DV) word-pairs.

	Closed	Open	Average (identified character ratio)
NV	91.53%	90.03%	91.01% (24.46%)
NN	91.41%	89.82%	90.92% (27.79%)
VV	88.80%	86.96%	87.67% (12.20%)
AN	88.00%	86.04%	86.89% (10.67%)
DV	88.98%	86.51%	88.03% (10.03%)
NV+NN+VV+AN+DV	91.33%	89.99%	90.70% (38.63%)

From Tables 4a and 4b, the average tonal and toneless STW accuracies of the MWP identifier with the MWP (NV+NN+VV+AN+DV) data for the closed and open test sets are 98.46% and 90.70%, respectively. Between the closed and the open test sets, the differences of the tonal and toneless STW accuracies of the MWP identifier with the MWP (NV+NN+VV+AN+DV) data are 0.49% and 1.34%, respectively. These results strongly support our belief that meaningful word-pairs can be used as application independent knowledge to effectively convert Chinese STW on the MWP-related portion.

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

We selected Microsoft Input Method Editor 2003 for Traditional Chinese (MSIME 2003) as our experimental commercial IME system. In addition, a bigram model called BiGram was developed. The BiGram STW system is a bigram model using Lidstone's law [Manning 1999], as well as forward and backward longest syllable-word first strategies. The system dictionary of the BiGram is comprised of CKIP lexicon and those unknown words found automatically in the *UDN* 2001 corpus by a Chinese word auto-confirmation (CWAC) system [Tsai *et al*. 2003b]. All the bigram probabilities were calculated by the *UDN* 2001 corpus.

MSIME 2003, which uses a statistical trigram-like model [Microsoft Research Center in Beijing], is one of the most widely available input methods. Table 5a compares the results of MSIME 2003, and MSIME 2003 with the MWP identifier on the closed and open test sentences. Table 5b compares the results of BiGram, and BiGram with the MWP identifier on the closed and open testg sentences. In this experiment, the STW output of MSIME with the MWP identifier, or BiGram with the MWP identifier, was collected by directly replacing the identified meaningful word-pairs from the corresponding STW output of MSIME or BiGram. On the MWP-portion in Table 5a,, the tonal and toneless STW accuracies of the MWP word-pair identifier are, respectively, 1.59% and 1.30% greater than those of MSIME 2003. Meanwhile, on the MWP-portion in Table 5b, , the tonal and toneless STW accuracies of the MWP word-pair identifier are, respectively, 1.17% and 1.47% greater than those of BiGram.

Table 5a. The results of tonal and toneless STW experiments for closed and open test sentences, using MSIME, MSIME with MWP identifier, and MWP identifier.

Identified-word	MSIME [a]	MSIME + MWP [b]	MWP identifier [c]
MWP portion	96.87%, 89.40%	-	98.46%, 90.70%
Overall	95.05%, 86.94%	96.30%, 89.79%	-

[a] STW accuracy of the words identified by the Microsoft Input Method Editor (MSIME) 2003

[b] STW accuracy of the words identified by the MSIME 2003 with the MWP identifier

[c] STW accuracy of the words identified by the MWP identifier

Table 5b. The results of tonal and toneless STW experiments for closed and open test sentences, using BiGram, BiGram with MWP identifier, and MWP identifier.

Identified-word	BiGram [a]	BiGram + MWP [b]	MWP identifier [c]
MWP portion	97.29%, 89.23%	-	98.46%, 90.70%
Overall	96.27%, 85.47%	96.75%, 87.74%	-

[a] STW accuracy of the words identified by the BiGram

[b] STW accuracy of the words identified by the BiGram with the MWP identifier

[c] STW accuracy of the words identified by the MWP identifier

To sum up the results and observations of this experiment, we conclude that the MWP identifier can achieve better MWP-portion STW accuracy than the MSIME 2003 and BiGram STW systems. The results show that the MWP identifier can help both MSIME 2003 (trigram-like) and BiGram (bigram base) to increase their tonal and toneless STW accuracies to achieve 96.30%/96.75% and 89.79%/87.74%, respectively. Furthermore, the results indicate that the meaningful word-pairs, or contextual information, can be used to effectively overcome the unseen event and over-weighting problems of SLM models in Chinese STW conversion.

3.4 Error Analysis of STW Conversion

We examine the Top 300 cases in the *tonal* and *toneless* STW conversion from the open testing results of BiGram with the MWP identifier and classify them according to the following three major types of error (see Table 6):

(1) ***Unknown words***: For any NLP system, unknown word extraction is one of the most difficult problems. Since proper names are major types of unknown words, we classify the cases of unknown words into two sub-types and calculate their corresponding percentages, as shown in Table 6.

(2) ***Inadequate syllable segmentation***: When an error is caused by word overlapping, instead of an unknown word problem, we call it *inadequate syllable segmentation.*

(3) ***Homophones***: These are the remaining errors.

Table 6. Three major error types of tonal/toneless STW conversion

Types	Sub-Types	Percentage within this type (%)	Examples	Overall Percentage (%)
Unknown word	Proper names	50 / 50	蘿拉、戴維絲、美邦	11.7 / 10.5
	Other cases	50 / 50	筍農、腋下	
Inadequate syllable segmentation			Tonal cases: 經/嘗試；經常/是、 這些/原油/是；這些/元/郵市、 Toneless cases: 意思/是；以/四史、 的/國家；德國/家	35.9 / 50.8
Homophone			Tonal cases: 需/須、心形/新型、美股/每股 Toneless cases: 主義/注意、雖是/隨時、提醒/體型	52.4 / 38.7

From Table 6, we make the following observations:

(1) The percentages of unknown word errors for the tonal and toneless STW systems are 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.. However, even though unknown words of the first error type have been incorporated in the system dictionary, they could still face the problems of *inadequate syllable segmentation* or *failed homophone disambiguation*.

(2) The major error types of tonal and toneless STW systems are different. To improve tonal STW systems, the major targets should be cases of failed homophone disambiguation. For toneless STW systems, on the other hand, cases of inadequate syllable segmentation should be the focus for improvement..

To sum up the above observations, the bottlenecks of the STW conversion lie in the second and third error types. To resolve these issues, we believe one possible approach is to extend the size of MWP data. Our experiment results show that the MWP identifier can achieve better tonal and toneless STW accuracies than those of BiGram and MSIME 2003 on the MWP-related portion (see the examples given in Appendix B).

4. Conclusions and Directions for Future Research

In this paper, we have applied a MWP identifier to the Chinese STW conversion problem and obtained a high degree of STW accuracy on the MWP-related portion. All of the MWP data was generated fully automatically by using AUTO-MWP on the *UDN* 2001 corpus.

The experiments on STW conversion in [Tsai *et al*. 2002a] and on WSD in [Tsai *et al*. 2002b], as well as the STW experiments in this study, demonstrate that meaningful word-pairs (i.e. contextual information) are key linguistic features of NLP/NLU systems. We are encouraged by the fact that MWP knowledge can achieve tonal and toneless STW accuracies of 98.46% and 90.70%, respectively, for the MWP-related portion of the testing syllables.

The MWP identifier can be easily integrated into existing STW conversion systems by identifying meaningful word-pairs in a post-processing step. Our experiment shows that, by applying the MWP identifier together with BiGram (an optimized bigram model) and MSIME 2003 (a trigram-like model), the tonal and toneless STW accuracies can be improved from 96.27%/85.47% to 96.75%/87.74% and from 95.05%/86.94% to 96.30%/89.97%, respectively.

Currently, our approach is quite basic when more than one MWP occurs in the same sentence (Step 3 in Section 2.2). 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 MWP knowledge to parsing or speech understanding.

The MWP-based approach has the potential to provide the following information for a

given syllable sequence: (1) better word segmentation; and (2) MWP-sentence including the information of five types of MWPs. Such information will be useful for general NLP and NLU systems, especially for syllable/speech understanding and full/shallow parsing. According to our computations, the collection of MWP knowledge can cover approximately 50% of the characters in the *UDN* 2001 corpus.

We will continue to expand our collection of MWP knowledge to cover more characters in the *UDN* 2001 corpus. In other directions, we will try to improve our MWP-based STW conversion with other statistical language models, such as HMM, and extend it to other areas of NLP, especially Chinese shallow parsing and syllable/speech understanding.

5. Acknowledgements

We would like to thank Prof. Zhen-Dong Dong for providing us with the Hownet dictionary.

References

Becker, J.D., "Typing Chinese, Japanese, and Korean," Computer 18(1), 1985, pp. 27-34

Chang, J.S., S.D. Chern and C.D. Chen, "Conversion of Phonemic-Input to Chinese Text Through Constraint Satisfaction," *Proceedings of ICCPOL'91*, 1991, pp. 30-36

Chen, C.G., K.J. Chen and L.S. Lee, "A model for Lexical Analysis and Parsing of Chinese Sentences," *Proceedings of 1986 International Conference on Chinese Computing*, Singapore, 1986, pp. 33-40.

Chen, B., H.M. Wang and L.S. Lee, "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*, 2000.

Chung, K.H., *Conversion of Chinese Phonetic Symbols to Characte*rs, M. Phil. thesis, Department of Computer Science, Hong Kong University of Science and Technology, Sept. 1993.

CKIP. Technical Report no. 95-02, *the content and illustration of Sinica corpus of Academia Sinica*. Institute of Information Science, Academia Sinica, 1995, http://godel.iis.sinica.edu.tw/CKIP/r_content.html

Dong, Z. and Q. Dong, *Hownet*, 1999, http://www.keenage.com/

Fan, C. and P. Zini, "Chinese Character Processing system based on character-root combination and graphics processing," *Document Manipulation and Typography, Proc. of the Int. Conf. on Electronic Publishing, Doc. Manipulation and Typography*, Nice (France), Cambridge University Press, 1988.

Fong, L.A. and K.H. Chung, "Word Segmentation for Chinese Phonetic Symbols," *Proceedings of International Computer Symposium*, 1994, pp. 911-916.

Fu, S.W.K, C.H. Lee and Orville L.C., "A Survey on Chinese Speech Recognition," Communications of COLIPS 6 (1), 1996, pp.1-17.

Gao, J, J. Goodman, M. Li and K.F. Lee, "Toward a Unified Approach to Statistical Language Modeling for Chinese," *ACM Transactions on Asian Language Information Processing*, Vol.1, No.1, 2002, pp. 3-33.

Gu, H.Y., C.Y. Tseng and L.S. Lee, "Markov modeling of mandarin Chinese for decoding the phonetic sequence into Chinese characters," *Computer Speech and Language*, Vol. 5, No. 4, 1991, pp.363-377.

Ho, T.H., K.C. Yang, J.S. Lin and L.S. Lee, "Integrating long-distance language modeling to phonetic-to-text conversion," *Proceedings of ROCLING X International Conference on Computational Linguistics*, 1997, pp. 287-299.

Hsu, W.L. and K.J. Chen, "The Semantic Analysis in GOING - An Intelligent Chinese Input System," *Proceedings of the Second Joint Conference of Computational Linguistics*, Shiamen, 1993, pp. 338-343.

Hsu, W.L., "Chinese parsing in a phoneme-to-character conversion system based on semantic pattern matching," *Computer Processing of Chinese and Oriental Languages*, Vol. 8, No. 2, 1994, pp. 227-236.

Hsu, W.L. and Y.S. Chen, "On Phoneme-to-Character Conversion Systems in Chinese Processing," *Journal of Chinese Institute of Engineers*, 5, 1999, pp. 573-579.

Huang, J.K., "The Input and Output of Chinese and Japanese Characters," *IEEE Computer*, 18, 1, 1985, pp. 18-24.

Kuo, J.J., "Phonetic-input-to-character conversion system for Chinese using syntactic connection table and semantic distance," *Computer Processing and Oriental Languages*, Vol. 10, No. 2, 1995, pp. 195-210.

Lee, Y.S., "Task adaptation in Stochastic Language Model for Chinese Homophone Disambiguation," *ACM Transactions on Asian Language Information Processing*, Vol. 2, No. 1, 2003, pp. 49-62.

Lin, M.Y. and W.H. Tasi, "Removing the ambiguity of phonetic Chinese input by the relaxation technique," *Computer Processing and Oriental Languages*, Vol. 3, No. 1, 1987, pp. 1-24.

Lua, K.T. and K.W. Gan, "A Touch-Typing Pinyin Input System," *Computer Processing of Chinese and Oriental Languages*, 6, 1992, pp. 85-94.

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

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

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

Qiao, J., Y. Qiao and S. Qiao, "Six-Digit Coding Method," *Commun. ACM*, 33, 5, 1984, pp. 248-267.

Sproat, R., "An Application of Statistical Optimization with Dynamic Programming to Phone-

mic-Input-to-Character Conversion for Chinese," *Proceedings of ROCLING III*, 1990, pp. 379-390.

Tsai, J.L. and W.L. Hsu, "Applying an NVEF Word-Pair Identifier to the Chinese Syllable-to-Word Conversion Problem," *Proceedings of 19th COLING 2002*, Taipei, 2002, pp.1016-1022.

Tsai, J.L, W.L. Hsu and J.W. Su, "Word Sense Disambiguation and Sense-based NV Event-Frame Identifier," *Computational Linguistics and Chinese Language Processing*, Vol. 7, No. 1, February 2002, pp.29-46.

Tsai, J.L, G. Hsieh and W.L. Hsu, "Auto-Discovery of NVEF word-pairs in Chinese," *Proceedings of ROCOLING XV*, 2003, pp.143-160.

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

Tsai, J.L, G. Hsieh and W.L. Hsu, "Auto-Generation of NVEF knowledge in Chinese," *Computational Linguistics and Chinese Language Processing*, Vol. 9, No. 1, February 2004, pp.1-24.

Appendix A. Exclusion Word List

I. Monosyllabic exclusion words

/之/的/不/與/兩/再/以/了/較/就/次/得/於/已/把/都/太/一/某/最/
/內/均/原/由/被/全/初/及/將/該/總/塊/項/和/二/從/三/凡/尚/前/
/十/極/番/元/件/甚/因/甲/向/才/四/本/若/先/便/五/粒/常/卅/後/
/左/曾/竟/廿/八/支/六/著/首/剛/應/篇/能/七/終/依/位/暫/共/須/
/中/九/時/可/俱/整/謹/宜/邊/往/批/夥/在/唔/年/諸/略/束/特/磅/

II. Polysyllabic exclusion words

/所以/不能/不會/是否/之間/終於/不必/唯一/西方/恐怕/連續/
/必須/不妨/大家/不得/一旦/初步/據說/看來/全面/臨床/無數/
/依法/國立/過度/突然/通常/一同/單一/大力/純粹/大都/當然/
/種種/大概/國有/順便/總是/不再/默默/無不/那麼/黑白/個人/
/四處/自行/恰好/終究/最佳/一心/十分/甚爲/私立/一起/可以/
/多元/所有/依然/現成/正好/針對/一般/難怪/等到/到底/應該/
/貿然/獨家/原先/根據/微微/不勝/國產/整整/衷心/好些/安然/
/慈善/爲什麼/一下子/一塊兒/非正式/

Appendix B. Three tonal and toneless STW results used in this study (The pinyin symbols and English words in parentheses are included for explanatory purposes only)

I.
Tonal STW results for the Chinese tonal syllable input “jin4 cheng2 jie1 duan4 dou bu2 hui4 gai3 bian4” of the Chinese sentence “近程階段都不會改變”

Methods	STW results	Identified MWP
MWP		不會(V)-改變(V)/VV (key MWP);
		近程(N)-階段(N)/NN;
	近程階段　不會改變	階段(N)-改變(V)/NV
MSIME	進程階段都不會改變	
MSIME+MWP	**近程階段**都**不會改變**	
BiGram	禁城階段都不會改變	
BiGram+MWP	**近程階段**都**不會改變**	

Toneless STW results for the Chinese toneless syllable input “jin cheng jie duan dou bu hui gai bian” of the Chinese sentence “近程階段都不會改變”

Methods	STW results	Identified MWP
MWP		不會(V)-改變(V)/VV (key MWP);
		金城(N)-改變(V)/NV;
	金城階段　不會改變	階段(N)-改變(V)/NV
MSIME	進程階段都不會改變	
MSIME+MWP	**金城階段**都**不會改變**	
BiGram	盡城階段都不會改變	
BiGram+MWP	**金城階段**都**不會改變**	

II.

Tonal STW results for the Chinese tonal syllable input “you2 qi2 zai4 cheng2 shou2 qi2 dao4 gu3 bao3 shi2 lv4 bu4 jia1” of the Chinese sentence “尤其在成熟期稻穀飽實率不佳”

Methods	STW results	Identified MWP
MWP		尤其(ADV)-成熟(V)/DV (key MWP);
	尤其　成熟　稻穀飽實	稻穀(N)-飽實(V)/NV
MSIME	尤其再成熟期稻穀保十率不佳	
MSIME+MWP	**尤其**在**成熟**期**稻穀飽實**率不佳	
BiGram	尤其在成熟期稻穀保時率不佳	
BiGram+MWP	**尤其**在**成熟**期**稻穀飽實**率不佳	

Toneless STW results for the Chinese toneless syllable input “you qi zai cheng shou qi dao gu bao shi lv4 bu jia” of the Chinese sentence “尤其在成熟期稻穀飽實率不佳”

Methods	STW results	Identified MWP
NVEF		尤其(ADV)-成熟(V)/DV (key MWP);
	尤其　成熟　稻穀飽實	稻穀(N)-飽實(V)/NV
MSIME	尤其再成熟期稻穀保十率不佳	
MSIME+MWP	**尤其**再**成熟**期**稻穀飽實**率不佳	
BiGram	尤其在成熟期稻穀保時率不佳	
BiGram+MWP	**尤其**在**成熟**期**稻穀飽實**率不佳	

III.

Tonal STW results for the Chinese tonal syllable input “yi3 li4 gong1 ke4 guan1 ka3” of the Chinese sentence “以利攻克關卡”

Methods	STW results	Identified MWP word-pairs
MWP		攻克(V)-關卡(N)/NV (key MWP);
	以利攻克關卡	以利(V)-攻克(V)/VV
MSIME	以利公克關卡	
MSIME+MWP	**以利攻克關卡**	
BiGram	以利公克關卡	
BiGram+MWP	**以利攻克關卡**	

Toneless STW results for the Chinese toneless syllable input “yi li gong ke guan ka” of the Chinese sentence “以利攻克關卡”

Methods	STW results	Identified MWP word-pairs
MWP		攻克(V)-關卡(N)/NV (key MWP);
	以利攻克關卡	以利(V)-攻克(V)/VV
MSIME	以理工科關卡	
MSIME+MWP	**以利攻克關卡**	
BiGram	以利公克關卡	
BiGram+MWP	**以利攻克關卡**	