Applying an NVEF Word-Pair Identifier to the Chinese Syllable-to-Word Conversion Problem

Jia-Lin Tsai
Intelligent Agent Systems Lab.
Institute of Information Science, Academia Sinica,
Nankang, Taipei, Taiwan, R.O.C.
tsaijl@iis.sinica.edu.tw

Wen-Lian Hsu
Intelligent Agent Systems Lab.
Institute of Information Science, Academia Sinica,
Nankang, Taipei, Taiwan, R.O.C.
hsu@iis.sinica.edu.tw

Abstract

Syllable-to-word (STW) conversion is important in Chinese phonetic input methods and speech recognition. There are two major problems in the STW conversion: (1) resolving the ambiguity caused by homonyms; (2) determining the word segmentation. This paper describes a noun-verb event-frame (NVEF) word identifier that can be used to solve these problems effectively. Our approach includes (a) an NVEF word-pair identifier and (b) other word identifiers for the non-NVEF portion.

Our experiment showed that the NVEF word-pair identifier is able to achieve a 99.66% STW accuracy for the NVEF related portion, and by combining with other identifiers for the non-NVEF portion, the overall STW accuracy is 96.50%.

The result of this study indicates that the NVEF knowledge is very powerful for the STW conversion. In fact, numerous cases requiring disambiguation in natural language processing fall into such "chicken-and-egg" situation. The NVEF knowledge can be employed as a general tool in such systems for disambiguating the NVEF related portion independently (thus breaking the chicken-and-egg situation) and using that as a good fundamental basis to treat the remaining portion. This shows that the NVEF knowledge is likely to be important for general NLP. To further expand its coverage, we shall extend the study of NVEF to that of other co-occurrence restrictions such as noun-noun pairs, noun-adjective pairs and verb-adverb pairs. We believe the STW accuracy can be further improved with the additional knowledge.

1. Introduction

More than 100 Chinese input methods have been created in the past [1-6]. Currently, the most popular input method is based on phonetic symbols. Phonetic input method requires little training because Chinese are taught to write the corresponding pinyin syllable of each Chinese character in primary school. Since there are more than 13,000 distinct Chinese characters (with around 5400 commonly-used), but only 1,300 distinct syllables, the homonym problem is quite severe in phonetic input method. Therefore, an intelligent syllable-to-word (STW) conversion for Chinese is very important. A comparable (but easier) problem to the STW conversion in English is the word-sense disambiguation.

There are basically two approaches for the STW conversion: (a) the linguistic approach based on syntax parsing or semantic template matching [3,4,7,8] and (b) the statistical approach based on the n-gram model where n is usually 2 or 3 [9-12]. The linguistic approach is more laborious but the end result can be more user friendly. On the other hand, the statistical approach is less labor intensive, but its power is dependent on training corpus and it usually does not provide deep semantic information. Our approach adopts the semantically oriented NVEF word-pairs (to be defined formally in Section 2.1) plus other statistical methods so that not only the result makes sense semantically, but the model is also fully automatic provided that enough NVEFs have already been collected.

According to the studies in [13], good syllable sequence segmentation is crucial for the STW conversion. For example, consider the syllable sequence "zhe4 liang4 che1 xing2 shi3 shun4

chang4" of the Chinese sentence "這輛車行駛順暢 (This car moves well)." By dictionary, the two possible segmentation results are (we use "/" to indicate syllable word boundary)
(1) "zhe4/liang4/che1 xing2/shi3/shun4 chang4"
(2) "zhe4/liang4/che1/xing2 shi3/shun4 chang4" using the longest-syllabic-word-first strategy [14]. The two ambiguous portions are /che1 xing2/shi3/ (/{車形,車型}/{使,始,史,駛,矢,屎,豕}/) and /che1/xing2 shi3/ (/{車,硨,蚱}/{行駛,行使}/), respectively. In this case, if the system has the information that "車-行駛(car, move)" is a permissible NVEF word-pair and its corresponding syllable-pair "che1-xing2shi3" has been collected, then the correct segmentation and word-pair "車-行駛(che1-xing2shi3)" of this syllable sequence can be determined simultaneously.

Since NVEF word-pairs are usually the key features of a sentence, if identified correctly, they become good reference words for the *n*-gram models to predict the remaining unconverted syllables. We [15] showed that the knowledge of NVEF sense-pairs and their corresponding NVEF word-pairs (NVEF knowledge) are useful for effectively resolving word sense ambiguity and getting highly accurate word-segmentation for those ambiguous NVEF word-pairs in Chinese.

In this paper, we shall show that the NVEF knowledge can be used effectively in the STW conversion for Chinese. Section 2 describes our approach. The experimental result is presented in Section 3. Directions for future research will be discussed in section 4.

2. Development of an NVEF-based Word Identifier

Hownet [16] is adopted as the system's word-sense dictionary, which provides the knowledge of Chinese lexicon (58,541 words), parts-of-speech (POS) and word senses. We have integrated Chinese words in Hownet, Sinica corpus [17], Cilin (tong2yi4ci2ci2lin2"同義詞詞林") [18], Chinese dictionary (guo2yu2ci2dian3"國語辭典") [19] and Chinese word lists in [20] into a commonly-used machine-readable dictionary (MRD) called *common MRD*, which provides the knowledge of

Chinese lexicon (in which the top 60,000 words are selected from the list of 252,307 words in descending order of word frequency), word frequencies and syllable words. The syllable of each word in common MRD was translated by the inversed process of phoneme-to-character system presented in [4,8]. Word frequency is computed according to a fixed size training corpus consisting of 4,539,624 Chinese sentences obtained from the on-line United Daily News [21] (during the period of 17 January, 2001 to 30 December, 2001).

2.1 Definition of the NVEF Sense-Pair, Word-Pair and Syllable Word-Pair

The sense of a word is defined as its DEF (concept definition) in Hownet. Table 1 lists three different senses of the Chinese word "車 (Che/car/turn)." In Hownet, the DEF of a word consists of its main feature and secondary features. For example, in the DEF "character|文字, surname|姓, human|人, ProperName|專" of the word "車(Che)," the first item "character|文字" is the main feature, and the remaining three items, "surname|姓," "human|人," and "ProperName|專," are its secondary features. The main feature in Hownet can inherit features in the hypernym-hyponym hierarchy. There are approximately 1,500 features in Hownet. Each feature is called a *sememe*, which refers to a smallest semantic unit that cannot be further reduced.

Table 1. Three different senses of the Chinese word "車(Che/car/turn)"

Word	POS/Sense (i.e. DEF in Hownet)
車 Che	N/character|文字,surname|姓,human|人,ProperName|專
車 car	N/LandVehicle|車
車 turn	V/cut|切削

The Hownet dictionary used in this study contains 58,541 words, in which there are 33,264 nouns, 16,723 verbs and 16,469 senses (including 10,011 noun-senses and 4,462 verb-senses). In our experiment, we have also added the DEFs for those words not in Hownet.

A permissible NV word-pair such as "車-行駛 (car-move)" is called a *noun-verb event-frame* (*NVEF*) word-pair. According to the sense of the word "車(Che/car/turn)" and the word "行駛

(move)," the only permissible NV sense-pair for the NV word-pair "車-行駛(car, move)" is "LandVehicle|車"-"VehicleGo|駛." We call such a permissible NV sense-pair an NVEF sense-pair. Note that an NVEF sense-pair is a class that includes the NVEF word-pair instance "車-行駛" as well as the corresponding NVEF syllable word-pair "che1-xing2 shi3."

2.2 Definition of the NVEF KR-Tree

A knowledge representation tree (KR-tree) of NVEF sense-pairs is shown in Fig.1. There are two types of nodes in the KR-tree: *concept nodes* and *function nodes*. Concept nodes refer to words and features in Hownet. Function nodes are used to define the relationships between their parent and children concept nodes. If a concept node A is the child of another concept node B, then A is a subclass of B. Following this convention, we shall omit the function node "subclass" between A and B. Noun-sense class is divided into 15 subclasses according to their main features. They are bacteria, animal, human, plant, artifact, natural, event, mental, phenomena, shape, place, location, time, abstract and quantity.

Three function nodes are used in the KR-tree as shown in Fig. 1:

Figure 1. An illustration of the KR-tree using "人工物(artifact)" as an example noun-sense subclass.

(1) Major-Event (主要事件): The content of its parent node represents a noun-sense subclass, and the content of its child node represents a verb-sense subclass. A noun-sense subclass and a verb-sense subclass linked by a Major-Event function node is an NVEF subclass sense-pair, such as "&LandVehicle|車" and "=VehcileGo|駛" in Fig. 1. To describe various relationships between noun-sense and verb-sense subclasses, we have designed three subclass sense-symbols, in which "=" means "*exact*," "&" means "*like*," and "%" means "*inclusive*." An example using these symbols is given below. Given three senses S_1, S_2 and S_3 defined by a main feature A and three secondary features B, C and D, let $S_1 = A$, B, C, D, $S_2 = A$, B, and $S_3 = A$, C, D. Then, we have that sense S_2 is in the "=A,B" *exact*-subclass; senses S_1 and S_2 are in the "&A,B" *like*-subclass; and senses S_1 S_2, and S_3 are in the "%A" *inclusive*-subclass.

(2) Word-Instance (實例): The content of its children are the words belonging to the sense subclass of its parent node. These words are learned automatically by the NVEF sense-pair identifier according to sentences under the Test-Sentence nodes.

(3) Test-Sentence (測試題): Its content includes several selected test sentences in support of its corresponding NVEF subclass sense-pair.

2.3 An NVEF Word-Pair Identifier

We [15] have developed an NVEF sense-pair identifier for word-sense disambiguation (WSD). This sense-pair identifier is based on the NVEF KR-tree and the techniques of *longest syllabic NVEF-word-pair first* (LS-NVWF) and *exclusion word list* (EWL) checking. By modifying this identifier, we obtain our NVEF word-pair identifier described below.

Step 1. Input a syllable sequence.
Step 2. Generate all possible NV word-pairs whose corresponding NV syllable word-pairs are found in the input sequence. Exclude certain NV word-pairs based on EWL checking.
Step 3. Check each NV word-pair to see if its corresponding NV sense-pairs (there can be several such pairs) can be matched to an NVEF subclass sense-pair in the KR-tree. If one such NV sense-pair matches an NVEF subclass sense-pair in the KR-tree, then this permissible NVEF sense-pairs and their corresponding NVEF word-pairs can be used for the sentence. Resolve conflicts using the LS-NVWF

strategy.

Step 4. Arrange all remaining permissible NVEF sense-pairs and their corresponding NVEF word-pairs in a sentence-NVEF tree. If no NVEF word-pair can be identified from the input sequence, a null sentence-NVEF tree will be produced.

A system overview of the NVEF word-pair identifier is given in Fig. 2. The output of this NVEF word-pair identifier is called a *sentence-NVEF tree*, as shown in Fig. 3.
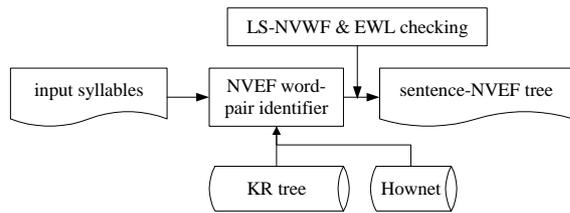

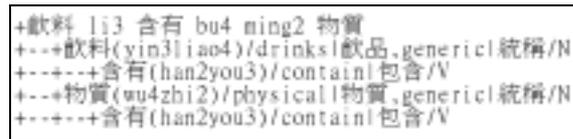Figure 2. A system overview of the NVEF word-pair identifier.

Figure 3. A sentence-NVEF tree for the Chinese syllables "yin3 liao4 li3 han2 you3 bu4 ming2 wu4 zhi2(　　　　　　|There are uncertain matters in the drink)."

2.4 A Word Identifier for the non-NVEF portion

To supplement the NVEF word-pair identifier for the portion of syllable sequence that is not converted by the NVEF knowledge, a separate word identifier is developed. A system overview of the identifier for the NVEF portion and non-NVEF portion is given in Fig. 4. Our word identifier for the non-NVEF portion includes four sub-identifiers whose details are given below:

(1) **Number-classifier-noun phrase (NCN phrase) identifier**: There are many specific linguistic units, such as names, addresses, determinatives-measure compounds (DM) etc. in syllables which need to be recognized in order to supplement the NVEF word-pair identifier (which works in a top-down fashion) Although the number of these linguistic units are infinite, they can be recognized by finite regular

expressions [22]. Following this fact and Chinese grammar, we have developed an NCN phrase identifier to identify phrases consisting of the numbers, classifiers, and nouns, in particular, the commonly-used number-classifier-noun syllable pattern, such as syllables "yi1 bai3 wu3 shi2 ge4 guan1 zhong4 (一百五十個觀眾|one hundred and fifty audience)."

To develop this identifier, we first divide the related words in Hownet into three subclasses for the construction of the NCN phrase, i.e. numbers (the POS is "NUM"), classifiers (the POS is "CLAS") and nouns (the POS is "N".) Secondly, to enrich the knowledge in Hownet, 12 new numbers and 172 new classifiers are added into the original Hownet. Then we create a table listing 13,366 classifier-noun word-pairs (CN word-pairs) and their corresponding CN syllable word-pairs, such as "ge4-guan1 zhong4 (個 - 觀眾)." This table is called the *CN word-pair list*, which is generated by training corpus (Monosyllabic nouns are not considered in this table).
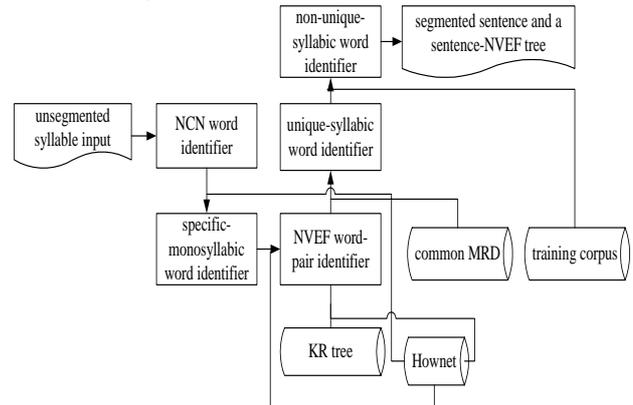

Figure 4. A system overview of the NVEF-based word identifier.

Now, take the syllables "yin1 yue4 ting1 you3 yi1 bai3 wu3 shi2 ge4 guan1 zhong4 (音樂廳有一百五十個觀眾|There are one hundred and fifty audiences in concert hall)" as an example. The NCN phrase identifier will first identify the words of number syllables "yi1 bai3 wu3 shi2(一 百 五 十 |one hundred and fifty)" combined by combining two matching number syllables "yi1 bai3(一百|one hundred)" and "wu3 shi2(五十|fifty)." Secondly, if the number of characters of the recognized number syllables is greater than 1, the NCN word identifier will

continue on checking the following syllables with the CN word-pair list. In this case, since the following syllables "ge4 guan1 zhong4" are found in the CN word-pair list, it will be identified as CN word-pair "個-觀眾."

(2) **Specific-monosyllabic word identifier**: When a monosyllabic word in Hownet has exactly one POS, and that POS is in the set {ADJ (adjective), ADV (adverb), AUX (auxiliary), COOR (coordinator), CONJ (conjunctive), PREP (preposition), STRU (structure word)}, we call this word a *specific-monosyllabic word*. There are 525 specific-monosyllabic words found in the used Hownet.

Consider the following monosyllabic word "已 |already(yi3)." We shall use the POS information of polysyllabic words immediately preceding and following this word to decide if "yi3" should be identified as "已(already)". According to the training corpus, the top 3 preferred POSs of words following "已 (already)" are V (verb), ADV (adverb) and ADJ (adjective). Therefore, the top 3 preferred POSs of syllable words following "yi3" should also be V, ADV and ADJ provided that "已(already)" is to be identified. The top 3 preferred POSs of syllable words preceding and following a specific-monosyllabic word will be called the *top 3 preceding and following preferred POSs*, respectively.

Now, consider the syllable sequence "gong1 cheng2 yi3 wan2 cheng2 le5 (工程已完成了 |The project has been done)" as an example.

First, by checking syllable-by-syllable from left to right, our algorithm recognizes that there is a specific-monosyllabic word "yi3" in this sentence. Then, it will use the longest-syllabic-word-first strategy to identify the syllable word "wan2 cheng2" following "yi3" and the syllable word "gong1 cheng2" preceding "yi3". It will check whether at least one of the distinct POSs of the preceding and following syllable words are within the set of top 3 preceding and following preferred POSs of "yi3", respectively. Since this is indeed the case, the word "已" will be identified.

After the input syllables have been processed by the NVEF word-pair identifier, the NCN word

identifier, and the specific-monosyllabic word identifier, the remaining unconverted syllables will be segmented in a *right-to-left* fashion using the LS-NVWF strategy in the following process.

(3) **Unique-syllabic word identifier**: When a given syllable word maps to exactly one word in the common MRD, we call the mapped word in MRD a unique-syllabic word, e.g. the word "音樂會 /yin1 yue4 hui4/". These unique-syllabic words will be identified directly from right to left.

(4) **Non-unique-syllabic word identifier**: This identifier is used at the very end to deal with those remaining unconverted syllables. It is an *n*-gram based approach. Define the *NVEF frequency* to be the number of sentences including a given NVEF word-pair in the training corpus. First of all, the identifier will select, from the sentence-NVEF tree, the NVEF word-pair having the largest NVEF frequency as the main NVEF word-pair. Recall that the unconverted syllables have been segmented by the longest-syllable-word-first strategy from right to left. Finally, it will convert each segmented syllable word to its corresponding word by the following steps: (a) find all distinctly mapped words of a given syllable word from the common MRD, (b) compute the co-occurrence frequency of each mapped word with the key NVEF word-pair one-by-one in descending order of mapped words' frequencies, (c) whenever the co-occurrence frequency is greater than 0, then convert the given syllable word to this mapped word, (d) if all the computed co-occurrence frequencies in step (b) are 0, the given syllable word will be convert to its mapped word with the largest word frequency.

Take the non-unique syllable word "jin4" in Table A1 as example, the list of its mapped words in descending order of word frequency were "進(enter)/212,481", "近(near)/115,913", "盡(exhaustive)/58,387", "禁(forbid)/17,702", "勁(strongly)/8,089", "晉(Jin Dynasty)/4,524", "浸 (soak)/1,677", "燼 (cinder)/722", "靳 (Jin)/114" and "縉 (red silk)/41." Since the co-occurrence frequency of the mapped words with the key NVEF word-pair "現場-湧入 (locale, enter)" is first greater than 0 at the word

"近(near)", the non-unique syllabic word "jin4" will be converted to the word "近."

Table A1 in the appendix illustrates the complete process of our STW conversion based on the NVEF word identifier.

3. Experimental Results

Define the STW accuracy to be the ratio of the # of correct characters identified over the total # of characters. We use the inverse translator of phoneme-to-character system in [3] to convert a test sentence into a syllable sequence, then apply our STW algorithm to convert this syllable sequence back to characters and calculate its accuracy.

If a sentence contains an NVEF word-pair, this sentence is called an NVEF *identified sentence*. Since the purpose of this study is to demonstrate the effect of applying NVEF word-pair identifier to the STW conversion, we shall focus on converting NVEF identified sentences.

10,000 NVEF identified sentences are randomly selected from the test sentences in the KR-tree to be the *closed test set*; and another 10,000 sentences are randomly selected from Sinica corpus to be the *open test set*. Note that sentences in open test set are not necessarily NVEF identified sentences.

The results of the STW experiment are shown in Table 2 listed in three columns: (1) the NVEF word-pair identifier; (2) the other four sub-identifiers for the non-NVEF protion; and (3) the combination of (1) and (2).

Table 2. The results of the STW experiment.

	(1)	(2)	(3)
Closed test	99.76%	94.65%	97.10%
Open test	99.55%	93.64%	95.97%
Average	99.66%	94.08%	96.50%

For more details, the accuracies of the four identifiers in Section 2.4 are listed in Table 3 below.

Table 3. The STW accuracies of the four sub-identifiers for the non-NVEF portion

	(1)	(2)	(3)	(4)
Closed test	100.00%	94.68%	97.45%	89.01%
Open test	97.25%	94.02%	97.37%	86.10%
Average	98.31%	94.32%	97.41%	87.35%

4. Conclusions and Directions for Future Research

In this paper, we have applied an NVEF word-pair identifier to the Chinese STW conversion problem and obtained excellent rates as shown in Table 2. The knowledge used in this study includes: (1) the NVEF knowledge, (2) the CN word-pair list, (3) the top 3 preferred POSs following or preceding the specific-monosyllabic words, (4) the unique-syllabic word list and (5) the co-occurrence frequency of words with a selected key NVEF word-pairs. Besides the NVEF knowledge in (1), which can be (and has been) generated semi-automatically, the other knowledge can all be trained automatically.

Our database for the NVEF knowledge has not been completed at the time of this writing. The NVEFs are constructed by selecting a noun-sense in Hownet and searching for meaningful verb-sense associated with it. Currently, only 66.34% (=6,641/10,011) of the noun-senses in Hownet have been considered in the NVEF knowledge construction. This results in 167,203 NVEF subclass sense-pairs and 317,820 NVEF word-pairs created in the KR-tree. In the training corpus, about 50% of the sentences includes at least one NVEF word-pair in it.

Based on this experiment, we find that the NVEF-based approach has the potential to provide the following information for a given syllable sequence: (1) well-segmented Chinese sentence, (2) sentence-NVEF tree including main verbs, nouns, NVEF word-pairs, NVEF sense-pairs, NVEF phrase-boundaries, and (3) the CN word-pairs. This information will likely be useful for general NLP, especially for sentence understanding.

The NVEF knowledge is a general linguistic key-feature for sentence analysis. We are encouraged to note that the NVEF knowledge can achieve a high STW accuracy of 99.66% for the NVEF related portion. Our NVEF word identifier can be easily integrated with other existing STW conversion systems by using the NVEF word identifier as a first round filter, namely, identifying words in the NVEF related portion (thus, providing a good fundamental basis) and leaving the remaining unconverted

syllables to other systems with a good potential to enhance their accuracies.

We shall continue our work on covering all the noun-senses in Hownet for the NVEF knowledge construction. This procedure can now be done fully automatically with 99.9% of confidence. The study of NVEF will also be extended to that of other co-occurrence restrictions such as noun-noun (NN) pairs, noun-adjective (NA) pairs and verb-adverb (ND) pairs. Note, however, that the study of these latter pairs will be much simpliefied once the key-feature NVEFs of a sentence have been correctly extracted. We shall also try to improve our NVEF-based approach for the STW conversion and further extend it to speech recognition.

The results in [15] indicate that the NVEF knowledge can also be used effectively for word sense disambiguation. In the future, we shall apply the NVEF knowledge to other fields of NLP, in particular, document classification, information retrieval, question answering and speech understanding.

Acknowledgements

References

1. Huang, J. K. 1985. The Input and Output of Chinese and Japanese Characters. IEEE Computer, 18(1):18-24.

2. 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.

3. 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, 338-343.

4. 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.

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

6. 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.

7. 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.

8. 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.

9. 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.

10. 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.

11. 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.

12. 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.

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

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

15. Tsai, J. L, W. L. Hsu and J. W. Su. 2002. Word sense disambiguation and sense-based NV

event-frame identifier. Computational Linguistics and Chinese Language Processing, 7(1):29-46.

16. Dong, Z. and Q. Dong, Hownet,

http://www.keenage.com/

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

http://godel.iis.sinica.edu.tw/CKIP/r_content.html

18. Mei, J. *et al*. 1982. Tong2Yi4Ci2Ci2Lin2 （同義詞詞林）, Shanghai Dictionary Press.

19. Taiwan's Ministry of Education. 1998. Guo2Yu2Ci2Dian3 (Electronic Chinese Dictionary),

http://www.edu.tw/mandr/clc/dict/

20. Tsai, C. T. (2001) A Review of Chinese Word Lists Accessible on the Internet. Chih-Hao Tsai Research Page

http://www.geocities.com/hao510/wordlist/.

21. On-Line United Daily News,

http://udnnews.com/NEWS/

22. Huang, C. R. *et al*. 1996. Readings in Chinese Natural Language Processing. Journal of Chinese Linguistics, 9:1-174.

Appendix A (this is optional for the reviewers)

Table A1. An illustration of our STW conversion on the syllable input "yin1 yue4 hui4 de5 xian4 chang2 yong3 ru4 jin4 liang3 qian1 ming2 guan1 zhong4" of the Chinese sentence "音樂會的現場湧入近兩千名觀眾(There are near two thousands audience entering the concert hall)"

Identifier	Temp output	Temp not-converted syllables
NCN word	[A]=[兩千-名-觀眾]	yin1 yue4 hui4 de5 xian4 chang2 yong3 ru4 jin4 [A]
Specific-monosyllabic word	[B]=[的]	yin1 yue4 hui4 [B] xian4 chang2 yong3 ru4 jin4 [A]
NVEF word-pair	[C₁]=[現場-湧入] [C₂]=[觀眾] [A]=[兩千-名]	yin1 yue4 hui4 [B] [C₁] jin4 [A] [C₂]
Unique-syllabic word	[D]=[音樂廳]	[D] [B] [C₁] jin4 [A] [C₂]
Non-unique-Syllabic word	[E]=[近]	[D] [B] [C₁] [E] [A] [C₂]

Final output:
```
+      /  /   /   / /   / /
+-+    (guan1 zhong4)/human|  ,*look|  ,#entertainment|  ,#sport|    ,*recreation|    /N
+-+-+    (yong3 ru4)/GoInto|    /V
+-+    (xian4 chang2)/place|    ,#fact|    /N
+-+-+    (yong3 ru4)/GoInto|    /V
+-+-+-+      (yin1 yue4 hui4/unique-syllabic word)
+-+-+-+    (de5/specific-monosyllabic Word)
+-+-+-+    (jin4/non-unique-syllabic word)
+-+-+-+      (liang3 qian1 ming2/NCN word)
```