A KEY-PHRASE UNDERSTANDING FRAMEWORK INTEGRATING REAL WORLD KNOWLEDGE WITH SPEECH RECOGNITION WITH INITIAL APPLICATION IN VOICE MEMO SYSTEMS FOR MANDARIN CHINESE

Bor-shen Lin[1], Hsin-min Wang[2], and Lin-shan Lee[1],

[1] Department of Electrical Engineering, National Taiwan University
[2] Institute of Information Science, Academia Sinica
Taipei, Taiwan, Republic of China
e-mail: bsl@speech.ee.ntu.edu.tw

ABSTRACT

Automatic speech recognition by computers can provide the most natural and efficient method of communication between humans and computers. The key-issue in some specific applications such as electronic shopping on WWW is not just to improve the syllable or word recognition accuracy but to achieve correct understanding of the speech. Thus, key-phrase understanding, which integrates real world knowledge with key-phrase recognition techniques and then achieves partial speech understanding, plays an important role in such areas. In this paper, a general framework for such a key-phrase understanding problem is presented and an initial application to be used in voice memo systems for Mandarin Chinese has been developed.

1.INTRODUCTION

Just as the keyword spotting techniques which can be applied to specific tasks of speech recognition without recognition of the complete sentences, key-phrase understanding, which integrates real world knowledge with key-phrase recognition techniques is believed to be a good approach to achieve partial speech understanding or limited intelligence for specific application domains. A good example may be voice salesmen system on WWW, which needs only a very limited vocabulary and key-phrase patterns for the dialogues in electronic shopping. The major goal of speech recognition for that purpose should be simply to understand correctly the dollars to be paid and the goods to be purchased, so complete semantic analysis or a large vocabulary speech recognition system actually won't work better than a specially designed key-phrase understanding system. Thus, the framework which integrates the real world knowledge into speech recognition so as to achieve 'correct understanding' of the speech is the primary focus of this paper.

The rest of this paper is organized as follows. The overall architecture of the key-phrase understanding framework is introduced in section 2, while an initial application for date-time transcription to be used in Mandarin voice memo systems is discussed in section 3. Finally, some discussions and conclusion remarks are given in section 4.

2.OVERALL ARCHITECTURE OF THE KEY-PHRASE UNDERSTANDING FRAMEWORK

The overall view of the key phrase understanding framework is shown in Fig. 1. The kernel of the framework is based on a tree-spanning algorithm [1,2] with the leaf-nodes being the words together with a 3-level pruning scheme. First, a vocabulary set consisting of all words and word classes needed for possible key-phrases for the given task should be defined, and all words are tagged with their word classes and represented as a tree lexicon[1]. Second, a set of representative sentences containing such key-phrases should be generated, and used to obtain a word class transition matrix using an automatic learning algorithm probably with some possible exception rules. Because the above process does not need a large number of training sentences, task-portability can be maximized. The word class transition matrix, of course, does not have to be very rigid. It's simply used for the first level pruning (to delete illegal word transition) in the tree-spanning algorithm to reduce the search space. The second level pruning not only serves as the double-check, but includes a data structure representing the real world knowledge for the key-phrases, and a set of firing rules generated with the concept of production systems[2]. When a word node ramifies, its corresponding rule fires to update the real world knowledge (i.e. the content of the data structure). If the accumulated knowledge up to any spanned word-node is judged illegal by the firing rules (i.e., can't be understood), the spanning from that node simply stops.

With this second level pruning, all survival paths are 'temporarily legal' (accepted by the firing rules) up to the word nodes but not necessarily complete in the meaning from the view point of knowledge. So the knowledge reestimation serves as the third level pruning to delete the incomplete (in meaning) paths. After three-level pruning, all survival paths can be transcribed into the corresponding knowledge representations. The system finally chooses the path with the highest accumulated acoustic score among all candidate paths within the tree, and takes its word sequence with corresponding knowledge representation as the output.

3. AN APPLICATION: DATE-TIME TRANSCRIPTION FOR MANDARIN VOICE MEMO SYSTEMS

The date-time expressions in Chinese language are not trivial. Most of them are unit-based with many rules. For example, February is expressed as "Month two" and 15th as "day 15". So the words "month", "day" may belong to the same word class, while numbers to another word class. Moreover, these units are not necessary unique, e.g. the "weekday unit" may be pronounced as 'hsin-chih'(meaning 'week') , 'li-bai'(meaning 'worship'), or 'jou'(meaning 'week' or 'cycle') but with different usage constraints. There can also be various types of flexible date-time expressions, such as date alias(e.g. mothers' day), expression by reference(e.g. next week), time range(e.g. afternoon) or transformation between different calendar systems, etc. The following are some examples:

1. 'er'(two) 'shy'(ten unit) 'i'(one) 'rhy'(day unit) : 21th of this month
2. 'hsia'(next) 'ge'(transition) 'uie'(month unit) 'shy'(ten unit) 'wu'(five) hau(day unit) : 15th of next month
3. 'hsia'(next) 'hsia'(next) 'hsin-chih'(week) 'san'(three) : next next Wednesday
4. 'i'(one) 'jou'(nine) 'jou'(nine) 'chih'(seven) 'nien'(year unit) 'uan-dan'(Jan 1st): 1st Jan 1997
5. 'mu-chin-je'(mothers' day) 'hsia-wu' (afternoon) 'lian'(two) 'dien'(hour unit) 'ban'(half): 2:30 p.m. on mothers' day
6. 'shi-uan'(year type, A.D.) 'lian'(two) 'chien' (thousand unit) 'nien'(year unit) : year 2000
7. 'min-guo'(year type, local calendar) 'ba'(eight) 'shy'(ten unit) 'wu'(five) 'nien' (year unit) : year 1996 after transformation from local calendar

As described in the above examples, the date-time transcription framework must be able to accept various types of date-time expressions and transcribe them into the desired date-time knowledge representation (e.g. year/month/date/time). All these knowledge can be well represented and taken care of in the general framework discussed above. A partial list of the tree lexicon and an example to describe how the pruning scheme is applied in the word tree spanning for date-time transcription framework are shown in Fig. 2 and Fig. 3 respectively. This framework can be used in voice memo systems including applications such as automatic notification or information retrieval according to the date-time transcription.

3.1 Experimental Results

An initial application module has been successfully implemented using the acoustic recognition module previously developed for large vocabulary Mandarin speech recognition [3], which can transcribe any spoken date-time expression in Mandarin speech into the desired knowledge representation (e.g. year/month/date/time). After the topN syllables recognized and a syllable lattice constructed in the acoustic processor, the understanding framework described above was applied instead of the statistical Chinese language model. The preliminary experimental results of four speakers (2 male and 2 female), each uttered 25 date-time expressions in continuous Mandarin speech, are shown in Table 1. In our experiments, the real-time tests based on the speaker-independent (SI) recognition mode were first done, and then the recorded speech was tested again based on the speaker-dependent (SD) recognition mode off-line, in which the SD acoustic model was trained by 260 phonetically balanced Chinese short sentences uttered by each tester. These results indicate that around 82% and 68% of date-time expressions can be correctly transcribed under the speaker-dependent mode and the speaker-independent mode respectively, while the error rates are around 9% and 26% respectively. However, if the users speak more clearly, the accuracy can be further up to 90%. It was found that, among the errors, around 50% of them were in fact partially correct, i.e., they could be partially understood. It was also found that most errors were speech recognition errors. That is, if any correct syllable was not included in the syllable lattice, the following understanding framework would eventually fail. So, we are trying to integrate the speech recognition process and the understanding framework into a one-stage process, and hope to improve the accuracy.

4.DISCUSSIONS AND CONCLUDING REMARKS

The key-phrase understanding framework described above indeed gives up the conventional 'acoustic units(phones or other similar ones) -> words -> semantic' hierarchy usually assumed in speech understanding. In this framework, the selection of the topN word sequences and the semantic analysis are actually integrated in the

tree-spanning algorithm with higher level knowledge used for pruning. The final path selected is optimal in 'intelligibility' instead of acoustic scores. This framework is in fact with very good portability, i.e., it can be easily applied to many other applications such as a voice salesman system on WWW which needs only a very limited vocabulary and key-phrase patterns for dialogues in electronic shopping, a Chinese address input system for voice retrieval of Chinese digital maps which though needs a relatively large vocabulary yet very limited key-phrase patterns, and so on.

REFERENCES

[1]Hung-yun Hsieh, et al, "Use of Prosodic Information to Integrate Acoustic and Linguistic Knowledge in Continuous Mandarin Speech Recognition with Very Large Vocabulary", ICSLP, Vol. 2, pp. 809-812, Oct. 1996.

[2]George F., et al, "Artificial Intelligence", Chap3-Chap5, Benjamin/Cummings Publishing Company, Inc., 1993.

[3]Hsin-min Wang, et al, "Complete recognition of continuous Mandarin speech for Chinese language with very large vocabulary but limited training data", IEEE Trans. On Speech and Audio Processing, Vol. 5, No. 1, Jan. 1997.

[4]K. Sparck, Johns, G. J. F. Johns, J. T. Foote, and S. J. Young, "Experiments in spoken document retrieval", Information Processing & Management, Vol. 32, No. 4, pp. 399-417, 1996.

speaker	SI			SD		
	C	R	E	C	R	E
m1	60.0	8.0	32.0	68.0	16.0	16.0
m2	72.0	4.0	24.0	92.0	0	8.0
f1	72.0	8.0	20.0	88.0	8.0	4.0
f2	68.0	4.0	28.0	80.0	12.0	8.0
average	68.0	6.0	26.0	82.0	9.0	9.0

Table 1: The correct rate (C), reject rate (R), error rate (E) for transcription of the date-time expressions of 4 test speakers.

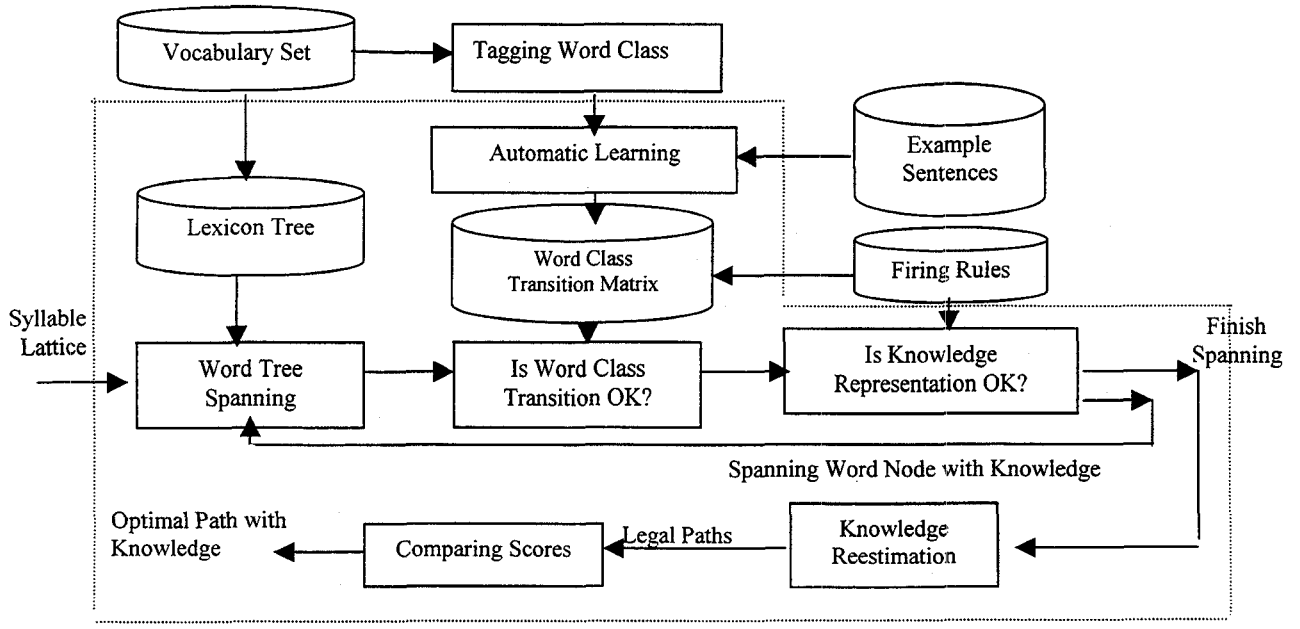

Fig. 1: Overall architecture of the understanding framework for key-phrase understanding (Dash-line enclosed area is the task-independent kernel.)

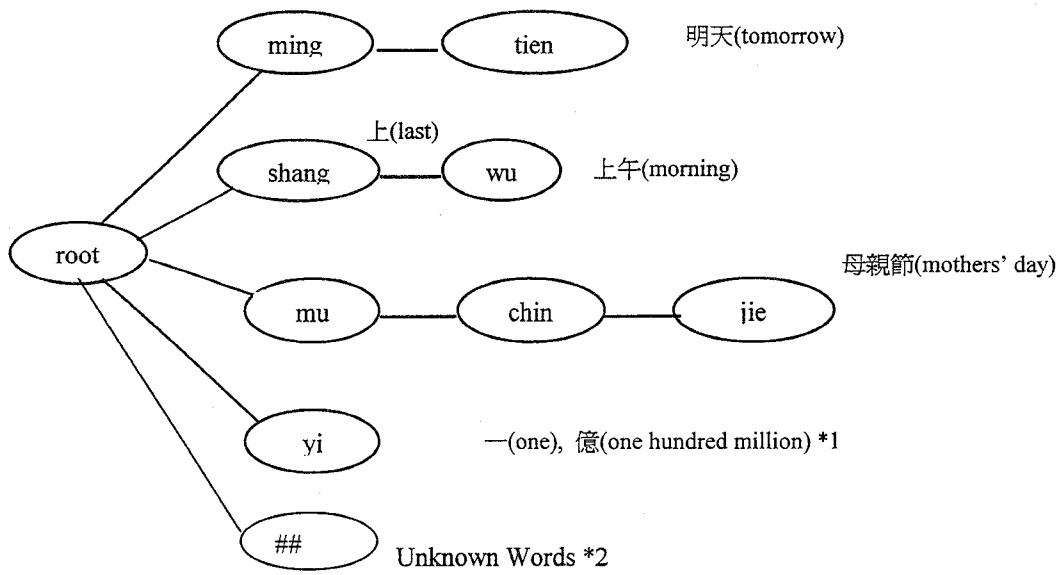

Fig. 2: Part of Lexicon Tree for Date-Time Transcription

*1. This is node with homonyms.

*2. This is node for insertions between words.

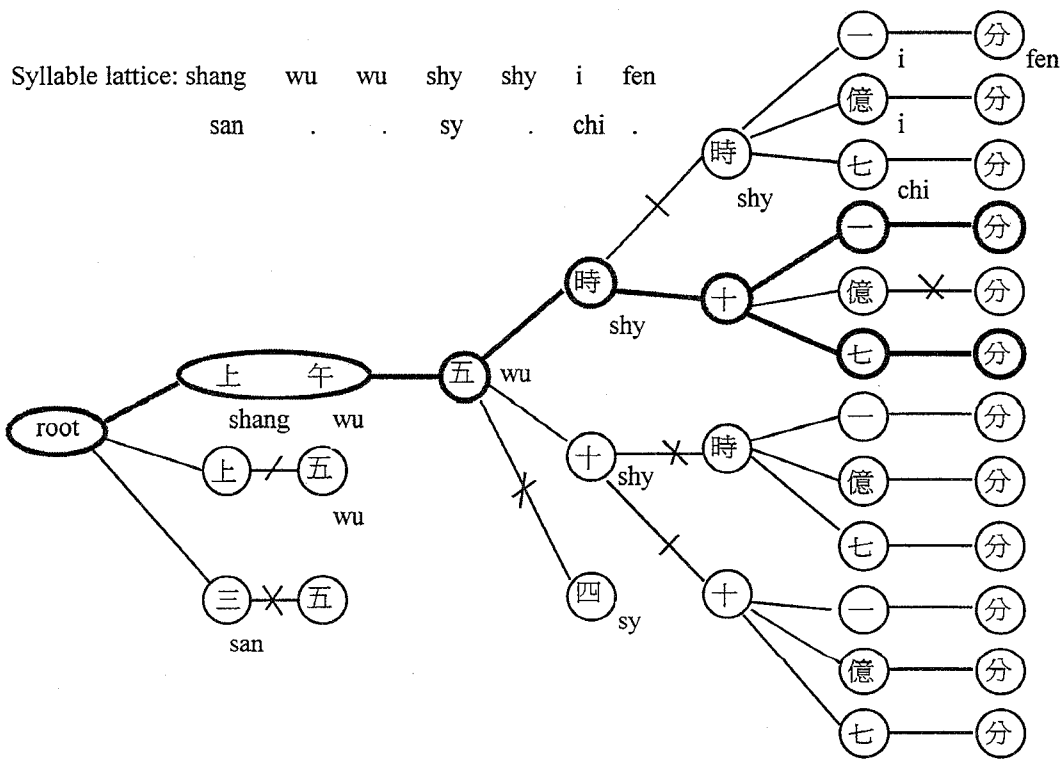

Fig.3: Pruning of word tree spanning by the word class transition matrix (/) and the firing rules (X).