7E-8

LANGUAGE PROCESSING FOR A LARGE VOCABULARY ISOLATED SPEECH RECOGNITION SYSTEM

K.H. Loken-Kim and Yasuhiro Nara

Fujitsu Laboratories

1. INTRODUCTION

While speaker-independent continuous speech recognition technology has been hobbling along with many technical difficulties, speaker-dependent isolated speech recognition technology has made steady progress and has been successfully applied to many commercial applications. This speaker-dependent isolated speech recognition technology, with its limitations, has reached to the point where a very large vocabulary word processor and a data entry system are conceivable.

In this paper, we describe the architecture of a postprocessor that supports a large vocabulary speaker-dependent isolated speech recognition system (Fujitsu Model F2361A), and report the result of the performance evaluation.

2. POSTPROCESSOR

The postprocessor (Figure 1) consists of a blackboard which is surrounded by several knoweldge sources. The activation of the knowledge sources is coordinated by the controller.

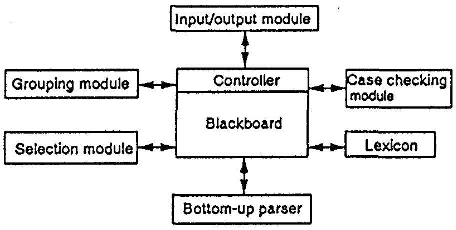

Figure 1. Postprocessor

The input/output module functions as a communication medium for the speech recognition system, controller, and a user. It transmits the recognition results, and the user's commands to the blackboard, and reports the hypothesized sentences to the user.

When a sentence is spoken, the result of the recognition (bunsetsu lattice) is placed on the blackboard. Then, the grouping module divides the bunsetsu lattice into several groups based on the magnitude of the distance score differences. Depth-first-search with grouping may find the sentence with a higher probability of correctness quicker than without grouping because: 1) higher ranked bunsetsu candidates (e.g. candidate number 1) is more likely the bunsetsu actually spoken than lower ranked candidates (e.g. candidate number 8), and 2) if the difference of distance scores between the first and the second candidate is very small, one of them is likely what actually was spoken.

Once the grouping of the bunsetsu lattice is completed, the select module searches through it to select a sentence based on a depth-first-search while higher priority is given to higher ranked candidate groups. The depth-first-search is conducted in two phases: 1) inter-group depth-first-search, and 2) intra-group depth-first-search. Search time for the correct sentence is often reduced significantly.

A sentence selected by the select module is placed on the blackboard to be parsed. In the parsing module, grammar rules are expressed in the Definite Clause Grammar (DCG) [1], and sentence parsing is conducted in bottom-up fashion [2] while generating multiple phrase structures through backtracking whenever possible. Generating multiple phrase structures for a sentence enables us to choose the intended phrase structure.

The case checking module receives the phrase structure and performs case analysis. In this study, case analysis is conducted based on the Lexical Functional Grammar (LFG) [3] due to its transparent manipulation of syntactic information, and it can be easily extended to evaluate the semantic compositionality of a sentence.

LFG is a unification grammar which minimizes the need for directional analysis. In LFG, two bunsetsu phrases can be unified if the case of one bunsetsu subcategorizes the other. Evaluating whether the subcategorization can take place or not starts by augmenting the phrase structure with metavariables which are then instantiated with actual variables. The actual variable carries syntactic features originated from both grammar rules and the lexicon. The resulting phrase structure is called a constituent structure (c-structure). This c-structure is, then, used to drive a functional structure through unification [4]. In the future, the functional structure will be used for the semantic analysis.

3. EVALUATION

The speech recognition system (Fujitsu Model F2361A) used for this study can handle up to 4000 words and phrases, allowing input in either connected or isolated mode. In connected mode, spoken input of up to 12 words every 3 seconds can be accepted. This experiment, however, was conducted in isolated mode.

The subject who participated in our experiment was a 35 year old male native Japanese speaker from the Tokyo area.

The selection of 4000 bunsetsu was accomplished by collecting frequently appearing words and phrases in the first year junior high school English textbooks, and 204 simple test sentences were composed using the subset of the selected words and phrases.

The test bunsetsu phrases and sentences were recorded in a quiet studio. A Sony condensor microphone (C-388B) was used by the subject, allowing 2 to 3 seconds between each bunsetsu, and was recorded with a Nippon Columbia reel-to-reel tape recorder (DN3301) on Sony magnetic tapes

Language Processing for a Large Vocabulary Isolated Speech Recognition System
K.H. Loken-Kim and Yasuhiro Nara
Fujitsu Laboratories

(PLN-370B).

After the recording was finished the contents were transferred to compact digital audio tapes (Technics RT-R120) using a Nagra (IV-S) and a Sony DAT (DTC-1000ES) tape recorder.

The training of the speech recognizer and generation of bunsetsu lattice was conducted by transmitting each bunsetsu directly from the line out jack of the DAT to the microphone jack of the speech recognizer via a connecting cable (Sony RK-C71). After the training was completed, each of the 204 test sentences was transmitted to the recognizer. The recognition system generated 8 candidate bunsetsu phrases for each bunsetsu.

After a sentence was spoken, all the bunsetsu candidates were saved as a bunsetsu lattice and the next sentence was transmitted. When bunsetsu matrices for all 204 sentences were collected, the performance of the speech recognition system was evaluated at both the bunsetsu and sentence levels.

Both bunsetsu and sentence level performance were obtained by simply counting the number of bunsetsu phrases and sentences recognized correctly. However, the sentence level recognition rate was obtained for three different categories as follows: 1) all the bunsetsu phrases in the spoken sentence were recognized as the first candidates, 2) not all the bunsetsu were recognized as the first candidates but were recognized as a candidates, and 3) one or more bunsetsu phrases were not recognized as candidates.

After the performance of the speech recognition system was evaluated, the entire bunsetsu lattices were fed to the postprocessor to analyze its behavior. The postprocessor was evaluated in four different ways: first with only a depth-first-search, second with a depth-first-search and grouping, third with a depth-first-search, grouping, and parsing, and fourth with a depth-first-search, grouping, parsing, and case checking module.

4. RESULTS

The recognition accuracy of the speech recognition system was evaluated at both the bunsetsu and sentence levels.

A total of 705 bunsetsu phrases were spoken: 501 (71%) were recognized as the first candidates, and 104 were recognized as one of the first 7 candidates (14.7%). This resulted in an 85.8% recognition rate. The remaining 100 bunsetsu phrases were not recognized. The probable causes for this low recognition rate are: 1) existence of many minimal pair bunsetsu phrases in the 4000 bunsetsu phrases that were registered, and 2) lack of sufficient voice patterns. Each of the 4000 bunsetsu phrases was registered with only one voice pattern.

A total of 204 sentences were spoken: 61 (29.9%) had all the bunsetsu phrases in each sentence recognized as the first candidates, and 60 (29.4%) had all the bunsetsu phrases in each sentence recognized within the first 7 candidates. This resulted in a 59% recognition rate. The remaining 83 sentences (40.6%) had one or more bunsetsu phrases not recognized.

The performance of the postprocessor was evaluated by observing how fast it could recover a spoken sentence from the candidate sentences. The depth-first-search through the bunsetsu lattice can generate many candidate sentences, and each of the knowledge sources of the postprocessor was designed to eliminate unlikely sentences, hence moving likely sentences toward the top of the candidate sentence list.

The result of the postprocessor performance evaluation is summarized in Table 1.

(unit: sentence)

K.S. RANK(R)	A	B	C	D
0 < R ≤ 5	10	25	33	37
5 < R ≤ 10	15	12	10	8
10 < R ≤ 50	9	9	6	5
50 < R	26	14	11	10

K.S.: Knowledge Source
A: After depth-first-search was performed
B: After grouping was performed
C: After parsing was performed
D: After case checking was performed
R: Rank in candidate sentence list

Table 1. Performance of the Postprocessor

When the postprocessor was provided with only depth-first-search capability, 10 of the test sentences were found within one of the first 5 candidate sentences provided by the postprocessor in each case. But after the grouping capability was added, the number of sentences increased to 25, suggesting that grouping helped to locate likely sentences. After parsing, many ungrammatical sentences were eliminated and the number of sentences recovered within the first 5 candidate sentences provided by the postprocessor was increased to 33. This number was further increased to 37 after case checking was performed on all the parsed sentences.

For our test, we checked the first 50 sentences generated as candidates by the postprocessor. Within this limitation, we recovered 50 out of the 60 sentences: the remaining 10 were unrecovered.

5. CONCLUSION

In this paper, we have presented the design philosophy of a postpocessor that supports a large vocabulary speech recognition system.

The postprocessor consists of knowledge sources that are applied to recover a spoken sentence by searching through a bunsetsu lattice. 60 sentences out of 204 test sentences were used to evaluate the sentence recovery capability of the postprocessor, and 50 out of the them were recovered. This indicates that each of the knowledge sources contributed to a speedier recovery of the spoken sentence, thus increasing the probability that the correct sentence would appear earlier in the candidate sentence list.

REFERENCES

[1] F. Pereira and D. Warren, "Definite Clause Grammars for Language Analysis-A Survey of the Formalism and a Comparison with Augmented Transition Networks", Artificial Intelligence 13, pp. 231-278, North-Holland Publishing Company, 1980.

[2] H. Tanaka, "Sentence Analysis Using Prolog", Computer Today, pp. 40-47, May, 1984.

[3] J. Bresnan, "The Mental Representation of Grammatical Relations", MIT Press, 1983.

[4] K.H. Loken-Kim and Y. Nara, "A Postprocessor for a Large Vocabulary Japanese Speech Recognition System", Euro Speech, September, 1989.