A POSTPROCESSOR FOR A LARGE VOCABULARY JAPANESE SPEECH RECOGNITION SYSTEM

K. H. Lokken-Smith and Y. Nara

Speech Recognition Group, Information Processing Division
Fujitsu Laboratories Ltd., Kawasaki 1015, Kamikodanaka,
Nakahara-ku, Kawasaki 211, Japan

ABSTRACT

In this paper, we explain the design philosophy of a postprocessor that supports a large vocabulary Japanese speech recognition system (Fujitsu Model 2361A). The postprocessor improves performance of the speech recognition system by selecting syntactically plausible sentences from the sentence candidate list. Selection of the sentences is achieved by 1) searching through a bunsetsu lattice (group of short Japanese phrases) which is grouped based on distance scores, 2) use of a bottom-up parser, and 3) case checking based on lexical functional grammar. The result of the performance evaluation was that the postprocessor recovered 21 sentences out of 26 resulting in an 80% recovery rate.

1. INTRODUCTION

Despite the recent strides made in speech technology, commercial applications are still limited due to low reliability. This poor performance is caused by such factors as speech variation. In many cases, sentences recognized by a speech recognition system are severely distorted and it is necessary to apply high level knowledge sources to recover them [1, 2]. However, most of the commercial speech recognition systems do not have the ability to verify the syntactic coherence of recognition results.

The recognition results in an isolated word recognition system are usually obtained as several candidate words for a spoken word. When a sentence is spoken, the candidate words form a group which is collectively called a word lattice. Searching through the word lattice for a sentence requires high level knowledge sources such as syntax, semantics, and pragmatics.

The purpose of this research is to develop a postprocessor that receives a recognized sentence in word lattice form from a speech recognition system, and applies syntactic knowledge to identify a grammatically plausible sequence of words in the lattice.

2. BUNSETSU AND BUNSETSU LATTICE

In Japanese, a bunsetsu is the smallest unit that carries a grammatical function within a sentence [3]. Although there is no clear definition of what a bunsetsu is, it is generally considered to be a unit spoken in a single breath, and consists of a conceptual word (e.g., noun, verb, adjective, etc.) and one or more functional words (e.g., particle, etc.) [4]. The details of the intra-bunsetsu structure can be found in Kido [5].

The advantages of using a bunsetsu as the basic speech recognition unit are twofold. First, most functional words are very short in Japanese, and it is not only awkward to utter a functional word separately but also difficult to obtain a high degree of recognition accuracy. When one or more functional words are combined with a conceptual word to make up a bunsetsu, it is easier to enunciate thus a higher degree of recognition accuracy can thus be obtained.

Second, there is a syntactic interdependency between two bunsetsu phrases. An analysis of this interdependency enables us to assess the structural soundness of a sentence recognized by a speech recognition system.

When a bunsetsu is spoken, the speech recognition system used for this study (Fujitsu Model 2361A) generates up to eight bunsetsu candidates with their corresponding distance scores (the similarity between input and reference voice patterns). For example, when the sentence “anata-wa hon-o yondeimasuka (Are you reading a book)” was spoken, the eight bunsetsu candidates listed in Table 1 were generated for each bunsetsu (Knowing the meaning of each Japanese bunsetsu is not important to read this paper). This collection of bunsetsu candidates is called a bunsetsu lattice [6]. It is the task of the postprocessor to identify the spoken sentence from the candidate bunsetsu phrases in the lattice.

<table>
<thead>
<tr>
<th>Bunsetsu</th>
<th>Corresponding Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>anata-wa</td>
<td>hon-o (a book)</td>
</tr>
<tr>
<td>yondeimasuka</td>
<td>reading</td>
</tr>
<tr>
<td>anata-tu</td>
<td>hon-o (2369)</td>
</tr>
<tr>
<td>yondeimasuka</td>
<td>suneimasuka(2407)</td>
</tr>
<tr>
<td>anata-sa</td>
<td>hon-no (2836)</td>
</tr>
<tr>
<td>yondeimasuka</td>
<td>yondeimsukaa(2706)</td>
</tr>
<tr>
<td>anata-ga</td>
<td>hon-to (3100)</td>
</tr>
<tr>
<td>yondeimasuka</td>
<td>oyodeimasuka(2847)</td>
</tr>
<tr>
<td>anata-ku</td>
<td>hon-no (2644)</td>
</tr>
<tr>
<td>yondeimasuka</td>
<td>yondeinsukaa(2779)</td>
</tr>
<tr>
<td>yonagi-go</td>
<td>koten-no (3176)</td>
</tr>
<tr>
<td>yondekepersa(2935)</td>
<td></td>
</tr>
<tr>
<td>yonagata(2804)</td>
<td>bona-o (3267)</td>
</tr>
<tr>
<td>yonodemaa(2595)</td>
<td></td>
</tr>
<tr>
<td>hon-o(2237)</td>
<td>mon-o-mo(3314)</td>
</tr>
<tr>
<td>kotsu-teimasuka(3042)</td>
<td></td>
</tr>
<tr>
<td>hon-to(3100)</td>
<td>mono-mo(3314)</td>
</tr>
<tr>
<td>yondeimasuka</td>
<td>kotsu-teimasuka(3042)</td>
</tr>
</tbody>
</table>

3. POSTPROCESSOR

The postprocessor currently being developed at Fujitsu Laboratories is a blackboard system (Figure 1). The blackboard is surrounded by several knowledge sources which communicate with each other through the blackboard. The contents of the blackboard are monitored by the controller which coordinates the activation of the knowledge sources.

![Figure 1. The Postprocessor](image-url)
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. This grouping of bunsetsu candidates works because higher ranked bunsetsu candidates (e.g., candidate number 1) is more likely the bunsetsu actually spoken than lower ranked candidates (e.g., candidate number 3). Depth-first-search with grouping may find the sentence with a higher probability of correctness quicker than without grouping.

For example, consider the bunsetsu ‘anata-wa’. An average distance score difference of 109 was calculated between candidates. The division of the candidates took place where the distance score difference was greater than the average distance score difference. The bunsetsu candidates thus were divided into three groups. The entire bunsetsu lattice was grouped by repeating the same procedure (Table 2).

<table>
<thead>
<tr>
<th>Bunsetsu Candidates</th>
<th>Distance Score</th>
<th>Average</th>
<th>Sub-groups</th>
</tr>
</thead>
<tbody>
<tr>
<td>anata-to(2058)</td>
<td>429</td>
<td>109</td>
<td>group 1</td>
</tr>
<tr>
<td>anata-wa(2497)</td>
<td>85</td>
<td>109</td>
<td></td>
</tr>
<tr>
<td>anatatagata-wa(2582)</td>
<td>124</td>
<td>109</td>
<td>group 2</td>
</tr>
<tr>
<td>anata-ga(2706)</td>
<td>26</td>
<td>109</td>
<td></td>
</tr>
<tr>
<td>hanako-wa(2731)</td>
<td>22</td>
<td>109</td>
<td></td>
</tr>
<tr>
<td>yamada-wa(2750)</td>
<td>51</td>
<td>109</td>
<td></td>
</tr>
<tr>
<td>anata-wa(2804)</td>
<td>17</td>
<td>109</td>
<td>group 3</td>
</tr>
</tbody>
</table>

Table 2. Bunsetsu Lattice

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 (Figure 2 a), d)) and 2) intra-group depth-first-search (Figure 2 b, c)). 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. The parsing module contains a Bottom-Up-Parser (BUP) [7], which is used to parse a selected sentence, and to generate multiple phrase structures through backtracking whenever possible. Generating multiple phrase structures for a sentence enables us to choose the intended phrase structure.

In the parsing module, grammar rules are expressed in the Definite Clause Grammar (DCG) [8], which works in top-down fashion. The existence of a left-recursive grammar rule may thus induce infinite loop. This can be avoided by dividing a left recursive grammar rule such as \( A \rightarrow A \ a \ b \) into \( A \rightarrow A' \ a \) and \( A' \rightarrow a \ A' \ b \). However, this increases complexity and reduces the transparency of the grammar rules.

Another way to avoid looping is to parse the sentences in bottom-up fashion. An example of a Prolog program that parses a sentence in bottom-up fashion is illustrated in Figure 3. A translator that converts DCG grammar rules to Prolog programs was developed. After parsing, BUP places the resulting phrase structure on the blackboard.

1) Input to the translator

\[
\text{goal(T, Tree, [L, X]):- !.}
\]
\[
\text{dictionary(verb, [verb, [yondeimasuka]]. [yondeimasuka], X):- !.}
\]
\[
\text{noun(verb, [noun, [anata-wa]]. [anata-wa], X):- !.}
\]

2) Output from the translator

\[
\text{link(X, Y):- !.}
\]
\[
\text{noun(verb, [noun, [anata-wa]]. [anata-wa], X):- !.}
\]

Figure 3. An Example

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) [9] 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 [10]. In LFG, two bunsetsu phrases can be unified if 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).
An example of the phrase structure and c-structure for the sentence "anata-wa hon-o yondeimasuka" are illustrated in Figure 4.

In this example, the bunsetsu "yondeimasuka" contains a major verb "yon(read)" which can subcategorize bunsetsu phrases with particles such as "ga", "o", and "de". The bunsetsu "yondeimasuka" can subcategorize the bunsetsu "hon-o" because of the particle "o" in "hon-o", thus they are unified. The result of the unification is expressed as a functional-structure (f-structure).

The bunsetsu "yondeimasuka" can also subcategorize the bunsetsu "anata-wa" because the particle "wa" functions in the same way as the particle "ga" (Figure 5, a). The final f-structure is expressed as in Figure 5 b) and the case analysis is completed.

Figure 4. Phrase and C-structure for Sentence "anata-wa hon-o yondeimasuka"

Sentence: "anata-wa hon-o yondeimasuka"

Figure 5. Induction of Functional Structure

4. EVALUATION

The speech recognition system used for this study is the Model F2361A commercial speaker-dependent speech recognition system manufactured by Fujitsu Limited of Japan. It is capable of handling 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 5 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 used in Japan, and 105 simple test sentences were composed using the subset of the selected words and phrases.

The test bunsetsu phrases and sentences were recorded in a studio where no ambient noise was present. A Sony condenser microphone (C-3848D) 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 (DN3801) on Sony magnetic tapes (PLX-376G).

After the recording was finished, all the tapes were brought to Fujitsu Laboratory, and the contents were transferred to compact digital audio tapes (Technics KT-R120) using a Magra (IV-5) 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 RX-C7I). After the training was completed, each of the 105 test sentences was transmitted to the recognizer leaving a gap of 2 to 3 seconds between each bunsetsu. 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 105 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 BUP, and fourth with a depth-first-search, grouping, BUP, and case checking module.

5. RESULTS

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

A total of 365 bunsetsu phrases were spoken: 259 (71.3%) were recognized as the first candidates, and 51 were recognized as one of the 7 candidates (14%). This resulted in an 85% recognition rate. The remaining 55 bunsetsu phrases were not recognized.
The probable causes for this low recognition rate are:
1. The existence of many minimal pair bunsetsu phrases in the 4000 that were registered, and
2. The lack of sufficient voice patterns. Each of the 4000 bunsetsu phrases was registered with only one voice pattern.

A total of 105 sentences were spoken: 34 (32%) had all the bunsetsu phrases in each sentence recognized as the first candidates, and 26 (24.5%) had all the bunsetsu phrases in each sentence not recognized as the first candidate but as one of the 7 candidates. This resulted in a 57% recognition rate. The remaining 45 sentences (42.5%) 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 3.

<table>
<thead>
<tr>
<th>Entry</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>9 &lt; R &lt; 15</td>
<td>5</td>
<td>9</td>
<td>15</td>
<td>16</td>
</tr>
<tr>
<td>5 &lt; R &lt; 10</td>
<td>8</td>
<td>6</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>10 &lt; R &lt; 50</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>50 &lt; R</td>
<td>11</td>
<td>8</td>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>

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

Table 3. Performance of the Postprocessor

When the postprocessor was provided with only depth-first-search capability, 5 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 9, 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 15. This number was further increased to 18 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 21 out of the 26 sentences; the remaining 5 were unrecovered. 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.

6. CONCLUSION

In this paper, we have presented the design philosophy of a postprocessor for a large vocabulary speech recognition system for the Japanese language.

The postprocessor consists of knowledge sources that are applied to recover a spoken sentence by searching through a bunsetsu lattice. 26 sentences out of 105 test sentences were used to evaluate the sentence recovery capability of the postprocessor. The recovery rate was 80% (21/26).

The performance of postprocessor, however, must be improved if it is to be practical.

First of all, the existing grammar rules should be expanded so that they can process complex sentences. Inputting complex sentences requires an efficient search algorithm since the length of the spoken sentence is directly related to the size of the search space. Reduction of the search space can be achieved by providing a bunsetsu lattice sieve that sifts only appropriate bunsetsu phrases.

Second, the postprocessor should be able to recover sentences of which one or more bunsetsu phrases in each sentence are not recognized. Automatic recovery of these sentences is difficult, but it can be achieved to a certain extent by creating a confusion matrix [11] which keeps track of commonly confused bunsetsu phrases. The confusion matrix can be applied when all the bunsetsu phrases in a bunsetsu lattice have been searched and the correct sentence has still not been found.

Third, semantic information should be implemented because an idiosyncratic sentence like "Watashi-wa ie-o tabeteimasu" (I am eating a house) is a legitimate sentence to the current system.

Fourth, a user interface has to be carefully designed. Development of a fully automatic sentence recovery system is not only extremely difficult but also undesirable. A user should be able to select bunsetsu, correct sentences, or delete the entire sentence using a variety of input devices. A careful allocation of the sentence recovery task between users and the system is the key to successful applications such as a speech word processor.

References