INTEGRATION OF GRAMMAR AND STATISTICAL LANGUAGE CONSTRAINTS FOR PARTIAL WORD-SEQUENCE RECOGNITION

Hajime Tsukada, Hirofumi Yamamoto, Yoshinori Sagisaka

ATR Interpreting Telecommunications Research Laboratories
Tel: +81 774 95 1374, Fax: +81 774 95 1308, E-mail: tsukada@itl.atr.co.jp

ABSTRACT

This paper proposes a novel spontaneous speech recognition approach to obtain not a whole utterance but reliably recognized partial segments of an utterance to achieve robust speech understanding. Our method obtains reliably recognized partial segments of an utterance by using both grammatical and n-gram based statistical language constraints cooperatively, and uses a robust parsing technique to apply the grammatical constraints. Through an experiment, it has been confirmed that the proposed method can recognize partial segments of an utterance with a higher reliability than conventional continuous speech recognition methods using an n-gram based statistical language model.

1. INTRODUCTION

In spontaneous speech recognition, statistical language models based on n-grams are widely used. This is because such models can significantly reduce the number of recognition candidates during a search as well as accept deviated utterances. On the other hand, in many speech dialogue systems including speech translation systems, the back-end of the speech recognizer uses a grammar to analyze syntactic structures, which is usually developed independent of an n-gram-based statistical language model used in speech recognition. Because n-gram-based statistical language models and grammars work as different types of linguistic constraints, to enhance the total performance of speech dialogue systems, it is necessary that not only statistical language models but also the grammar of the back-end are used cooperatively as constraints in speech recognition.

Many methods (e.g. [11]) that integrate both a statistical language model and a tight grammatical constraint have been proposed, where only utterances that do not deviate from the grammar are acceptable. Moreover, a recognition method that uses a tight grammatical constraint approximated by a back-end grammar has been proposed [7][8]. However, using a grammar as a tight constraint in spontaneous speech recognition like these studies involves a few drawbacks. First, spontaneous speech often deviates from the grammar because of peculiar linguistic phenomena in spontaneous speech such as filled pauses, hesitations and corrections, which do not appear in read speech. Second, a tight grammatical constraint is often not robust. Although a grammar can represent long distance dependencies, local recognition errors often result in bad effects globally.

In most cases, ignoring a small amount of deviation and recognition errors does not impact on the communications using speech dialogue systems. To achieve these robust recognition, we propose a novel approach to obtain not a whole utterance but partial reliably recognized segments in an utterance by applying not only an n-gram based statistical language constraint but also a back-end grammatical constraint using a robust parsing method. In our approach, the appropriateness of a whole utterance including reliable segments is constrained by an n-gram based statistical language model, and by using robust parsing the appropriateness of each partial reliable utterance segment is constrained by the grammar used in the back-end for speech recognition.

In the following section, we explain the outline of our method. Second, we explain a method that translates a back-end grammar into an efficiently applicable representation of a grammatical constraint. Third, our robust parsing method using this grammatical constraint representation is explained. Finally, we demonstrate the validity of our method by experiment.

2. OUTLINE OF METHOD

To achieve robust speech understanding, we propose a recognition method that obtains reliably recognized partial segments of an utterance by using both grammatical and n-gram based statistical language constraints cooperatively. First, spontaneous speech is recognized by using an n-gram-based statistical language model. Next, the results are robustly parsed by a grammatical constraint. Our robust parsing method applies a grammatical constraint assuming insertions, deletions and substitutions, and outputs partial word segments removing these insertions/deletions/substitutions. Through these two steps, the obtained partial segments are not only constrained by the statistical language model but also verified by the grammatical constraints using robust parsing. As a result, the partial segments obtained by our method are more reliable than a whole utterance recognized by only using an n-gram based statistical language model.

In a typical speech dialogue system, the back-end for speech recognition uses a context-free grammar (CFG) or an attribute grammar, which is an extension of a CFG made by adding attributes. In the back-end, a CFG or its extension is necessary to analyze syntactic structures from an utterance. To reflect the constraint of this back-end grammar, our method adopts a CFG as a grammatical constraint, and uses the constraint in speech recognition

by approximating to a finite-state automaton (FSA). The approximation method will be described in Section 3. Because an approximated FSA is used, we can effectively apply the grammatical constraint with a simple algorithm. The technique to use an approximated FSA from the back-end grammar for speech recognition comes from [7] and [8], but we extend this technique with robust parsing. In our method, the grammatical constraint represented by FSA is applied assuming insertions, deletions and substitutions. This robust parsing method will be described in Section 4.

Several methods that integrate a grammatical constraint and an n-gram-based statistical language model have been proposed to achieve robust speech recognition [1][3][5][9]. Among these studies, [3] and [5] aim to enhance the recognition rates of semantically important phrases. These approaches are similar to ours in that they attach much importance to segments of utterances. Moreover, by using linguistic knowledge such as syntax or semantics, methods to robustly recognize syntactic structures [2] or semantic representations [12] have been proposed. Compared with these studies, our method has the advantage of a simpler and more robust algorithm to apply a grammatical constraint. This is because we adopt an efficiently applicable representation of a grammatical constraint approximated beforehand, and all insertions, deletions and substitutions are considered when applying the grammatical constraint. Moreover, our method does not depend on semantic representations, which have a tendency to depend on the specific task and system. As a result, our approach is portable.

3. FSA GENERATION ALGORITHM

According to formal language theories, CFGs are more powerful than FSAs. A language is defined as a set of symbol strings generated by a grammar. An automaton is defined as a machine that determines whether a given symbol string is acceptable. A language generated by a certain CFG is called a *context-free language*. Also, a language accepted by a certain FSA is called a *regular language*. Every regular language is a context-free language. However, certain context-free languages are not regular languages. Therefore, it is generally impossible to convert a CFG into an equivalent FSA that can accept the same language generated by the original CFG.

Although FSAs theoretically have less power than CFGs, in practical applications, FSAs have enough power to express grammatical constraints.[1] Moreover, FSAs have significant features from the viewpoint of parsing: (1) parsing algorithms using FSAs are simpler than similar algorithms using CFGs; and (2) every FSA can be translated into a unique deterministic and minimized FSA. The deterministic FSA has a deterministic transition from each state according to a symbol. The minimized FSA is an FSA that accepts the same language and has the minimum number of states. Because of the deterministic feature, a sentence that is ambiguous from the viewpoint of syntactic structures is acceptable through a deterministic path. Also, because of the minimum number of states, the memory space for parsing using FSAs is optimum. These features prove most advantageous for parsing with a large grammar.

Many methods to produce FSAs approximately from CFGs have been developed. Among these works, we adopted Pereira's algorithm [7]. This algorithm guarantees that all symbol strings generated by a CFG are acceptable to the produced FSA. The approximation is exact for certain CFGs generating regular languages, including all left-linear and right-linear CFGs.

SENT → NP, VP, NP.
SENT → SENT, PP.
NP → **det**, **noun**.
NP → **pron**.
NP → NP, PP.
PP → **prep**, NP.
VP → **verb**.

Figure 1: Example CFG

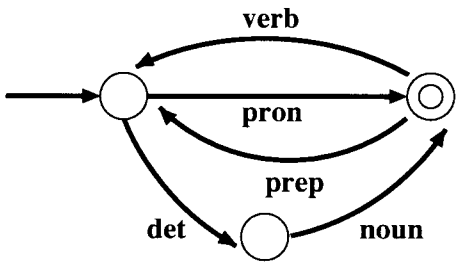

Figure 2: Approximated FSA

Figure 1 shows an example of a small English CFG, and Figure 2 shows the approximately produced FSA from the CFG by using the above mentioned algorithm, which is deterministic and minimized. In this example, the lower-case expressions stand for terminal symbols in the CFG. In the FSA, only these terminal symbols appear, and nonterminal symbols disappear. The sentence "I(pron) saw(verb) a(det) girl(noun) with(prep) a(det) telescope(noun)" is ambiguous in this CFG because the prepositional phrase "with a telescope" may modify either "a girl" or "I saw a girl". However, this sentence is accepted by a single deterministic path of the FSA in Figure 2.

4. ROBUST PARSING

The general meaning of *robust parsing* is to estimate the syntactic structures from a noisy word string. In this paper, we use *robust parsing* in a more narrow sense, i.e., obtaining acceptable paths of an FSA assuming insertions, deletions, and substitutions. To explain our algorithm we extend an FSA into a finite-state transducer (FST), where the output symbols are added to edges.

Formally, an FSA is defined as a five-tuple (Q, Σ, q_0, F, E), where Q is a finite set of states, Σ is a finite set of input symbols, $q_0 \in Q$ is the initial state, $F \subseteq Q$

[1] When an arbitrary depth of embedding is allowed in a sentence, an FSA has trouble treating *agreements* properly. However, when the depth is limited, an FSA can treat them.

is the set of final states and $E \subseteq Q \times (\Sigma \cup \{\epsilon\}) \times Q$ is the set of edges or transitions. The special symbol ϵ expresses a null transition that is allowed without reading input symbols. In contrast, an FST is defined as a slightly modified six-tuple $(Q, \Sigma, \Sigma', q_0, F, E')$, where Q is a finite set of states, Σ is a finite set of input symbols, Σ' is a finite set of output symbols, $q_0 \in Q$ is the initial state, $F \subseteq Q$ is the set of final states and $E' \subseteq Q \times (\Sigma \cup \{\epsilon\}) \times \Sigma'^* \times Q$ is the set of edges or transitions. Σ'^* is a set of strings composed of Σ' elements. An FST provides a function to translate input strings into output strings when the input strings are acceptable.

Our robust parsing method is realized by adding edges that represent insertions, deletions and substitutions to the FSA and extending it to an FST. Figure 3 shows an example FST made from the FSA in Figure 2 by adding insertion, deletion, and substitution edges. In this paper, we simply assume insertions/deletions/substitutions at any position of an input string. The left-hand side of the slash stands for an input symbol. The right-hand side of the slash stands for output symbols. Eliminating a slash means input and output symbols are the same. The edge whose input symbol is a question mark represents transitions corresponding to all input symbols in Σ. In Figure 3, insertions and substitutions are represented by edges with question marks. Deletions are represented by edges with epsilons.

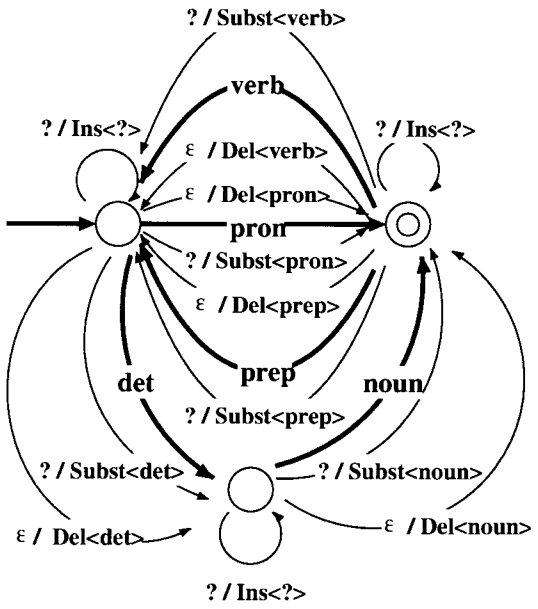

Figure 3: FST for robust parsing

For FSTs, the goal of robust parsing is to obtain the most appropriate accepting path as well as output strings. To define appropriateness, we set a penalty for each deletion/insertion/substitution edge, and the path that has the lowest number of penalties is defined as the most appropriate one. Naturally, an elegant approach would be an extension of probabilities in the FST formalism instead of penalties, but we have set out to prove the validity of a novel recognition approach by using a simpler way in this paper.

Assume that the result of speech recognition using an n-gram-based statistical language model was "hi(interj) saw(verb) girl(noun) with(prep) a(det) telescope(noun)." With the most appropriate path, which has two penalties, we can tag the words as follows: "hi(Subst⟨pron⟩) saw(verb) ϵ(Del⟨det⟩) girl(noun) with(prep) a(det) telescope(noun)." We can get reliable segments by ignoring the words marked as insertions, deletions, or substitutions. In this example, "saw(verb)," and "girl(noun) with(prep) a(det) telescope(noun)" are the reliable segments.

5. EXPERIMENT TO COMPARE RELIABILITIES

5.1. Experimental Purpose and Conditions

To clarify the effectiveness of our parital word segment recognition approach that is outlined in Section 2, we compared the reliabilities of top-best words recognized by using only a statistical language model based on n-grams with word segments obtained by robust parsing of the top-best words.

For speech-recognition experiments, we used a speaker-independent speech recognition system based on word-graphs [10]. For the speech-recognition task, we used 55 hotel-reservation dialogues included in the ATR spontaneous speech database [6]. In that database, the dialogues are bilingual and speakers talk to each other through an interpreter. For recognition experiments, we only used 1,535 Japanese utterances, which contained 22,695 words. Also, we used a context-free grammar [11] developed for speech recognition. The grammar consisted of 1,832 rules, and its unit was not a whole sentence but a segment that could be paused. The grammar was developed using nine dialogues from among the 55 dialogues used for the recognition experiments. As for the n-gram-based statistical language model, we used a variable-order n-gram [4] composed of 98 dialogues, which consisted of 1,132 words and included the 55 dialogues.

5.2. Measure of Reliability

To evaluate the reliability of word segments, we use the *relevance rate*, which is used in the research field of *information retrieval*. The relevance rate is defined as:

$$\text{Relevance Rate} = \frac{\text{Matching Words}}{\text{Recognized Words}} \times 100$$

The number of *matching words* is the maximum number of correspondences between recognized and correct words. Compared to a common recognition rate, the denominator is different. If the denominator were the number of correct words, the rate would be a normal recognition rate.

5.3. Experimental Results

5.3.1. Reliability

The relevance rate of the top-best recognition results using the variable-order n-gram was 68%. By contrast, the rate of robustly parsed partial segments was 73%. This experiment showed that we can obtain reliable segments of an utterance by using our proposed speech recognition method.

5.3.2. *Reliability and Coverage*

To achieve robust speech understanding with the proposed recognition method, not only the reliability of the obtained partial utterance segments but also the coverage of correct words by robust parsing must be enhanced.

$$\text{Correct Word Coverage} = \frac{\text{Output Correct Words of Robust Parsing}}{\text{Input Correct Words of Robust Parsing}} \times 100$$

In general, however, there is a trade-off between the reliability of the obtained segments and the correct word coverage from the flexibility inherent in robust parsing when a grammatical constraint is given. To investigate this trade-off relationship, we studied *tight* robust parsing.

Tight robust parsing is achieved by ignoring the neighboring words of insertions, deletions and substitutions. This is because the neighboring words are thought to be unreliable since they are affected by the insertions, deletions and substitutions. With this tight parsing method, for example, we can obtain the segment "with(prep) a(det) telescope(noun)" from "hi(Subst⟨pron⟩) saw(verb) ϵ(Del⟨det⟩) girl(noun) with(prep) a(det) telescope(noun)" for the example from Section 4.

Figure 4 shows details of words that were rejected by grammatical constraints. In this figure, in contrast to tight robust parsing, the original robust parsing method in Section 4 is described as a *loose* method. Using the tight parsing method, we can improve the reliability from 73% to 81%. On the contrary, the correct word coverage decreases from $89\% \simeq (47\% + 14\%)/68\%$ to $69\% \simeq 47\%/68\%$.

In the grammar we now use, one problem is that the lexical items are insufficient to handle the recognition task. An out-of-vocabulary word for a grammar is always regarded as an insertion or substitution by our robust parsing method. Reflecting this lack of vocabulary, as well as the appearance of an ungrammatical utterance in the recognition task, the coverage of the task by our grammar is 89% with a loose robust parsing method and 71% with a tight one. These rates are nearly equal to the above mentioned correct word coverages of 89% and 69%, respectively. Therefore, we should be able to better cover correct words if the lexical items in the grammar are sufficient.

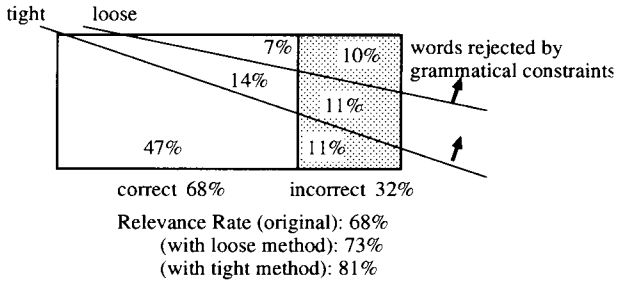

Figure 4: Rejected words by robust parsing

6. CONCLUSION

To achieve robust speech dialogue systems, we proposed a recognition method to obtain reliably recognized partial segments of an utterance by robustly paring the result recognized by using an n-gram based statistical language model. Our method uses an efficiently applicable representation of a grammatical constraint approximated by a CFG in robust parsing. Using the back-end grammar as a grammatical constraint we can enhance the total performance of speech understanding. Through an experiment on spontaneous speech recognition, we showed that our method can obtain partial utterance segments with higher reliability than conventional continuous speech recognition using an n-gram based statistical language model.

Our recognition approach can be applied also to multiple-pass search methods for robust recognition. Such methods use the information of reliable segments after their first pass. For speech recognition tasks that involve out-of-vocabulary words in particular, this type of search method should be indispensable.

ACKNOWLEDGMENT

We are grateful to T. Takezawa et al., who provided us with their subtree grammar. This grammar was necessary for our experiments.

REFERENCES

[1] W. Eckert, F. Gallwitz, and H. Niemann, Combining stochastic and linguistic language models for recognition of spontaneous speech, In Proc. of ICASSP, 1996

[2] A. Lavie, GLR*: A robust parser for spontaneously spoken language, Proceedings of ESSLLI-96 Workshop on Robust Parsing, 1996

[3] H. Lloyd-Thomas, J. H. Wright, and G. J. F. Jones, An integrated grammar/bigram language model using path scores, In Proc. of ICASSP, 1995

[4] H. Masataki and Y. Sagisaka, Variable-order n-gram generation by word-class splitting and consecutive word grouping, In Proc. of ICASSP, 1996

[5] M. Meteer and J. R. Rohlicek, Statistical language modeling combining n-gram and context-free grammars, In Proc. of ICASSP, 1993

[6] T. Morimoto et al., Speech and language database for speech translation research, In Proc. of ICSLP, 1994

[7] F. C. N. Pereira and R. N. Wright, Finite-state approximation of phrase-structure grammars, In 29th Annual Meeting of the Association for Computational Linguistics, pp. 246–255, 1991

[8] D. B. Roe, et al., A spoken language translator for restricted-domain context-free languages, Speech Communication, Vol. 11, pp. 311–319, 1992

[9] S. Seneff, H. Meng, and V. Zue, Language modeling for recognition and understanding using layered bigrams, In Proc. of ICSLP, 1992

[10] T. Shimizu, H. Yamamoto, H. Masataki, S. Matsunaga and Y. Sagisaka, Spontaneous dialogue speech recognition using cross-word context constrained word graphs, In Proc. of ICASSP, 1996

[11] T. Takezawa and T. Morimoto, Dialogue speech recognition method using syntactic rules based on subtrees and pretermianl bigrams, Transactions of the Institute of Electronics and Communication Engineers of Japan, Vol. J79-D-II, No. 12, pp. 2078–2085, 1996

[12] W. Ward, Understanding spontaneous speech: the Phoenix system, In Proc. of ICASSP, 1991