GRAMMATICAL WORD GRAPH RE-GENERATION FOR SPONTANEOUS SPEECH RECOGNITION

Hajime Tsukada, Hirofumi Yamamoto, Toshiyuki Takezawa and Yoshinori Sagisaka

ATR Interpreting Telecommunications Research Laboratories
2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0288 Japan
E-mail: tsukada@itl.atr.co.jp

ABSTRACT

We propose a novel recognition method for generating an accurate grammatical word-graph allowing grammatical deviations. Our method uses both an n-gram and a grammar-based statistical language model and aligns utterances with the grammar by adding deviation information during the search process. Our experiments confirm that the word-graph obtained by our proposed method is superior to the one obtained by only using the n-gram with the same word-graph density. In addition, our recognition method can search enormous hypotheses more efficiently than the conventional word-graph based search method.

1. INTRODUCTION

A lot of speech recognition systems [7][9] adopt multi-pass search methods mediated by effective representations for enormous hypotheses of continuous speech recognition such as word-graphs. Also, the conventional speech understanding process of speech dialogue systems, including speech translation systems, can be considered as a similar multi-pass search process where speech recognition components generate word-graphs, and natural language understanding (NLU) components generate understanding-level hypotheses from the word-graphs.

If we consider the multi-pass modeling of the speech understanding process, we find that a different language model constraint is applied for each pass. Statistical language models based on n-grams are widely used in the speech recognition component, because these models can significantly reduce the number of recognition hypotheses during a search as well as accept utterances that deviate from conventional written grammars. On the other hand, the NLU component uses a grammar to analyze syntactic structures as well as semantic knowledge, and the grammar and semantic knowledge are usually developed independently from an n-gram used by the speech recognition component. Since the both language models in these two components work as different types of linguistic constraints, the understanding-level hypotheses are constrained by both language models.

However, the direct connection of the conventional speech recognition and NLU components often is not robust because of the grammar in the NLU components. The conventional NLU components often reject whole utterance hypotheses because of two reasons. One reason is because the utterance hypotheses are often ungrammatical with the result that a small number of words are misrecognized in the speech recognition components, even when the input utterance is grammatically correct. The other reason is because spontaneous speech is often poorly modeled by conventional grammars, which tend to be based on the written form of a language and do not adequately deal with certain linguistic phenomena that frequently occur in spontaneous speech, such as filled pauses, hesitation and correction. To overcome this problem, a lot of speech dialogue systems use robust parsers [3][8][13][12] that align the recognition hypotheses with their grammar and considers deviated parts from the grammar.

However, these robust parsing methods have trouble dealing with a large amount of recognition hypotheses as their inputs due to the high computational cost. As a result, most of the enormous hypotheses generated by the speech recognition component are discarded in the stage of the robust parsing. We propose a recognition method where a finite-state machine is used for an robust parser to overcome this problem. Our robust parsing method can deals with enormous hypotheses represented by word-graphs as the input and generate a grammatical word-graph from input utterances even when the utterances are ungrammatical. The method proposed here is an extension that we proposed in 1997 [11] where we assumed n-best style utterance hypotheses. Experiments will show that our robust parsing method can recover recognition errors of word-graphs.

2. PROPOSED METHOD

2.1. Formulation

A speech recognition problem is conventionally formulated to obtain the most probable word sequence W that maximizes $P(W|O)$, where O is the input speech as shown in (1).

$$\begin{aligned} & \underset{W}{argmax}\ \ P(W|O) \\ = & \underset{W}{argmax}\ \ P(O|W)P(W). \end{aligned} \quad (1)$$

In contrast, we formulate a speech recognition problem to obtain the tagged word sequence W_T as well as W as follows.

$$\underset{W_T,W}{argmax}\ \ P(W_T, W|O)$$

$$= \underset{W_T,W}{argmax} \;\; P(O|W_T,W)P(W_T,W). \qquad (2)$$

Furthermore, we can rewrite (2) assuming that O is independent of W_T.

$$\begin{aligned} &\approx \underset{W_T,W}{argmax} \;\; P(O|W)P(W_T,W) \\ &= \underset{W_T,W}{argmax} \;\; P(O|W)P(W)P(W_T|W). \qquad (3) \end{aligned}$$

Comparing (3) to the conventional formula (1), $P(W_T|W)$ is added, which denotes the likelihood of tags for W.

In the proposed method, we use a conventional n-gram based model for $P(W)$, and use a grammatical model which considers deviations for $P(W_T|W)$. We use a grammar that generates part-of-speech sequences for the grammatical model. We use the tags denoting part-of-speech as well as grammatical deviations, namely insertions, deletions or substitutions. For example, suppose that the input utterance is "I(pron) saw(verb) a(det) girl(noun) with(prep) a(det) telescope(noun)," and that $P(W_T|W)$ is modeled by a grammar that generates the input utterance. We may obtain the most probable W and W_T as follows because of the local recognition error.

W: hi saw girl with a telescope

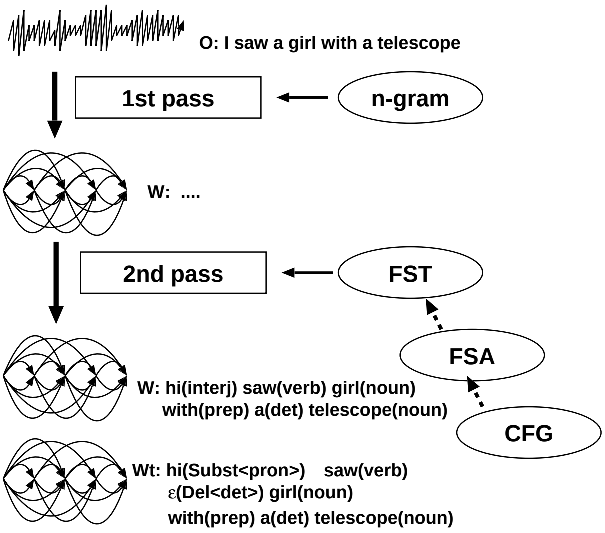
W_T: hi(Subst⟨pron⟩) saw(verb) ϵ(Del⟨det⟩) girl(noun)
with(prep) a(det) telescope(noun)

We can consider the obtained W_T to be mapped from W so that the target will be aligned to a grammar. When using the part-of-speech grammatical model, the grammatical parts are tagged by the part-of-speech, and the deviated parts are tagged by estimated part-of-speech with deviation type markers such as Ins, Del and Subst. If $P(W_T|W)$ properly models the grammatical deviations caused by recognition errors as well as other factors, W obtained by (3) can be more accurate than that obtained by (1).

Furthermore, we have another interpretation of W_T. If the underlying grammar of $P(W_T|W)$ properly models the input utterance and $P(W_T|W)$ properly models grammatical deviations caused by recognition errors, W_T can be considered as a recovered word sequence from W. Since the method here use part-of-speech grammar, the recovered parts can possibly be adjusted to the correct part-of-speech. The following experiments will show the effectiveness of recovering errors by W_T.

2.2. Multi-Pass Search Strategy

Figure 1 shows the outline of our method. We adopt a multi-pass search strategy to obtain the probable hypotheses of W and W_T in equation (3). In the first pass, a word-graph that represents W hypotheses are obtained by using $P(O|W)P(W)$, namely by the conventional manner based on an n-gram. In the second pass, a W word-graph is re-scored with $P(O|W)P(W)P(W_T|W)$, and a W_T word-graph is simultaneously generated using a grammar based model.



O: I saw a girl with a telescope
1st pass
n-gram
W: ....
2nd pass
FST
FSA
CFG
W: hi(interj) saw(verb) girl(noun) with(prep) a(det) telescope(noun)
Wt: hi(Subst<pron>) saw(verb) ε(Del<det>) girl(noun) with(prep) a(det) telescope(noun)

Figure 1: Method Outline

2.3. Grammatical Deviation Model

For efficiency, we adopt a finite-state machine as the representation of the grammar in the second pass. After the method we have proposed [11], we use a finite-state transducer (FST) to mark the deviated parts from the grammar, and the FST is an extension of a finite-state automaton (FSA), where output symbols are added. First, the underlying grammar of $P(W_T|W)$ is described by hand as a context-free grammar. Next, an FSA approximating the CFG is automatically generated. Finally, an FST is constructed by adding transitions that represent deviations from the underlying grammar such as insertions, deletions and substitutions.

We use the Poisson distribution shown in equation (4) to model deviations from a grammar. This model assumes that each type of grammatical deviations is independent and that the duration of each utterance is almost equal. Although these assumptions are rather rough, the modeling is adequate to show the effectiveness of our proposed recognition method.

$$P(W_T|W) = \frac{\lambda_I^{k_I}\lambda_D^{k_D}\lambda_S^{k_S}}{k_I!k_D!k_S!}e^{-\lambda_I-\lambda_D-\lambda_S} \qquad (4)$$

$\lambda_I, \lambda_D, \lambda_S$: the average number of insertions, deletions and substitutions in an utterance,

k_I, k_D, k_S : the number of insertions, deletions and substitutions in an utterance

2.4. Word-Graph Re-Generation

The word-graph re-generation process in the second pass can be actualized by integrating an FSA intersection algorithm with an A^*

search. A word-graph obtained in the first pass can be seen as a kind of an FSA if the score is ignored. Similarly, an FST that represents a deviation model from a grammar is also an FSA if the output symbols are ignored. As the intersection of these two FSAs, the W word-graph can be generated in the second pass. In the same way, the W_T word-graph can be generated as the intersection of FSAs except that the symbols of its transitions correspond to the output symbols of the FST, not the input symbols. We can prune the intersection word-graphs by an A^* search considering scores while generating them.

3. EXPERIMENTS

3.1. Experimental Purpose and Measures

As mentioned in 2.1, our method can recover the recognition errors of a conventional recognition method. To confirm this advantage, we compare the accuracy of W obtained by the conventional method based only on the n-gram with that of W_T obtained by our method. We use the *network word accuracy* for the accuracy measure. The network word accuray is the highest *word accuracy* of all word-graph paths, where the word accuracy is defined as (5).

$$word\, accuracy \;=\; \frac{N-(I+D+S)}{N} \times 100 \qquad (5)$$

N: Number of correct words

I: Number of insertion errors

D: Number of deletion errors

S: Number of substitution errors

There are non-word parts for the W_T word-graph because the error parts are recovered as estimated part-of-speech. The W_T *word accuracy* is obtained by only considering if the part-of-speech is correct for the part-of-speech parts. In order to evaluate fairly, we also generate a typical word-graph by extending part-of-speech to words in a W_T word-graph and comparing the network word accuracies. The substituted parts are extended to the words that have the most similar pronunciations, and the deleted parts are extended to the words that have the shortest pronunciations.

Since the network word accuracy depends on the size of the word-graph, we should take its size into account when comparing the network word accuracies. We use a *graph density* defined in (6) for the measure of the word-graph size.

$$word\; graph\; density \;=\; \frac{W}{N} \qquad (6)$$

W: Number of words in the word graph

N: Number of correct words

3.2. Experimental Conditions

We used 428 utterances from Japanese hotel-reservation dialogues in the ATR spontaneous speech database [5][6] for a test set. The database consists of human-to-human roleplays. One person is a customer and the other is a clerk in the dialogue. Utterances are spontaneous and overlap between utterances is prohibited.

We use a part-of-speech based FSA that has 263 states and 4,963 edges for the underlying grammar of the deviation model. The FSA is approximately generated from a CFG [10] developed for speech recognition. All utterances in the test set, i.e., the pronounced word sequences are acceptable to the FSA. Disfluencies are outside of the scope of the grammar. In other words, test-set utterances were selected with the condition that the utterance is acceptable to the FSA.

We used a variable-order n-gram [4] for the n-gram based language model trained from 32,074 utterances in the ATR spontaneous speech database, which did not overlap with the test utterances. The vocabulary size of the recognition dictionary was 7,219 words, and the test set did not contain any out-of-vocabulary words.

The parameters of the grammatical deviation model were also trained by the recognition results of the training utterances. In this experiment, we intended to model all recognition errors as grammatical deviations. Therefore, we could obtain the parameters by just counting insertions, deletions and substitutions in an alignment between references and hypotheses obtained by only using the n-gram.

As described in 2.2, our method re-scores the word-graph with $P(O|W)P(W)P(W_T|W)$ in the second path. Since we only consider network word accuracy in this experiment, we do not take re-scoring into account in order to avoid too much complexity. We only use the $P(W_T|W)$ score to limit the search space in the second pass. As a result, only one part consisting of continuous deviations is allowed in the W_T generated by the second pass.

3.3. Experimental Results

For 398 of 428 test utterances, we could obtain the recovered word-graph in the second pass. In other words, the word-graph of the rest of the utterances obtained in the first pass has more than one part consisting of continuous deviations.

Figure 2 shows the network accuracy of the 398 utterance word-graphs obtained in the first and second passes corresponding to the word-graph density. We define the 398 utterance as set0 and 156 utterances in set0 as set1, whose top hypotheses of the first pass deviate from the grammar. The upper dots are for set0 and the lower dots are for set1 in Figure 2. "◇" denotes a word-graph obtained in the first pass. "□" denotes a W_T word-graphs obtained from the highest density word-graph of the first pass. "×" denotes a word-graph extended from the W_T word-graph. "+" denotes the pruned word-graph by the $P(W_T|W)$ score from the highest density word-graph of the first pass.

In both set0 and set1, "+" is on the line of "◇" dots. This indicates that the pruning with the $P(W_T|W)$ score is adequate. Also, "□" and "×" is higher than the prolongation of the "◇" dots line. This indicates that the second pass improves the word-graphs obtained in the first pass. Improvement in the second pass is more drastic in

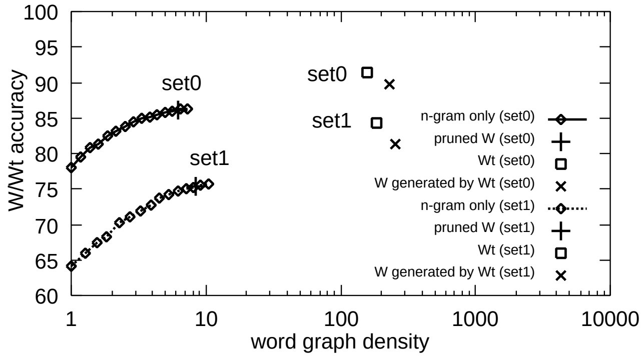


Figure 2: Experimental Results

set1, probably because there are more deviations from a grammar in the first pass word-graph.

Efficiency is another advantage of our method. Usually, the computational cost of a search linearly increases according to the word-graph density. Although it takes 2,012 sec. in CPU time for the first pass in set0, it takes only 1,790 sec. for the second pass and the part-of-speech extension from the W_T word-graph. The word-graph density of W_T is 10 times more than that of the word-graph obtained by the first pass. The size of the word-graph obtained by our proposed method is so large that the conventional search method [9] cannot obtain such a high density word-graph.

4. DISCUSSION

The current experiment gives less consideration to the score of each hypothesis. We can probably give a score to the re-generated word-graph and prune it to an adequate density in future work. In addition, we should improve the top word accuracy as well as the network word accuracy.

Our proposed method still has room for improvement. Our method gives less consideration to pronunciations and word length for modeling the grammatical deviations due to recognition errors. For example, Kaki et al. [2] consider character co-occurrence that implicitly reflect pronunciations, and Ishikawa et al. [1] consider pronunciations for recovering errors. Although these two methods cannot recover enormous hypotheses so far, the constraints used for recovering errors are useful.

5. CONCLUSION

A novel recognition method that uses both an n-gram and a grammatical deviation model has been proposed. Our proposed method can generate grammatical word-graphs that are more accurate than those obtained by the conventional method. In addition, our method can search huge hypotheses more efficiently than the conventional search method.

6. ACKNOWLEDGMENTS

We thank all the people who supported our experiment. H. Masataki et al. in Dept. 1 provided the variable-order n-gram language model, H. Singer et al. in Dept. 1 provided the acoustic model, J. Etoh constructed the CFG and T. Amakasu performed the CFG approximation.

7. REFERENCES

1. K. Ishikawa, E. Sumita, H. Iida, “Example-based error recovery method for speech translation: repairing sub-trees according to the semantic distance,” in this conference.
2. S. Kaki, E. Sumita and H. Iida, “A method for correcting errors in speech recognition using the statistical features of character co-occurrence,” Proc. COLING-ACL'98, 1998.
3. A. Lavie, “GLR*: A robust grammar-focused parser for spontaneously spoken language,” PhD thesis (Technical Report CMU-CS-96-126), Carnegie Mellon University, 1996.
4. H. Masataki and Y. Sagisaka, “Variable-order n-gram generation by word-class splitting and consecutive word grouping,” ICASSP'96, Vol. 1, pp. 188–191, 1996.
5. T. Morimoto, N. Uratani, T. Takezawa, O. Furuse, Y. Sobashima, H. Iida, A. Nakamura, Y. Sagisaka, H. Higuchi and Y. Ymazaki, “Speech and language database for speech translation research,” ICSLP'94, Vol. 4, pp.1791-1794, 1994.
6. A. Nakamura, S. Matsunaga, T. Shimizu, M. Tonomura and Y. Sagisaka, “Japanese speech databases for robust speech recognition,” ICSLP'96, Vol. 4, pp.2199–2202, 1996.
7. M. Oerder and H. Ney, “Word Graphs: an efficient interface between continuous-speech recognition and lauguage understanding,” Proc. ICASSP'93, Vol 2, p.119–122, 1993.
8. S. Seneff, H. Meng and V. Zue, “Language modeling for recognition and understanding using layered bigrams,” ICSLP'92, Vol. 1, pp. 317–320, 1992.
9. T. Shimizu, H. Yamamoto, H. Masataki, S. Matsunaga and Y. Sagisaka, “Spontaneous dialogue speech recognition using cross-word context constrained word graphs,” Proc. ICASSP'96, Vol. 1, pp. 145–148, 1996.
10. T. Takezawa and T. Morimoto, “Dialogue speech recognition method using syntactic rules based on subtrees and preterminal bigrams,” Systems and Computers in Japan, Vol. 28, No. 5, pp. 22–32, 1997.
11. H. Tsukada, H. Yamamoto, Y. Sagisaka, “Integration of grammar and statistical language constraints for partial word-sequence recognition,” Eurospeech'97, 1997.
12. Y. Wakita, J. Kawaki and H. Iida, “Correct part extraction from speech recognition results using semantic distance calculation, and its application to speech translation,” Proc. of Spoken Language Translation Workshop in conjunction with ACL'97/EACL'97, pp. 24–31, 1997.
13. W. Ward, “Understanding spontaneous speech: the Phoenix system,” ICASSP'91, Vol. 1, pp. 365–367, 1991.