SPEECH RECOGNITION OF A NAMED ENTITY

Tatsuhiko Tomita [a)], Yoshiyuki Okimoto [b)], Hirofumi Yamamoto [c)], Yoshinori Sagisaka [d)]

a, c, d) GITI, Waseda University, 3-14-9 Okubo, Shinjuku-ku, Tokyo 169-0072, Japan

b) Advanced Technology Research Laboratories, Matsushita Electric Industrial Co., Ltd.

3-4 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0237, Japan

c) ATR Spoken Language Translation Research Laboratories

2-2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0288, Japan

a) 61403675@suou.waseda.jp, b) okimoto.yosh@jp.panasonic.com, c) yama@slt.atr.co.jp,

d)sagisaka@giti.waseda.ac.jp

ABSTRACT

A hierarchical language model is newly applied to identify a named entity consisting of multiple word sequences for continuous speech recognition. By redesigning an out-of-vocabulary model of a single word using phonotactic constraints for a named entity, a hierarchical model is composed harmoniously with conventional word and word-class N-grams. Continuous speech recognition experiments aiming at movie-title identification showed the effectiveness of this modeling in the task of inquiries on these titles. These results ensure that the proposed hierarchical language modeling architecture is applicable to multiple word successions for speech recognition to cope with unregistered expressions and enables the mix use of different statistics harmoniously.

1. INTRODUCTION

For the application of conventional statistical speech recognition frameworks to the real world, unregistered expressions are one of the most serious problems. Here, an unregistered expression corresponds not only to a single unregistered word but also to a named entity consisting of multiple words including unregistered words. Though out-of-vocabulary (OOV) problems have been recognized as an important problem in speech recognition, as its name implies, mostly they have been regarded as burdens to carry out the recognition of a target task. As informative expressions such as movie-titles, book names and product names usually consist of multiple word sequences, they are not well modeled in conventional plain language models. Sometimes these unregistered expressions themselves play crucial roles in tasks where novel expressions are inevitable such as speech information retrieval or inquiries on new information by speech. It is useful to identify these unregistered expressions even if we could not perfectly identify their phonetic expressions. If some of key words contained in speech are correctly detected, they are useful for the following process.

There exist a few attempts to correctly identify OOV single words by using statistical models specially designed for target OOV word classes [1][2][3][4]. We have proposed a hierarchical language model to accommodate inter-word statistics and intra-word statistics in one statistical model [1][2]. These studies showed the possibilities of merging different statistics in different levels, i.e. words and phones harmoniously in a conventional N-gram framework. Though the usefulness of this hierarchical language model has been confirmed in a single-word level, it is not sure if we can expand this model to the above mentioned general unregistered expressions. As unregistered expressions consist of not only single words but also multiple words which give another word level statistics, it is not certain that the hierarchical language model is still effective for merging the same level (word level) different statistics.

In this paper, we apply our hierarchical model to unregistered expressions consisting of multiple words. In the following Section 2, a hierarchical model is introduced for unregistered expressions by expanding our model for single OOVs. In Section 3, experimental results are shown on continuous speech recognition to identify unregistered movie-titles. Finally, we summarize the results and discuss further modeling in Section 4.

2. A HIERARCHICAL STATISTICAL LANGUAGE MODEL FOR UNREGISTERED EXPRESSIONS

As shown in Figure1, we generalize a hierarchical statistical language model[1][2] used for single OOV words to unregistered expression consisting of multiple words. As shown in the figure, this model has multiple layers to express linguistic constraints representing different statistics. In our original hierarchical model, word-level linguistic constraints between words are represented by the conventional word and word-class N-grams and phoneme-level phonotactic constraints by the sub-word models of corresponding OOV word classes . As shown in Figure1), these two constraints are expressed as the upper-layer model and the lower-layer one respectively. In this formulation, $P(MOV)$ the probability of the unregistered movie-title "godzilla no gyakushuu"(the revenge of godzilla) can be represented by the following:

$$P(MOV) = P(C_{MOV}|C_{MOV-1})P(R|C_{MOV})$$

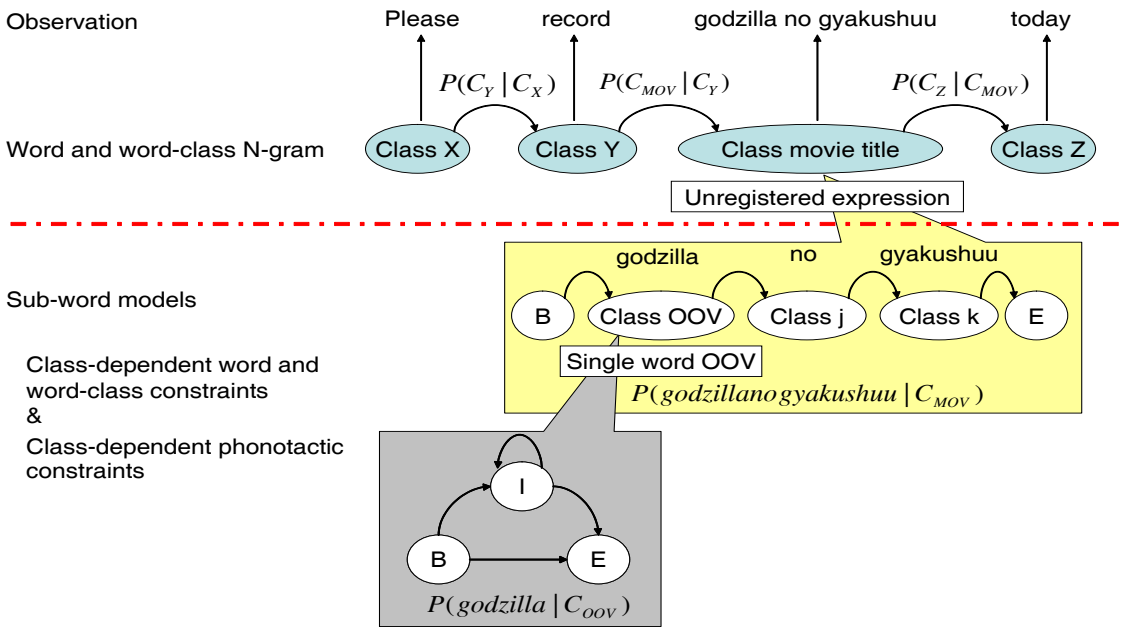

Figure 1: Hierarchical language model for unregistered expression

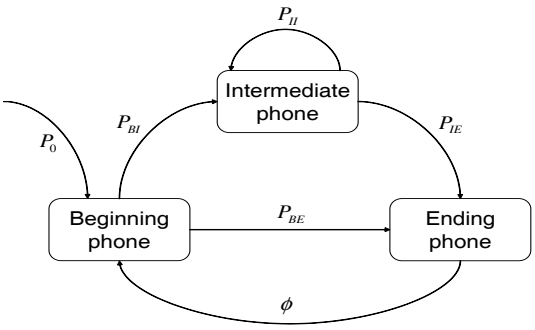

Figure 2: Word-structure model using mora as a unit

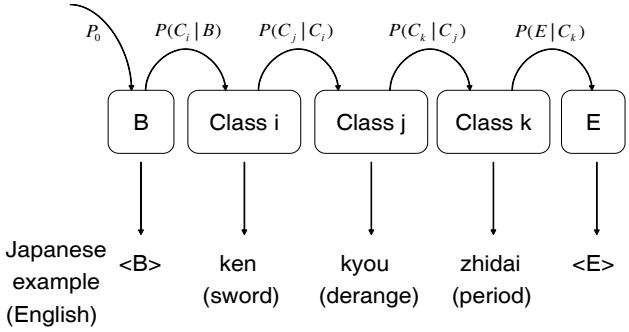

Figure 3: Unregistered expression model using word as a unit

where a word class is represented by C_i (i=1,$\cdots$,N) and R stands for a word sequence corresponding to the current word "godzilla no gyakushuu" belonging to the word class C_{MOV}. The corresponding models which give these values are shown in Figure 1)

For a single word, it has been confirmed that a simple word-structure model shown in Figure 2 works for a specific word class such as personal names and city names in Japanese speech recognition[2]. In this modeling, not only single-phone units but also variable length multi-phone sequences were employed to reflect finer phone statistics $P(R|C_{OOV})$.

As we are taking care of unregistered expressions consisting of multiple words, we generalized this word-structure model to an unregistered expression model as shown in Figure3. In this model, instead of using phone-level statistical phonotactic constraints within a word, word transitions in unregistered expressions are employed. As named entities such as movie or book titles, it is likely that we can get some statistics on their constituent words but it may be difficult to up date all title expressions as recognition entries.

3. EXPERIMENTS

3.1. A named entity identification task and speech data

For a task of speech recognition and named entity identification, we chose a task of inquiries on movie-titles. In the experiments, all movie-titles were not given as individual entries but treated as unregistered expressions only by providing constituent word entries. Though we can even assume OOV words in unregistered expressions by embedding single OOV word models to an unregistered expression model as shown at the bottom in Figure 1), as a first step, we decided to measure the performance of the proposed unregistered expression model by providing all constituent word entries.

We have recorded 1400 utterances on movie-title inquiries consisting of 100 sentences uttered by fourteen subjects (7 males and 7 females). These 100 sentences were selected separately from 3344 sentential expressions on movie-title inquiries used for language model training. As a named entity, 25143 movie-title data were used for the training of the unregistered expression model. These titles cover 22244 Japanese movie-titles and the 2899 Americans produced from 1945 to 2000.

3.2. Experiment conditions on model training and recognition

For a word and word-class language model, we adopted multi-composite class 2-grams (MCC2-grams) where not only single-word entries but also frequently used word successions and statistically derived word-classes were used as units[5]. In the experiments, MCC2-grams were trained for 970 single-word entries, 245 word successions and 800 word classes using 3344 sentential expressions on movie-title inquiries.

For an unregistered expression model, we also employed MCC2-grams. It was trained for 13846 single-word entries, 1161 word successions and 7000 word classes using 25143 movie-title data. As the experiments were conducted to evaluate the proposed hierarchical model design, we only assumed " unregistered movie entries " but no OOV words. This means that the words constituting a movie-title were all listed in the dictionary.

For comparison, we employed a word-structure model[2] shown in Figure 2 where unregistered expressions were expressed by mora unit successions. Though this word-structure model cannot express constituent word constraints so directly as the unregistered expression model, it is expected that the hierarchical structure can reflect phonetic differences between inquiry task sentences and movie-titles. The word-structure model was built based on Japanese mora units using the same 25143 movie-title training data. The model employed 482 single-unit (mora) entries, 716 mora successions, 1 beginning class, 1 ending class, and 1720 intermediate classes. Thus, replacing a movie-title class in the word and word-class MCC2-grams by either the unregistered expression model or the word-structure model, two hierarchical statistical language models were built.

We employed the following acoustic parameters, the ML-SSS tied tri-phone HMM acoustic models[6] and the decoding conditions in the first pass of two-pass search[7].

- Acoustic features
  - Sampling rate 12 kHz
  - Frame shift 10 msec
  - MFCC 12 + their delta and delta power, total 25
- Acoustic models
  - 1400-state 5-mixture HMnet model based on ML-SSS [6]
  - Automatic selection of gender dependent models
- Decoder [7]
  - frame-synchronized viterbi search word lattice output

For other search conditions such as a beam width, a language score weight, and an insertion penalty, they were decided so that word accuracy gave the best score using each language model. In the calculation of the word corrects and accuracies, a movie-title class was treated as one of word entries.

3.3. Results

The word perplexity of the upper-layer MCC2-gram (word and word-class) language model commonly used in all experiments was 15.50. The mora perplexities of lower-layer language models were calculated for open data after back-off smoothing by putting all sub-words in a test set to a sub-word (mora) unit dictionary. The mora perplexity of the unregistered expression model was turned out to be 13.16. This value is considerably lower than the mora perplexity 20.70 of the conventional word-structure model for single-word OOVs. The perplexity reduction supports the effectiveness of this structuring.

The speech recognition results are shown in Table 1 (4) Hierarchical (word) for the proposed hierarchical model using the unregistered expression model. For comparison, the recognition results of the following three models are respectively shown in Table 1(1)-(3).

(1) A non-hierarchical plain model disregarding unregistered expressions, i.e. the upper-layer MCC2-gram (word and word-class) language model only without movie-name class, denoted as "Conventional" in the table.

(2) A non-hierarchical plain model trained with all movie-titles, i.e. upper-layer MCC2-gram (word and word-class) language model trained with both inquiries on movie-titles and all movie-titles including those in the test set. As all words are treated as known words in this model, it gives the upper bound of the proposed model.

(3) A hierarchical model using the word-structure model for single-word OOVs instead of the proposed unregistered expression model, which is denoted as "Hierarchical (mora)" in the table.

As show in the table, two hierarchical models (3) and (4) gave remarkable improvements on the conventional one (1). In particular, the proposed hierarchical model (4) using word units in the lower model performed better than the other hierarchical one (3) using mora units. Though there still exist a gap between the upper bound one, the word accuracy of the proposed model reached to 66.30 % which is close to 71.50 % of the upper bound model.

As the purpose of these recognition experiments is to identify movie-titles, we calculated the named entity identification rate and phone recognition rate for the hierarchical models. As shown in Table 2, the identification correct/accuracy rates are quite high in both of these models, though phone correct/accuracy rates are not so much. (Please be aware that the phone correct/accuracy rates were calculated only for correctly identified movie-titles.) Error characteristics in movie-title identification were different

Table 1: Results of speech recognition

Language model	word correct [%]	word accuracy [%]
(1) Conventional	68.71	32.86
(2) Upper-bound	91.77	71.50
(3) Hierarchical (mora)	68.80	55.66
(4) Hierarchical (word)	72.71	66.30

Table 2: Named entity identification rate and phone recognition rate

Model	Named entity (cor/acc) [%]	Phone (cor/acc) [%]
Hierarchical(mora)	98.32/53.43	74.69/53.50
Hierarchical(word)	98.23/65.03	82.85/63.47

in these models. Deletion errors were frequently observed when a word unit model was employed, while insertion errors were conspicuous when a mora unit model was used.

In many speech recognition applications such as information retrieval or inquiries on new information where named entity identification is useful, the accuracy of constituent word detection is one of the most important indexes. We have also calculated content word accuracy and correct rates for movie-titles to roughly estimate the current performance of key word detection in the named entities. Table 3 shows the content-word detection rate of the hierarchical model using the word unit model. As Table 3 shows, the proposed model can detect 60% content words with insertions errors. Though this score is far from satisfaction, this score is encouraging as the first trial of detection where only conventional word neighboring statistics are employed. We are expecting that the identification rate will be improved with further use of statistical linguistic characteristics on word occurrences.

4. CONCLUSION

In this paper, we have presented experimental results on unregistered expression detection using a hierarchical statistical language model. A hierarchical language model was built to cope with unregistered expressions consisting of multiple word sequences. Speech recognition experiments on movie-title identification in their inquiry utterances showed the high identification rate of 98 % correct and 65 % accuracy by the proposed hierarchical language model. Though the phonetic recognition rate and content word detection rate are not still satisfactory, these experimental results are quite encouraging in two points.

First, the hierarchical structuring of a language model is effective to identify quite long segments in an utterance without being bothered by its poor recognition capability of constituent phones or words. This fact suggests another possibility of further linguistic unit detections under insufficient acoustic model performance by introducing appropriate phonetically different linguistic word sequence categories.

Next, the hierarchical model succeeded the merge of different statistics in the same linguistic level. Though our previous works showed the possibility of different statistical properties, their statistical characteristics exist in different levels, that is, phonotactic constraints and word level constraints. It was not so sure if the same level statistics can be merged without any side effects. The current study employs different statistics overlapping multiple word sequences. We have not observed any remarkable recognition rate decrease by appending two different statistics. This fact suggests the possibilities of a unified language model by integrating multiple different statistics in hierarchical fashion. A lot of unregistered expressions consisting of two or more words are sure to exist in any languages. Though we conducted recognition experiments only for Japanese, it is quite sure that the same kind of modeling can be applicable to other languages. For future steps, we would like not only improve the current modeling by introducing further linguistics constraints but also design the integration of different statistics towards task free speech recognition.

Table 3: Content word detection rate in unregistered expressions using the hierarchical model

Recognition rate[%]	word correct	word accuracy
restore rate of contents word	59.05	38.20

ACKNOWLEDGEMENT

Work supported in part by the Grant-in-Aid for Scientific Research(B)(2)-14380168, JSPS and by the Waseda University Grant for Special Research Projects 2002B-038.

REFERENCES

[1] Koichi Tanigaki, Hirofumi Yamamoto, Yoshinori Sagisaka, "A Hierachical Language Model Incorporating Class-dependent Word Models for OOV Words Recognition," Proc. ICSLP 2000, Vol.3, pp.123-126, 2000

[2] S. Oonishi, H. Yamamoto, G. Kikui, Y. Sagisaka, "A Statistical Word Model Using Word-class Specific Constraints for Handling Out-of-vocabulary Words in Speech Recognition," SNLP-Oriental COCOSDA 2002, pp. 37-42, 2002

[3] I. Bazzi, R. Glass, "Modeling Out-Of-Vocabulary Words For Robust Speech Recognition," Proc. ICSLP, Vol1. pp. 401-404, Beijing, 2000.

[4] T. Hazen, I. Bazzi, "A COMPARISON AND COMBINATION OF METHODS FOR OOV WORD DETECTION AND WORD CONFIDENCE SCORING," Proc. of ICASSP, Salt Lake City, 2001.

[5] Hirofumi Yamamoto, Shuntaro Isogai, Yoshinori Sagisaka "Multi-Class Composite N-gram Language Model for Spoken Language Processing Using Multiple Word Clusters, Proc," ACL 2001, Vol.1, pp. 531-538, 2001

[6] M. Ostendorf, H. Singer, "HMM topology design using maximum likelihood successive state splitting," Computer Speech and Language, 11(1):17–41. 1997.

[7] T. Shimizu, H. Yamamoto, H. Masataki, S. Matsunaga, Y. Sagisaka, "Spontaneous dialogue speech recognition using cross-word context constrained word graphs," Proc. ICASSP1996, pp. 17–41. 1996.