Word Class Modeling for Speech Recognition with Out-of-Task Words Using a Hierarchical Language Model

Yoshihiko OGAWA †, Hirofumi YAMAMOTO †‡, Yoshinori SAGISAKA †‡, Genichiro KIKUI ‡

†Graduate School of Global Information and Telecommunication Studies,
Waseda University Nishi-Waseda 1-3-10, Shinjuku-ku, Tokyo 169-0051 Japan
ogawa@fuji.waseda.jp, sagisaka@giti.waseda.ac.jp,

‡ATR Spoken Language Translation Research Laboratories
Hikaridai 2-2-2, Seika-cho, Soraku-gun, Kyoto 619-0288 Japan
{hirofumi.yamamoto, genichiro.kikui}@slt.atr.co.jp

Abstract

Out-of-vocabulary (OOV) problems are frequently seen when adapting a language model to another task where there are some observed word classes but few individual words, such as names, places and other proper nouns. Simple task adaptation cannot handle this problem properly. In this paper, for task dependent OOV words in the noun category, we adopt a hierarchical language model. In this modeling, the lower class model expressing word phonotactics does not require any additional task dependent corpora for training. It can be trained independent of the upper class model of conventional word class N-grams, as the proposed hierarchical model clearly separates Inter-word characteristics and Intra-word characteristics. This independent-layered training capability makes it possible to apply this model to general vocabularies and tasks in combination with conventional language model adaptation techniques. Speech recognition experiments showed a 19-point increase in word accuracy (from 54% to 73%) in the with-OOV sentences, and comparable accuracy (85%) in the without-OOV sentences, compared with a conventional adapted model. This improvement corresponds to the performance when all OOVs are ideally registered in a dictionary.

1. Introduction

As linguistic constraints for continuous speech recognition, task-specific word and word class N-grams have been widely used. For the training of N-grams, a task specific language corpus is needed, which highly restricts the portability of speech recognition technology. There are many cases when language data are not available to attain word neighboring statistics or the full list of candidate word entries. For data scarcity, language model adaptation techniques have been widely used [1][2]. Though these adaptation techniques can contribute to the improvement of recognition performance when some amount of language data is available, they cannot cope with out-of-vocabulary (OOV) words. To cope with OOV problems, OOV models have been proposed for some specific word classes [3][4].

The hierarchical language model [4][5] that we had proposed enabled the identification of OOV word classes for personal names and places without degrading the recognition performance of words in the lexicon. As the proposed language model has a hierarchical structure where Inter-word characteristics and Intra-word characteristics are modeled independently, it can be easily applied to cover more general OOV words. If this model proves to be useful for general OOV words, it offers a fundamental solution to the problem of exhaustive listing of task specific words.

In this paper, we generalize OOV words to all noun classes when adapting a language model to another task. As most task specific OOV words belong to the noun category and a training corpus for task adaptation is usually too small to cover them all, the OOV problem is quite serious. In Section 2, the structure of the hierarchical language model is explained with emphasis on word structure to reflect phonotactic constraints and its advantages. In Section 3, the experimental conditions and setup are described. The experimental results are set forth in section 4, and our conclusions are set forth in section 5.

2. Hierarchical Language Model

2.1. Model overview

Our proposed hierarchical language model consists of two layers, an Inter-word class N-gram model and an Intra-word structure model. The Inter-word class N-gram model is formalized by a class N-gram model [6] [7], and provides constraints for word to word (including OOVs) transitions. On the other hand, the Intra-word structure model is formalized by a word structure model [5], and provides phonotactic constraints for OOVs. In the proposed model, these two language models are hierarchically combined, thus ensuring supporting that all OOVs belong to some specific class. Figure 1 shows an overview of the proposed language model.

The hierarchical language model as a management measure to OOV words separates clearly and uses independently the following two constraints:

- A constraint on the morae connection characteristics that participate in word formation.
- A constraint on the word neighboring characteristics as grammar.

In this model, the morae connection characteristic is specified independently of the task for recognition. Therefore, it can be used for a different task, without any changes. Furthermore, improvement in performance can be expected for a more general vocabulary and a more general task.

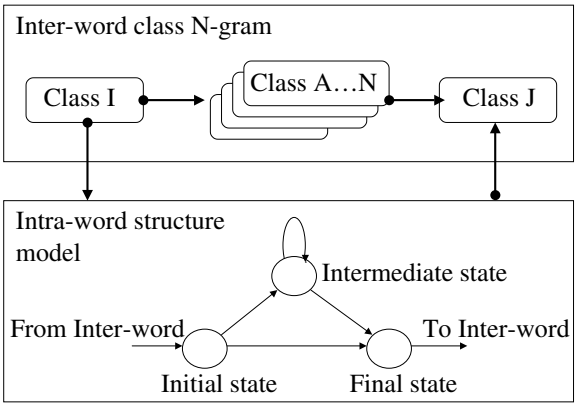

Figure 1: Experimental system configuration

2.2. Inter-word class N-gram model

A class N-gram language model is proposed to avoid the data sparseness problem. This Model is also effective for rapid adaptation [8] [9], and training on a small data set. In the class N-gram model, the transition probability to word w_i ($P(w_i)$) is calculated by the following equation:

$$P(w_i) = P(c_i|c_{N-i+1}, ..., c_{i-2}, c_{i-1}), P(w_i|c_i) \quad (1)$$

where c_i denotes the word class in which word w_i belongs. In this model, all of the OOVs belong to a single word class c_{OOV}.

2.3. Intra-word structured model

2.3.1. Constraints for morae sequence considering its length

The word structure model can be represented by a probabilistic Finite-State-Automaton with three states, beginning, intermediate and ending. For each state, morae or morae successions are assigned. This structure can represent not only morae to morae (or morae successions) transition probabilities in the OOV, but also can give the OOV's morae length constraints. The word morae length can be approximated by Gamma distribution [4]. For the OOV with a short length morae, the probability for morae successions in the beginning directly gives the value of the Gamma distribution. For the OOV with a long length morae, a self-loop transition in the intermediate state provides the Exponential distribution value. Exponential distribution $Exp(X)$ provides good approximation for the Gamma distribution $Gamma(X)$ if X is large.

2.3.2. Extraction of morae successions

For extracting morae successions, the following two issues are considered:

- Frequent morae successions must be extracted to avoid the data sparseness problem. Additionally, frequent morae successions can give a good approximation of the Gamma distribution. This means that frequent morae sequences can have a good approximation of its probabilities.
- Position dependent morae successions must be extracted. These morae successions can represent frequent morae sequences in the word head or tail, such as "re-" or "in-" in the word head and "-tion" in the tail.

2.3.3. Expand to N-gram representation

The above structure can be expanded to a N-gram representation. The occurrence probabilities of OOVs with a length-N morae sequence M, $P(OOV_M)$, can be Represented by the following N-gram style equations:

$$P(OOV_M) = P(m_B)P(m_2|m_B)\prod(P(m_i|m_{i-1}))P(m_E|m_{N-1}) \quad (2)$$

when $N > 2$, and

$$P(OOV_M) = P(m_B)P(m_E|m_B) \quad (3)$$

when $N = 2$,
where m_B and m_E represent the beginning and ending morae or morae succession, and m_i represents the i-th intermediate morae or morae succession.

2.4. Model composition

In the hierarchical language model, transition target words are both OOV and In-lexicon words. Transition probabilities to In-lexicon words can be trained according to training corpus. However, transition probabilities to the OOV cannot be obtained. Therefore, transition probabilities to In-lexicon words must be distributed to OOV words. In this model, we assume that all of the OOV words belong to the noun category. The transition probabilities to In-lexicon noun words, $P(Noun|X)$, are discounted, and the discounted probabilities are assigned to the OOV words as shown in the following equations:

$$\hat{P}(Noun|X) = (1 - \lambda)P(Noun|X) \quad (4)$$

$$P(c_{OOV}|X) = \lambda \sum_{Noun} P(Noun|X) \quad (5)$$

where $\hat{P}$ represents the transition probability after discounting.

The transition probabilities for the OOV class are obtained according to the Above equation. The occurrence probabilities of the OOV words from the OOV class are obtained according to equations (2) and (3). The total probabilities of OOV words with morae sequence M from word X are obtained according to the following equation:

$$P(OOV|X) = P(c_{OOV}|X)P(OOV_M) \quad (6)$$

3. Speech Recognition Experimental Setup

3.1. Tasks, training and evaluation sets

We have evaluated the proposed model in continuous speech recognition with language model adaptation. In the experiment, we used two sets of different corpora. The first is the Japanese travel conversation corpus, which is composed of 432,639 sentences [10]. The second is the Japanese appointment conversation corpus, which is composed of 8,020 sentences [11]. All of the sentences in the first corpus are used for the model training. The second corpus is split into 5,480 sentences for the training and 2,540 sentences for the evaluation. Two evaluation sets are created with the above 2,540 sentences. The first is 663 sentences that contain at least one OOV word (a total of 897 OOV words). The second is 626 sentences without OOV words.

3.2. Generation of hierarchical N-grams

3.2.1. Inter-word class N-gram

For creating an Inter-word class N-gram, 432,639 sentences of the travel conversation corpus are used as the adaptation source (general) data. 5,480 sentences of the appointment conversation corpus are used as the adaptation target data. These 432,639 + 5,480 sentences include 4,1732 words, and the lexicon size is also 4,1732 words. For the adaptation, occurrences in the adaptation target data are weighted by parameter α. The weighted occurrence of word (or word sequence) W is given by the following equation:

$$C(W) = C_{source}(W) + \alpha C_{target}(W) \quad (7)$$

In this experiment, α is set to 50 based on a preliminary experiment. Automatic word clustering [7] is performed using the above occurrence. The number of created word classes is 880. The class N-gram is created from these word classes and the above occurrences. The Good-Turing discount is used for smoothing.

3.2.2. Intra-word structure model

For creating the Intra-word structure model, nouns in the adaptation source data that occur less than 200 times. The reason that only words occurring less than 200 times are used is that words with a high occurrence are usually general words that can be observed in small adaptation target data. For the training and extraction of morae successions, the occurrence of each noun is considered. Special symbols are added to the first and the last morae to assign different probabilities in the beginning and ending states. For example, for the noun with the morae sequence "a b c d e", the following symbol sequence is used for the training:

<s> Ba b c d eE <e>

where, "<s>" and "<e>" are the symbols that represent the data start and end. For the no-observed morae to morae transition, The Good-Turing discount is used for smoothing. For the extraction of morae successions, we select morae sequences with a greater than 200 times occurrence with a length from 3 to 8. The total number of morae and morae successions is 155 morae and 212 morae successions for the beginning state, 159 morae and 94 morae successions for the intermediate state and 157 morae and 252 morae successions for the ending state.

3.3. Parameter in model combining

To combine the Inter-word class N-gram and Intra-word structure model, discount Parameter λ given in equations (4) and (5) must be fixed. The OOV rate in the evaluation set of 2,540 sentences is 0.022(2.2%). The total word 1-gram of nouns in the training data is 0.18 (18%). Therefore, we set λ to $0.12 = 0.022/0.18$.

3.4. Decoding conditions

The decoding conditions in the first pass are as follows:

- Acoustic features
  - Sampling rate 16 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 [12]
  - Automatic selection of gender dependent models
- Decoder [13]
  - frame-synchronized viterbi search word lattice output

In the second pass, pruning for the outputted word lattice is performed. Pruning is necessary because, in the simple combining of the Inter-word class N-gram and Intra-word structure model, back-off smoothing gives non-zero probabilities of unreasonable transitions, such as from the word to the intermediate morae or ending morae. To avoid these unreasonable transitions, Finite-State-Automaton is used. In this Finite-State-Automaton, only reasonable transitions are accepted. Therefore, the intersection of the outputted word lattice and the Finite-State-Automaton includes only reasonable transition passes. The final 1-best result is selected from this intersected word lattice.

4. Speech Recognition Results

4.1. Comparison target models

We prepare the following three language models for comparison with the proposed model.

- Word 2-gram
  This is the most conventional model. The training set is the same as The Inter-word class N-gram in the proposed model, and the lexical entry is 4,1732, which is also the same as the Inter-word class N-gram.
- Class 2-gram
  This model is only an Inter-word class N-gram model. The number of classes is 880, which is the same as the proposed model.
- Upper Limit Model
  This model is the same as the above word 2-gram model without the training data and lexicon. For the training of this model, additional adaptation data is used. In this data, all of the OOV words in the evaluation set are observed once. Therefore, all of the OOV words have 1-gram probabilities calculated by $\alpha = 50$ occurrence. This additional adaptation data simulates the ideal adaptation data. Therefore, this model will give almost the upper limit performance in an OOV test set.

4.2. Performance comparison in proposed model

The proposed model is evaluated in comparison with the above three models in two evaluation sets, the with-OOV set and the without-OOV set. In the evaluation, two measures are used. The first measure is the conventional word accuracy and correct. In the proposed model, if the OOV words are recognized as OOV words, we consider the OOV words to be correctly recognized if its morae sequence is not correct. The word accuracy and correct of each model in each evaluation set is shown in Table 1. The second measure is the OOV and In-lexicon word recall. This measure is only used With the with-OOV set. The word recall is given by

$$Recall = R/M \quad (8)$$

where M represents the number of OOV (or In-lexicon) words and R represents the number of correctly recognized OOV (or

Table 1: Accuracy and Correct in each evaluation set

Kind of Model	Without-OOV set	With-OOV set
	Accuracy/Correct	Accuracy/Correct
Word 2-gram	85.1/88.6	52.8/73.7
Class 2-gram	84.9/88.6	54.2/73.6
Upper Limit	85.1/88.6	74.3/81.8
Propose Model	85.2/87.9	73.1/83.2

Table 2: Recall of OOV and In-lexicon words in the with-OOV set

Kind of Model	OOV	In lexicon words
Word 2-gram	-	82.2
Class 2-gram	-	82.1
Upper Limit	57.3	84.6
Propose Model	68.6	84.9

In-lexicon) words. The word recall of each model in the with-OOV evaluation set is shown in Table 2.

In the without-OOV evaluation set, word accuracy and correct for each model are almost comparable, even though some words are recognized as OOV words in the proposed model. This result means that the word structure model in the proposed model perform satisfactorily, if an OOV word is not included. Furthermore, many mis-recognized OOV words have the same or similar morae sequence as the correct words.

In the with-OOV evaluation set, each model gives quite different accuracy. The accuracy in the upper limit case is (74.3%), which is about 10 points lower than The without-OOV set. The reason for this difference is that long or complex sentences tend to include OOV words, and long or complex sentences are difficult to recognize. The accuracies in the word 2-gram and class 2-gram model are quite low (52.8% and 54.2%). These accuracies are almost 20 points lower than the upper limit model. On the other hand, the proposed model resulted in 73.1% accuracy, which is only 1.2 points lower than the upper limit model. The proposed model resulted in 1.4 points higher word correct than upper limit model, even though the accuracy is 1.2 points lower. In Japanese, almost all proper nouns (almost OOV words belong), such as company names, are compound nouns. In the proposed model, these compound noun OOVs are sometimes separately recognized as OOVs and In-lexicon words (or OOVs and OOVs). Segmentation of Japanese compound nouns is very difficult, and is sometimes carried out as a "Named entity problem" in post-processing. Therefore, the OOV detection capacity of the proposed model is comparable or higher than the upper limit model.

Next, we evaluate the word recall in the with-OOV set shown in Table 2. The word 2-gram and class 2-gram models resulted in a 2.4 points lower recall for the In-lexicon words, since the OOV words damage the neighboring words. The proposed model resulted in comparable recall for the In-lexicon words and an 11-point higher recall for the OOV words. This means that the proposed model can correctly detect OOV words without damage to neighboring In-lexicon words.

5. Summaries

In this paper, we adopted a hierarchical language model for general OOV words in the noun category. The speech recognition experiments showed the efficiency of this modeling. OOV words were identified correctly without degrading the recognition performance of other word candidates in the lexicon. This modeling seems to be quite effective, particularly for task adaptation with small data where a full listing of the task specific words is not expected. As a next step, we will improve the performance of the phonetic recognition accuracy of the OOV words by splitting an Intra-word class into multiple sub-classes where the phonotactic differences of the task specific OOVs can be modeled properly. We believe that these hierarchical language models will integrate different language constraints properly and enable task-free speech recognition using linguistically rich, multiple statistical data harmoniously.

6. References

[1] S. Matsunaga, T. Yamada and K. Shikano", "Task Adaptation In Stochastic Language Models For Continuous Speech Recognition,"*Proc. ICASSP1992*, pp. 165–168, 1992.

[2] H. Masataki, Y. Sagisaka, K. Hisaki and T. Kawahara, "Task adaptation using MAP estimation in n-gram language modeling," *Proc. ICASSP-97*, Vol.2, pp. 783–786, 1997.

[3] F. Gallwitz, E. Noeth and H. Niemann, "A Category Based Approach for Recognition of Out-of-Vocabulary Words," *Proc. ICSLP*, Vol.1, 1996.

[4] K. Tanigaki, H. Yamamoto and Y. Sagisaka, "A hierarchical language model incorporating class-dependent word models for oov words recognition,"*Proc. ICSLP2000*, VOL.3, pp. 123–126, 2000.

[5] S. Onishi, H. Yamamoto, G. Kikui and Y. Sagisaka, "A statistical word model using word-class specific constraints for handling out-of-vocabulary words in speech recognition," *Proc. SNLP-Oriental COCOSDA 2002*, pp. 37–43, 2002.

[6] S. Bai, H. Li and B. Yuan, "Building Class-based Language Models with Contextual Statistics," *Proc. ICASSP-98*, pp. 173-176, 1998.

[7] H. Yamamoto, S. Isogai and Y. Sagisaka, "Multi-class composite N-gram language model for spoken language processing using multiple word clusters," *Proc. ACL*, 2001.

[8] G. Moore and S. Young, "Class-based language model adaptation using mixture of word-class weight," *Proc. ICSLP-2000*, vol. 4, pp. 512-515, 2000.

[9] H. Yamamoto and Y. Sagisaka, "A language model adaptaion considering both topic and sentence style variance,"*Proc. ASRU*, 2001.

[10] T. Takezawa, E. Sumita, F. Sugaya, H. Yamamoto and S. Yamamoto, "Toward a broad-coverage bilingual corpus for speech translation of travel conversations in the real world," *Proc. 3rd International Conference on Language Resources and Evaluation*, Vol. I, pp. 147–152, 2002.

[11] A. Nakamura, S. Matsunaga, T. Shimizu, M. Tonomura and Y. Sagisaka, "Japanese speech database for robust speech recognition," *Proc. of ICSLP1996*, pp. 2199–2202, 1996.

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

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