TWO-STEP GENERATION OF VARIABLE-WORD-LENGTH LANGUAGE MODEL INTEGRATING LOCAL AND GLOBAL CONSTRAINTS

Shoichi Matsunaga and Shigeki Sagayama

NTT Human Interface Labs.,
1-1, Hikari-no-oka, Yokosuka-shi, Kanagawa 238 Japan

ABSTRACT

This paper proposes two-step generation of a variable-length class-based language model that integrates local and global constraints. In the first-step, an initial class set is recursively designed using local constraints. Word elements for each class are determined using Kullback divergence and total entropy. In the second step, the word classes are recursively and words are iteratively recreated, by grouping consecutive words to generate longer units and by splitting the initial classes into finer classes. These operations in the second step are carried out selectively, taking into account local and global constraints on the basis of a minimum entropy criterion. Experiments showed that the perplexity of the proposed initial class set is superior to that of the conventional part-of-speech class, and the perplexity of the variable-word-length model consequently becomes lower. Furthermore, this two-step model generation approach greatly reduces the training time.

1. INTRODUCTION

Word n-gram models are effective and widely used as language models for large vocabulary continuous speech recognition[1] and broadcast news transcription[2]. However, they represent only local constraints within a few successive words and lack the ability to capture global or long-distance dependencies between noncontiguous words[3, 4]. Efforts to make more powerful models integrating longer-distance constraints, or variable-length modeling[5, 6] has been carried out. These powerful models have a huge number of parameters to be estimated. Word-grouping or class-based modeling[7, 8] is a representative method to limit this increase in the number of parameters for a large vocabulary, and a key point in making a compact and powerful model is how to design the class set adequately. Usually for sophisticated models, optimal class setting requires a tremendous amount of time. We devised a variable-length class-based language model (indicated as one-step) and showed its superiority to the word bigram or trigram for middle-vocabulary-size spontaneous dialogue[9]; however, this model also needed much computation to generate sub-optimal classes for a large vocabulary task and uses an initial class-set based on part-of-speech (POS) information, which resulted in performance being insufficient.

To overcome these problems of the performance and the training time, we propose a multi-step modeling approach to generate a powerful language model. We devise the following two-step strategy for a finer variable-length class-based language model using local and global constraints to keep the amount of computation from exploding.

In the first step of making this model, an initial class set for the second-step is recursively designed taking account of local constraints, which have the strongest effect on the performance of the language model. Word elements for each class are determined using Kullback divergence of distribution of the nearest words and total entropy of the class-based model. In the second step, compound word sequences are iteratively recreated and a word-class set is recursively recreated, by grouping consecutive words to generate longer units and by splitting the initial classes into finer classes. The main characteristic of the second step is that these operations of grouping and splitting are carried out selectively, taking into account local and global constraints between noncontiguous words on the basis of a minimum entropy criterion. To capture the global constraints, the model takes into account the sequences of the function words and of the content words, which are expected to respectively represent the syntactic and semantic relationships between words.

This paper indicates that the perplexity of the proposed initial class set is superior to that of the conventional part-of-speech class, that the perplexity of the proposed model for the test corpus consequently becomes lower than that of the one-step model generation, and that this two-step model generation approach greatly reduces the training time.

2. INITIAL CLASS SETTING (FIRST STEP)

Suppose a sentence S consists of a word sequence $w_1, w_2, \ldots, w_N$ (indicated as w_1^N), then the probability of S is written as

$$P(S) = P(w_1, w_2, \ldots, w_n) = \prod_{i=1}^{N} P(w_i \,|\, w_1^{i-1}). \quad (1)$$

As neighboring words are the strongest factors to predict the next word occurrence, the following bigram model and its class-based approximation are adopted to design an initial class-set:

$$\begin{aligned} P(w_i \,|\, w_1^{i-1}) &\simeq P(w_i \,|\, w_{i-1}) \\ &\simeq P(w_i \,|\, C_i) P(C_i \,|\, C_{i-1}), \end{aligned} \quad (2)$$

where C indicates a word-class ($w_i \in C_i$).

In our initial class setting, word clustering using Kullback divergence, and word-class setting based on an entropy criterion, are used in the procedure of recursive word-class division to obtain a suitable initial word-class set for the second step (Figure 1). The class division algorithm is as follows:

- (1-1) Select a class C_h and its division into two classes C_i and C_j, giving the minimum total distances:

$$\arg\min_{C_h, g_i, g_j} \left(\frac{1}{K_i} \sum_{w_s \in C_i} d(g_i, w_s) + \frac{1}{K_j} \sum_{w_s \in C_j} d(g_j, w_s)\right), \quad (3)$$

where K_i is the number of words belonging to class C_i, g_i is the center word (centroid) of class C_i, and $d(x, y)$ is the distance value between words x and y. Membership of each word to each class is decided on the bases of smaller distance value: if $d(w_s, g_i) \leq d(w_s, g_j)$, then $w_s \in C_i$, else $w_s \in C_j$. Distance $d(x, y)$ is defined using both Kullback divergence $d1$ of previous words and that $d2$ of succeeding words. Namely, to calculate the difference of distribution of succeeding words w for words x and y, $d1$ is defined as

$$d1(x, y) = \sum_{w}^{V} (P(w \,|\, x) - P(w \,|\, y))(\log P(w \,|\, x) - \log P(w \,|\, y)). \quad (4)$$

Calculation of $d2$ for preceding words is the same as Eq. (4), and $d(x, y)$ is defined as follows;

$$d(x, y) = d1(x, y) + d2(x, y) \quad (5)$$

- (1-2) Divide the class C_h into C_i and C_j according into the result of (1-1), and rearrange membership of word elements of C_i and C_j based on the basis of minimum entropy criterion.

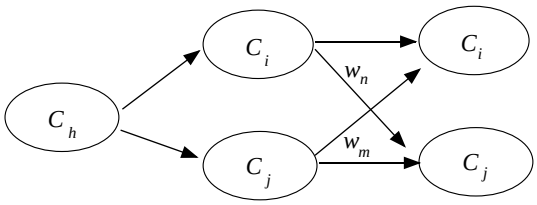


Figure 1: Initial class setting procedure

These procedure are repeatedly carried out until the number of classes reaches the pre-defined number. In our modeling, for the use of global constraints in the second step, all words in the lexicon were divided into content words such as nouns, verbs, adjectives, and adverbs, and function words such as auxiliary verbs and case markers, and then the class division procedure begins with these two classes. Then, each word class in the generated initial set is composed of either function or content words, which is a necessary condition for the second step.

3. CLASS-BASED MODELING USING GLOBAL CONSTRAINTS (SECOND STEP)

3.1. Global constraints

For simplicity, only a single preceding word is taken into account, both for global and local relationships. Let f_i denote the last function word and h_i the last content word in the substring w_1^i. Taking f_{i-1} and h_{i-1} into consideration as well as w_{i-1}, the occurrence probability of a word w_i is, represented approximately as follows[10]:

$$P(w_i \,|\, w_1^{i-1}) \simeq P(w_i \,|\, w_{i-1}, h_{i-1}, f_{i-1}). \quad (6)$$

As w_{i-1} is identical to h_{i-1} or f_{i-1},

$$P(w_i \,|\, w_{i-1}, h_{i-1}, f_{i-1}) = \begin{cases} P(w_i \,|\, w_{i-1}, f_{i-1}), \\ \quad \text{if } w_{i-1} \text{ is a content word} \\ P(w_i \,|\, w_{i-1}, h_{i-1}), \\ \quad \text{if } w_{i-1} \text{ is a function word.} \end{cases} \quad (7)$$

Figure 2 shows an example of how global constraints are taken into account for each word in the sentence. Arcs indicate global constraints in this figure. Word w_8, which is a function word, has global constraints with function word w_5, because its preceding word w_7, which has local constraints with w_8, is a content word. Thus its probability is $P(w_8 \mid w_7, w_5)$.

In our formalization, to reduce the amount of memory required and to cope with the sparseness of training data class-based modeling is introduced[9],

$$\begin{aligned} &P(w_i \,|\, w_{i-1}, h_{i-1}, f_{i-1}) \\ &\quad \simeq P(w_i \,|\, C_i) \cdot P(C_i \,|\, C_{i-1}, R_{i-1}), \end{aligned} \quad (8)$$

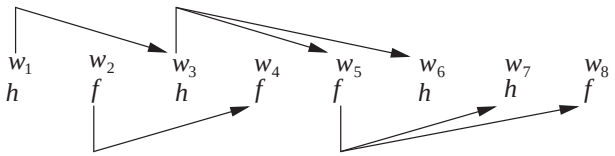
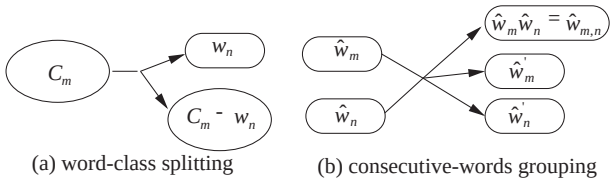


Figure 2: Examples of global constraints in a sentence



(a) word-class splitting (b) consecutive-words grouping

Figure 3: Two procedures in variable-length modeling

where C and R are word classes ($w_i \in C_i$, $R_{i-1} \ni f_{i-1}$ or h_{i-1}).

3.2. Iterative word-class setting

An appropriate set of word-classes should be designed to generate powerful variable-length models keeping the number of total parameters small. In our approach, starting from initial given classes, the following two types of procedures are carried out repeatedly (Figure 3)[6, 9]:

(a) Word-class splitting:

Split a class C_m into a word intrinsic class w_n and its complement class $C_m - w_n$.

(b) Consecutive word grouping:

Group a pair of consecutive words $\hat{w}_m$ and $\hat{w}_n$ into a concatenated word $\hat{w}_m\hat{w}_n$ ($=\hat{w}_{mn}$) and its complements $\hat{w}'_m$ and $\hat{w}'_n$.

Consequently, the probability of S is rewritten as

$$P(S) \simeq \prod_{i=1}^{K} P(\hat{w}_i|C_i) \cdot P(C_i|C_{i-1}, R_{i-1}), \qquad (9)$$

where $\hat{w}_i$ is a word sequence belonging to the word-class C_i, K is the number of word sequences contained in the sentence ($K \leq N$), and word classes C, R are sets of content words, function words, or word sequences of function and content words. When word w_{i-1} of word-class C_{i-1} is a new word generated by grouping content words and function words, $P(C_i|C_{i-1}, R_{i-1})$ is identical to $P(C_i|C_{i-1}, C_{i-2})$.

The above-mentioned procedures are carried out on the basis of a total minimum entropy criterion concerning each probability of $P(\hat{w}_i|C_i)P(C_i|C_{i-1}, R_{i-1})$ in the following way:

1. Setting of given initial word-classes

2. Calculation of total entropy for each selection

   (2-a) Select a split word $w_{n_{min}}$ and its class $C_{m_{min}}$ giving the minimum entropy (indicated as $H_{a_{min}}$) for each word split.

   (2-b) Select a pair of grouping words $\hat{w}_{p_{min}}$ and $\hat{w}_{q_{min}}$ giving the minimum entropy (indicated as $H_{b\,min}$) for each pair.

3. Selection of word-class split or word grouping

   Either split word-class for $C_{m_{min}}$, $w_{n_{min}}$ or group consecutive words for $\hat{w}_{p_{min}}$, $\hat{w}_{q_{min}}$, whichever gives the smaller entropy, $H_{a\,min}$ or $H_{b\,min}$, and go back to step 2.

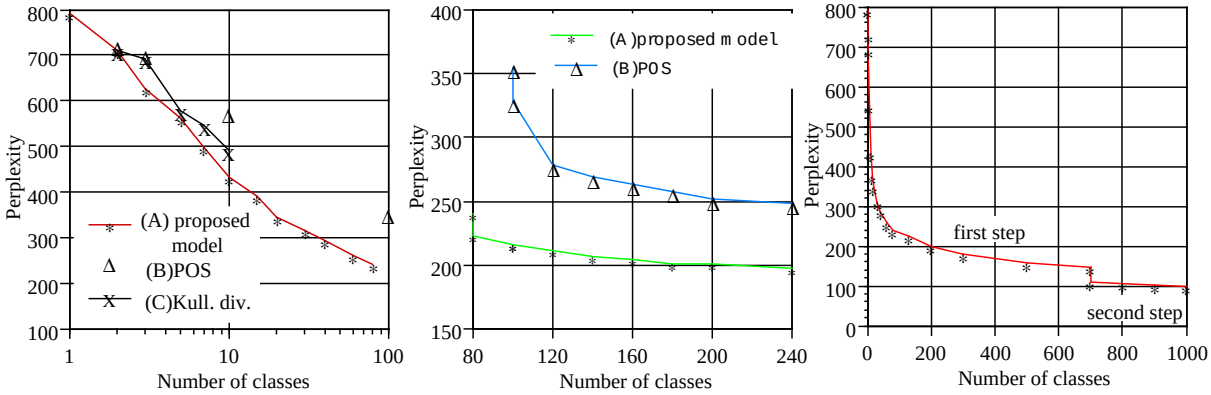
By repeating these processes, a finer variable-length class-based model taking account the global constraints is generated.

4. EXPERIMENTS

The proposed model was generated using a Japanese newspaper corpus (Nikkei). The training data consisted of about 1.4×10^5 sentences with about 2.6×10^6 words (9.2×10^3 different words, where 8% are function words and the rest are content words), and POS was tagged to each word of the lexicon. Test text consisted of 2.2×10^4 sentences (4.0×10^5 words) from another part of the newspaper corpus.

We evaluated the proposed initial model (A) using the test-set perplexity from Eq. (2). First, all words were divided into a class of content words and a class of function word, and the initial class setting in Section 2 was carried out. The proposed model was compared with a heuristic POS-based initial model (B) and a clustering-based model based on minimum distortion (C). Concerning the POS-based model, the model of ten classes is designed using POS information, and that of 100 classes is designed using both POS information and the syntactic and semantic property for each word. In the clustering-based method, class division was carried out using the distortion measure, which is described as Eq. (5) in Section 2, using Kullback divergence. Experimental results for each number of classes are shown in Figure 4-a. The proposed initial model achieved lower perplexity than the other two models (about 2/3 for the 100-class model compared to the POS model).

Next, the proposed class-based variable length model was generated for two initial set: 80 classes by the proposed initial class generation algorithm and the 100-class model based on POS information. Test-set perplexities from Eq. (8), shown in Figure 4-b, also show that the proposed initial class model is superior to the conventional POS model (about 20% reduction for the 240-class model).



(4-a) Perplexities for each initial models (4-b) Second step for each initial model (4-c) Two-step model generation

Figure 4: Reduction of perplexity for two-step approach of class-based variable-length language model

Then, we generated the proposed class-based variable-length model using the 700-class model resulting from the first step. Experimental results are shown in Table 1 and Figure 4-c. As the number of classes increased, perplexity decreased monotonically. When the number of classes exceeded 700, the perplexity of the test text became lower than that of bigram.

These results show that the two-step variable-length class-based modeling is very useful. Furthermore, this two-step model generation approach greatly reduced the training time compared to the one-step modeling (approximately 1/20).

5. CONCLUSIONS

In this paper, we proposed a fast two-step generation approach of a finer variable-length class-based language model that integrates local and global constraints. In the first-step, an initial class set is recursively designed using only local constraints. In the second step, the operations of grouping consecutive words and splitting the initial classes into finer classes were carried out selectively. Experiments showed that the perplexity of the proposed initial class set is superior to that of the conventional part-of-speech class (about 2/3), and the perplexity of the proposed model for the test corpus is lower than that of word bigram model (about 20% reduction). This two-step model generation approach greatly reduced the training time compared to the one-step modeling (approximately 1/20).

We will examine our modeling approach for larger vocabulary, and are planning to apply this language model to the second pass in a multi-pass speech recognition system.

Table 1: Comparison of test-set perplexity and ratio of number of parameters

	first step (initial set)	second step variable length		word bigram
No. of classes	700	700	1,000	9,212
perplexity	145	111	100	117
parameters	0.23	1.4	1.7	1

6. REFERENCES

[1] Dugast, C., et. al.; "Continuous speech recognition tests and results for the NAB'94 Corpus", *Proc. SLST Workshop*, 1995.

[2] Kubala, F., et al.; "Toward automatic recognition of broadcast news," *Proc. DARPA SR Workshop*, pp.55-60, 1996.

[3] Lau, R., Rosenfeld, R. & Roukos, S.; "Trigger-based language models: a maximum entropy approach," *ICASSP'93*.II-45–48, 1993.

[4] Bahl, L. R., Brown, P. F., de Souza, P. V. & Mercer, R. L. : "A tree-based statistical language model for natural language speech recognition," *IEEE ASSP* **37**, pp.1001–1008, 1989.

[5] Niessler, T. R. & Woodland, P. C.: "Variable-length category-based *n*-gram language model," *Proc. ICASSP-96*, pp.164-167, 1996.

[6] Masataki, H. & Sagisaka, Y.: "Variable-order N-gram generation by word-class splitting and consecutive word grouping," *Proc. ICASSP-96*, pp. 188-192 1996.

[7] Brown, P. F., et al; "Class-based *n*-gram models of natural language," *Computational Linguistics*, Vol.18, 4, pp. 467-479, 1992.

[8] Kneser, R. & Ney, H.; "Improved clustering techniques for class-based statistical language modeling," *Eurospeech'93*, pp.973-976, 1993.

[9] Matsunaga, S. & Sagayama, S.: "Variable-length language modeling integrating global constraints," *Proc. Eurospeech'97*, pp.2719-2722, 1997.

[10] Isotani, R. & Matsunaga, S.: "A stochastic language model for speech recognition integrating local and global constraints," *Proc. ICASSP-94*, pp. II-5-II-8, 1994.