EUROSPEECH '97
5th European Conference on Speech Communication and Technology
Rhodes, Greece, September 22-25, 1997

VARIABLE-LENGTH LANGUAGE MODELING INTEGRATING GLOBAL CONSTRAINTS

Shoichi Matsunaga and Shigeki Sagayama

NTT Human Interface Labs.,
1-1, Hikari-no-oka, Yokosuka-shi, Kanagawa 238 Japan.
E-mail: mat@nttspch.hil.ntt.co.jp

ABSTRACT

This paper proposes a novel variable-length class-based language model that integrates local and global constraints. In this model, the classes are iteratively recreated by grouping consecutive words and by splitting initial part-of-speech (POS) clusters into finer clusters (word-classes). The main characteristic of this modeling is that these operations of grouping and splitting is carried out selectively, taking into account 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. Experiments showed that the perplexity of the proposed model for the test corpus is lower than that of conventional models and that this model requires a small number of statistical parameters, showing the model's effectiveness.

1. INTRODUCTION

Word n-gram models are widely used as language models for continuous speech recognition[1]. However, they have two disadvantages for effective modeling. One is that they represent only local constraints within a few successive words and lack the ability to capture global or long-distance dependencies between noncontiguous words. The other is that, for a powerful model, even trigram, the higher-order n involves an increase in the number of parameters that must be estimated, which results in data being sparse. These disadvantages often prevent speech recognition performance from improving.

Some researchers have recently tried to cope with the former problem by introducing long-distance factors. Typical models including such factors are the extended-bigram[2], the trigger pair[3], and the tree-based[4] models. However, the extended-bigram uses either only local information or only long-distance information for each word. Other models require a great deal of computation and training data. We also proposed a model using the conventional word n-grams for local constraints, and using function- and content- word n-grams for long-distance factors[5]. In this model, function-word n-grams are intended mainly for syntactic constraints, while content-word n-grams are for semantic ones. We showed their effectiveness in Japanese speech recognition, but the sparseness of the training data for content words is a problem when dealing with a large vocabulary.

For the latter problem, word-grouping techniques[6, 7] and variable-length modeling[8, 9] were devised to compensate for the insufficient statistics for rare word sequences, and to cope with the variation of data distributions in a training language corpus. However, these methods deal with only consecutive word constraints.

Addressing these problems, we propose a novel variable-length class-based language model that integrates global constraints. The classes are repeatedly renewed by grouping consecutive words and by splitting initial POS clusters into more suitable classes. These procedures are carried out iteratively and selectively, taking account of long-distance factors based on a minimum entropy criterion.

This paper mainly focuses on how to generate this model effectively, and the model is evaluated with respect to test-set perplexity. Experiments show that the perplexity of the proposed model is lower than that of conventional models and that this model requires a small number of statistical parameters.

2. MODEL GENERATION

2.1. Class-based modeling using global constraints

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

$$\begin{aligned} P(S) &= P(w_1, w_2, \ldots, w_n) \\ &= \prod_{i=1}^{N} P(w_i \,|\, w_1^{i-1}). \end{aligned} \tag{1}$$

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

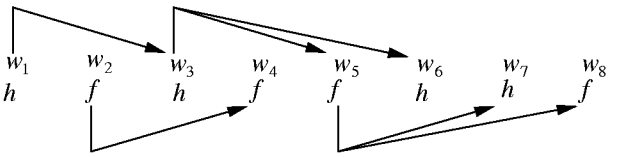


Figure 1: Examples of global constraints in a sentence

h_i the last content word in the substring $w_1^{i\,1}$. Taking f_{i-1} and h_{i-1} into consideration as well as w_{i-1}, the probability of a word w_i given w_1^{i-1} is, represented approximately as follows[5]:

$$P(w_i \,|\, w_1^{i-1}) \simeq P(w_i \,|\, w_{i-1}, h_{i-1}, f_{i-1}). \qquad (2)$$

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

$$P(w_i \,|\, w_{i-1}, h_{i-1}, f_{i-1}) = \begin{cases} P(w_i \,|\, w_{i-1}, f_{i-1}), \\ \quad \text{if } w_{i-1} \text{ is a content word} \\ P(w_i \,|\, w_{i-1}, h_{i-1}), \\ \quad \text{if } w_{i-1} \text{ is a function word.} \end{cases} \qquad (3)$$

Figure 1 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.

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

$$P(w_i \,|\, w_{i-1}, h_{i-1}, f_{i-1}) \simeq P(w_i|C_i) \cdot P(C_i|C_{i-1}, R_{i-1}) \qquad (4)$$

where C and R are word classes ($w_i \in C_i$, $R_{i-1} \ni f_{i-1}$ or h_{i-1}). POS word-grouping is employed for the initial classes, so each word class is composed of function or content words.

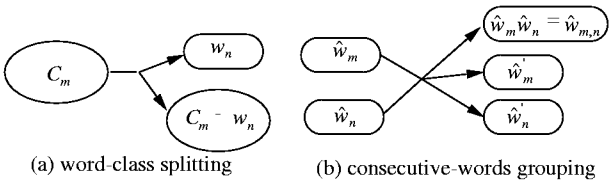
2.2. Iterative word-class setting

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 POS classes, the following two types of procedures are carried out repeatedly (Figure 2)[9]:

(a) Word-class splitting:

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

[1]All words in the lexicon were divided into content words such as nouns, verbs, adjectives, adverbs, and function words such as auxiliary verbs and case markers, on the basis of part-of-speech (POS).



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

Figure 2: Two procedures in variable-length modeling

(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: $\hat{w}'_m$, whose succeeding word is not $\hat{w}_n$, and $\hat{w}'_n$, whose preceding word is not $\hat{w}_m$ ($\hat{w}$: word sequence including single words). When a content word and a function word are grouped together, this sequence is treated as a new word which has properties of both a content word and a function word. Word-property transformation resultingof from groupings are listed in Table 1.

Consequently, the probability of S is rewritten as

$$P(S) \simeq \prod_{i=1}^{K} P(\hat{w}_i|C_i) \cdot P(C_i|C_{i-1}, R_{i-1}) \qquad (5)$$

where $\hat{w}_i$ is a word sequence belonging to the word-class C_i, K is a 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[9]:

1. Initialization of word class using POS grouping

Table 1: Transformation of word properties by consecutive word grouping

before grouping		after grouping
$\hat{w}_m$	$\hat{w}_{m+1}$	$\hat{w}_{m,m+1}$
content	content	content
function	function	function
content	function	con. & func.
content	con. & func.	con. & func.
con. & func.	function	con. & func.
con. & func.	con. & func.	con. & func.

Other types of grouping are prohibited.

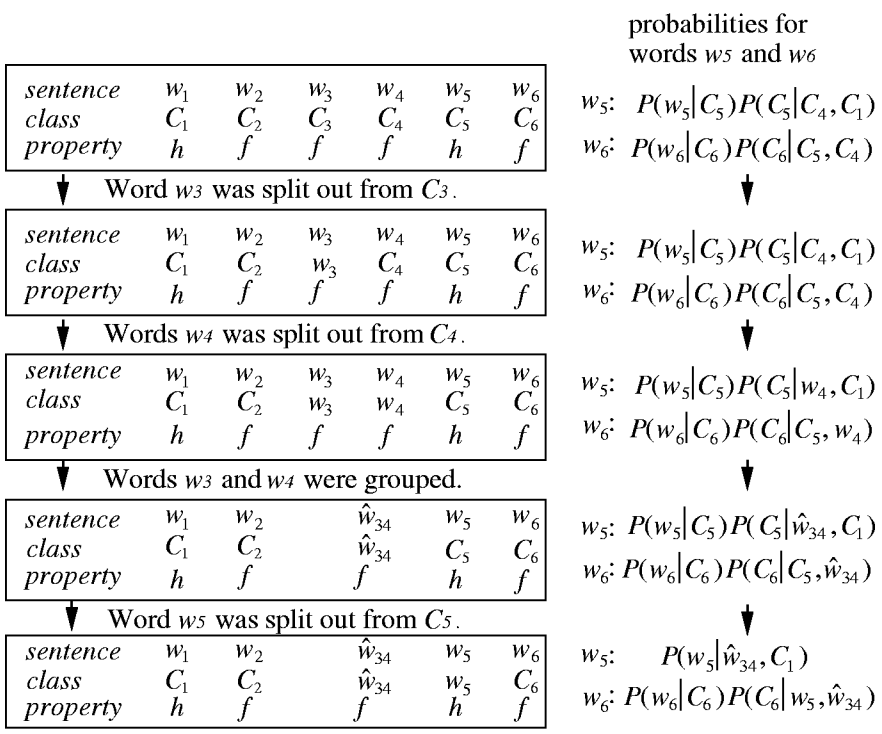


Figure 3: Examples of variable-length model generation integrating global constraints

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_{bmin}) for each pair.

3. Selection of a procedure; 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_{amin} or H_{bmin}, and go back to step 2.

By repeating these processes, a finer variable-length class-based model taking account the global constraints is generated. Figure 3 shows an example generation process and probabilities for w_5 and w_6 for each stage. At the initial stage, w_5 has global dependency on class $C_1(\ni w_1)$ and local constraints with $C_4(\ni w_4)$, and w_6 has global dependency on C_4 and local constraints with C_5. After spliitting w_3, w_4, and w_5 from each cluster C_3, C_4, or C_5, and generating $\hat{w}_{34}$ from consecutive words w_3 and w_4, at the last stage w_5 has global dependency on C_1 and local constraints with $\hat{w}_{34}$, and w_6 has global dependency on $\hat{w}_{34}$ and local constraints with w_5. The probability for w_5 at the last stage becomes
$P(w_5|w_5)P(w_5|\hat{w}_{34}, C_1) = P(w_5|\hat{w}_{34}, C_1)$.

3. EXPERIMENTS

The proposed model was generated using a Japanese text database of spoken dialogues concerning travel arrangements[10]. The data consisted of about 2.2×10^4 sentences with about 3.5×10^5 words (6.4×10^3 different words, where 9% are function words and others are contents words), and POS was tagged to each word. 800 classes were generated from 80 initial POS classes; the number of consecutive word groupings was 163 (23%), and the number of word splits was 557 (77%). The new words generated by these consecutive word groupings consist of 81 function words, 44 content words, and 38 words having both properties.

We evaluated the proposed variable-length model using test-set perplexity. Test dialogue text consisted of 490 sentences (7.4×10^3 words) concerning travel arrangement. Experimental results are shown in Table 2 and Figure 4, where deleted interpolation was carried out to obtain the perplexities. As the number of classes in the proposed model increased, perplexity decreased monotonically. When the number of classes reached 260, the perplexity of the test text (22.0) became lower than that of bigram (22.7), and at 460 classes the perplexity (18.2) became lower than that of trigram (19.0).

Table 2 also lists results obtained by using variable-length class-based n-gram (800 classes) without global constraints[9] (the probability $P(\hat{w}_i|C_i)P(C_i|C_{i-1})$ was used instead of the term in equation (5)), and by us-

Table 2: Comparison between the proposed model and conventional models

	proposed model		bigram	trigram	without global[9]*	word-based global[5]**
No. of classes	80 (POS)	**800**	-	-	800	-
Test-set perplexity	39.9	**16.4**	22.7	19.0	18.2	19.5
Ratio of number of parameters	0.28	**1.5**	1.0	3.0	0.9	1.9

- proposed model: variable-length class-based model *with* global constraints: $P(\hat{w}_i|C_i) \cdot P(C_i|C_{i-1}, R_{i-1})$
- without global*: variable-length class-based model *without* global constraints: $P(\hat{w}_i|C_i) \cdot P(C_i|C_{i-1})$
- word-based global**: *word-based bigram* (fixed-length) *with* global constraints: $P(w_i \mid w_{i-1}, h_{i-1}, f_{i-1})$

ing a word-based bigram (fixed-length) with global constraints[5] (the probability $P(w_i \mid w_{i-1}, h_{i-1}, f_{i-1})$ in equation (2), was used instead of the term in equation (5)). By comparison with the perplexities of these models (18.2 and 19.5), the proposed model is superior. This indicates that both global constraints and variable-length class-based modeling are very useful. Although the number of parameters in the proposed model (1.5) is larger than that in bigram (1.0) or the variable-length class-based model without global constraints (0.9), it is much smaller than in conventional trigram (3.0).

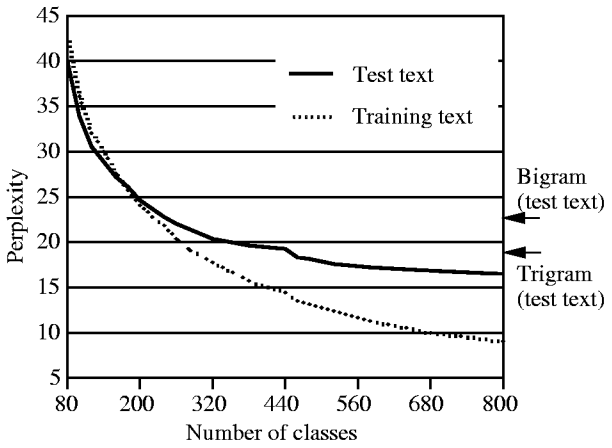


Figure 4: Reduction of perplexity for training text and test text

4. CONCLUSIONS

In this paper, we proposed a new variable-length class-based language model that integrates local and global constraints. This model was generated by grouping consecutive words and by splitting initial POS clusters into finer clusters (word-classes) taking account of local and global dependency on a minimum entropy criterion. The experiments showed that the proposed language model gives lower perplexity than conventional models while keeping the number of parameters comparatively small.

We are planning to apply this language model to the second pass in our multi-pass speech recognition system.

REFERENCES

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

[2] Wright, J. H., Jones, G. J. F. & Lloyd-Thomas, H.; "A consolidated language model for speech recognition," *Eurospeech'93*, pp.977–980. 1993.

[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] 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.

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

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

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

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

[10] T. Morimoto, et al.: "A Speech and Language Database for Speech Translation Research," *Proc. ICSLP-94*, pp.1791-1794, 1994.