文節境界情報を利用した N-gram 言語モデルの高精度化

鄭 聖曄[†],　広瀬 啓吉[†],　峯松 信明[††]

[†] 東京大学大学院 新領域創成科学研究科,　[††] 東京大学大学院 情報理工学系研究科
Tel.: 03–5841–6393,　Fax.: 03–5841–6648
{synim, hirose, mine}@gavo.t.u-tokyo.ac.jp

あらまし　既に音声認識に韻律情報を利用する方法として、n-gram ベースの言語モデルにアクセント句境界を用いることを提案し、perplexity の低下が得られることを示したが、アクセント句境界の特徴を得るために音声データベースを必要とするという問題点があり、それを品詞 n-gram により解決していた。アクセント句は発話の単位であるが、ほぼ文節に対応するもので、それを言語モデルに用いることを今回提案する。音声データベースを必要とせず、精度の高い記述が可能となるが、単語系列から文節境界を推定するプロセスが必要となる。毎日新聞記事データベース 1 年分（1997 年度）を用いて実験を行ったところ、テキストオープンで 3-gram の場合、約 8 %の perplexity の減少を得た。

キーワード　韻律境界情報、文節境界情報、音声認識、言語モデリング、パープレキシティ

N-gram Language Modeling of Japanese using *"Bunsetsu"* Boundaries

Sungyup Chung[†],　Keikichi Hirose[†]　and　Nobuaki Minematsu[††]

[†] Graduate School of Frontier Sciences, University of Tokyo,
[††] Graduate School of Information Science and Technology, University of Tokyo,
Tel.: 03–5841–6393,　Fax.: 03–5841–6648
{synim, hirose, mine}@gavo.t.u-tokyo.ac.jp

Abstract　As one of the methods to utilize prosodic features in speech recognition process, accent phrase boundary information was successfully incorporated into n-gram based language modeling. The method, however, possesses the problem that it requires speech database to acquire the accent phrase boundary features. This problem was solved by introducing part-of-speech n-grams, but that process degraded the model accuracy. This paper proposes the modeling using *bunsetsu* boundary information instead. A *bunsetsu* is a grammatical unit of Japanese and almost corresponds to an accent phrase. This method has the advantage that it needs no speech corpus and can get more reliable results with large-sized training text data. When we evaluated the proposed method using Mainichi Newspaper'97, about 8 % of perplexity reduction was observed in the 3-gram case.

Key words　prosodic boundary information, *bunsetsu* boundary information, speech recognition process, language modeling, perplexity

1 Introduction

A rather large number of research works have been conducted to incorporate prosodic information into speech recognition process. Among them, we have recently proposed a scheme to include accent phrase boundary information into n-gram based language modeling. The method is to separately counting n-grams when and not when the word transition is across the boundary. Although the method showed the validity in perplexity reduction and led to a slight improvement in recognition rate, it included a problem in that it required a speech corpus with accent boundary information. In Japanese, we have another unit called" *bunsetsu* ", which is defined as a unit consisting of content word(s) and its (their) following particle(s). Although *bunsetsu* is a linguistic unit and not a pronunciation unit and may differ from an accent phrase, they almost coincide. This fact inspired us to use *bunsetsu* boundary information in language model instead of accent phrase bundary information. Since *bunsetsu* boundaries are automatically detectable from the text corpus using a parser with rather high accuracy, two types of language models (crossing boundaries and not crossing boundaries) can be directly obtainable from a large text corpus. Instead the method requires a process to predict coming *bunsetsu* boundaries from word history. (In the case of accent phrase boundary, it is detectable directly from prosodic features.) To incorporate boundary information into language modeling is to count the structure of the sentence. From that viewpoint, the proposed method can be said to view a longer span in language modeling. The rest of the paper is constructed as follows; the method using accent phrase boundary is checked in section2. In section 3, we explain the basic idea of the method using *bunsetsu* boundary and show evaluation results in terms of perplexity. Section 4 concludes the paper.

Table 1. *Perplexities for baseline and the model using accent phrase boundary information. Accent phrase boundary labels of the corpus are used. Hereinafter, perplexity is abbreviated as PP.*

	Text Closed	Text Open
Baseline Model	117.0	117.4
PB Model	104.1	107.1
PP Reduction	11%	8.77%

2 The Model Using Prosodic Boundary Information

In this section, the language modeling method using prosodic (accent phrase) boundary information [1] will be explained briefly to show how we reached the idea of using *bunsetsu* boundary information instead. We will also discuss on the weak point and the limitation of the method.

2.1 Basic Idea & Evaluation Results

Two different language models are made from two different word transitions, inter & intra prosodic boundary transitions. They will be selected and used according to the existence & absence of accent phrase boundary during the decoding process of speech recognition. The evaluation in terms of perplexities was conducted. The conditions of evaluation are as follows:

- language model: word 2-gram
- vocabulary size: 20k
- morpheme analysis: Chasen [2]
- discounting method: Good Turing discount
- training data for baseline model: Mainichi Newspaper'97
- training data for POS (part of speech) 2-gram calculation: ATR503 sentence speech corpus[4]

Table 1 shows the results using the boundary information labeled in the speech corpus. Text closed represents the case that the whole of ATR 503 sentences are used for training and evaluation. 11 % of reduction rate of perplexity from the baseline model is observed, indicating

the validity of this modeling method. Text open represents the case where 453 sentences are used for training and remaining 50 sentences are used for evaluation. The reduction rate was still close to 9 %.

2.2 Modeling Scheme and Remaining Problems

This method has a major difficulty in collecting enough training data with prosodic information. Usually when training language models, a large-sized text corpus, such as a newspaper corpus for a year or more, is required. Therefore, to train two types of language models directly, a huge speech corpus with accent phrase boundary labels is required. Since to prepare such a corpus is impossible, an indirect way using only a small speech corpus should be adopted instead. It is based on the fact that differences in word transitions according to the existence and absence of accent phrase boundaries are well preserved in part-of-speech transitions; instead of directly constructing two types of models, part-of-speech (POS) transitions for the two cases were counted from a small speech corpus at first, and then original word n-gram counts of the text corpus were divided into the two cases, according to the counting result of POS transitions.

This modeling scheme was adopted in order to solve the training data insufficiency, one of the most serious problems when training the language models for speech recognition. Although this scheme yielded a good result with perplexity reduction, it still has the problem the real word n-gram counts are not used. We need to use real word n-gram counts to get more reliable results. Moreover, the scheme was evaluated only for *n=2* case, and the evaluation of *n>2* cases could not be conducted, because of training data limitation. From this consideration, we propose to use *bunsetsu* boundary instead of accent phrase boundary in the next section.

Table 2. *The comparison of the number of accent phrase boundary and bunsetsu boundary in ATR 503 sentences.*

The number of Accent Phrase Boundary labeled in ATR 503	The number of *Bunsetsu* Boundary parsed by KNP
2841	2835

3 The Model Using *Bunsetsu* Boundary Information

Here we propose a scheme to use *bunsetsu* boundary information instead of prosodic boundary information. Although larger syntactic boundaries can be used, in the current paper, we focus on *bunsetsu* boundaries only; the first step of including syntactic structure into language modeling.

First of all, we need to confirm that the merit of using accent phrase boundaries is kept when using *bunsetsu* boundaries. Actually, most *bunsetsu* boundaries coincide with accent phrase boundaries. In the next section, we first show this taking ATR speech corpus as an example. Then, we show the basic idea of using *bunsetsu* boundaries for n-gram modeling. The method is able to use the information which usual 3-gram cannot use, and therefore offers a better language modeling. The evaluation results in terms of perplexities are shown also.

3.1 The Similarity of Accent Phase Boundary and *Bunsetsu* Boundary

While right answer of accent phrase boundaries is labeled in ATR 503 sentence speech corpus, which was used in section 2.3, the result from KNP was used as right answer of *bunsetsu* boundary in the modeling. If the right answer of *bunsetsu* boundary is almost the same to that of accent phrase boundary, the use of *bunsetsu* boundaries may yield a good result. Also even the newspaper text corpus of more than ten years

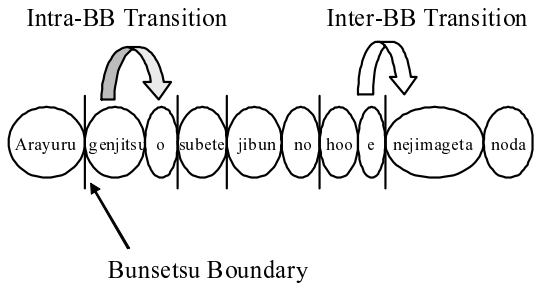

Fig 1. *Two types of word transitions. Transitions across bunsetsu boundaries and those not across are modeled separately. The word BB abbreviates "bunsetsu boundary".*

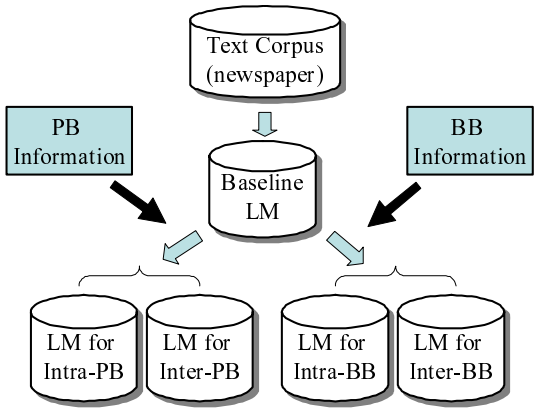

Fig 2. *Schematic illustration for two kinds of n-gram languge modeling method, using prosodic boundary information and bunsetsu boundary information. In the figure and also after part of this paper, LM denotes language models, PB for prosodic boundary and BB for bunsetsu boundary.*

can be used.

Table 2 shows the numbers of accent phrase boundaries and *bunsetsu* boundaries in the ATR corpus. Numbers are mostly coincide for both of the cases, indicating the similar effect is expected when we use *bunsetsu* boundaries, obtained using KNP as a parser, instead of accent phrase boundaries labeled in the corpus.

3.2 Basic Idea & Modeling Scheme

Basically two different language models would be made in the modeling method using *bunsetsu* boundary information. The first one is "the language model predicting next word" and the second one is "the language model predicting *bunsetsu* boundary". The former is normal n-gram language model, and the latter predicts if the *bunsetsu* boundary would appear or not after current word history. Also, we divide "the language model predicting next word" into two different models according to two different word transitions; inter and intra *bunsetsu* boundary transitions. Figure 1 shows two different word transitions according to *bunsetsu* boundaries. Finally, two different language models predicting next word will be combined by the appearance probability of *bunsetsu* boundary calculated from the language model predicting *bunsetsu* boundary, during the decoding process of speech recognition. Figure 2 illustrates the basic idea of the language modeling method using *bunsetsu* boundaries by comparing to the method using prosodic boundaries.

We can say that the vocabulary size of the language model predicting *bunsetsu* boundary is only 2, *bunsetsu* boundary and word (no boundary). In this case we can use the word history longer than 2, because it will be possible in the aspect of data quantity. Usually 3-gram uses 2 words of history to predict the next word, and it uses automatically the *bunsetsu* boundary information extracted from the 2 words of history. The language modeling method using *bunsetsu* boundary information, however, extracts the *bunsetsu* boundary information from the word history longer than 2, and predicts next word using 2 words of history.

First of all, we conducted an experiment in order to know the difference in the ability of predicting *bunsetsu* boundary between the case of *word history=2* and the case of *word history>2*. If the predicting ability of the *bunsetsu* boundary appearance is better in the case of *word history>2* than in *word history=2*, the proposed method using word history longer than 2 will make better

Table 3. *The difference in the ability of predicting bunsetsu boundaries between the case of "word history=2" and that of "word history>2".*

	Predicting Ability
History=2	83.65%
History=3	87.52%
History=4	93.42%

performance than normal 3-gram using 2 words of history. Mainichi Newspaper Corpus '97 was used for this experiment. We tried to count how many cases we could know if the *bunsetsu* boundary would appear or not after the current word history, by the probability of more than 80 %. As like the results from Table 3, in the case of *history>2*, the predicting ability of the *bunsetsu* boundary was better than that in the case of *history=2*. Here, "the predicting ability of the *bunsetsu* boundary" is defined as follows;

$$PA = \frac{RP}{T} * 100(\%)$$

where, PA stands for the predicting ability of the *bunsetsu* boundary, RP for the word sequence regarded as right prediction and T for total numbers of word sequence. We can say that normal 3-gram (2 word history) cannot predict perfectly the appearance of *bunsetsu* boundary. Getting longer word history, the predicting ability gets better.

Two kinds of language model predicting next word, inter-*bunsetsu*-boundary & intra-*bunsetsu*-boundary model, are combined by the language model predicting *bunsetsu* boundary, as like a formula below. If we define a variable q as the appearance probability of *bunsetsu* boundary after current word history,

$$\begin{aligned} P_{total}(w3|w1,w2) &= q * P_{inter}(w3|w1,w2) \\ &\quad +(1-q) * P_{intra}(w3|w1,w2) \end{aligned}$$

where, word sequence (w1,w2) stands for 2 words of history and w3 for next word to be predicted. P_{inter} represents the appearance probability of next word calculated from inter-*bunsetsu*-boundary model and P_{intra} from intra-*bunsetsu*-boundary model. Also P_{total} represents the word appearance probability of next word given as the combination of P_{inter} and P_{intra}.

The experiment for the model using *bunsetsu* boundary information is based on word 3-gram. When we found the probability q, if the language model prediciting *bunsetsu* boundary is based on $n< 3$, we did not use proposed models, but used baseline model for calculating the probability of word appearance. It means we did not adopt discounting method for predicting *bunsetsu* boundary model yet. This is left as a future work. Also, if we represent the event existing the *bunsetsu* boundary as c,

$$\begin{aligned} P(w3|w1,w2) &\equiv P(w3,c|w1,w2) + P(w3,\bar{c}|w1,w2) \\ &\equiv P(w3|w1,w2,c)P(c|w1,w2) \\ &\quad +P(w3|w1,w2,\bar{c})P(\bar{c}|w1,w2) \end{aligned}$$

where, $P(c|*) + P(\bar{c}|*) = 1.0$

According to the formula above, if we predict the existence or absence of *bunsetsu* boundary by *history=2*, it will be the same as normal word 3-gram. Therefore we should used word history longer than 2 in order to find $P(c|*)$. It could be said that the aim of proposed method is to improve normal 3-gram by the higher precision of $P(c|*)$ from longer word history.

3.3 Evaluation in terms of Perplexities

The conditions of evaluation are like below:

- language model: word 3-gram
- vocabulary size: 20k
- morpheme analysis: JUMAN & KNP[3]
- discounting method: Good Turing discount
- training data: Mainichi Newspaper '97
- evaluation text:
  -5 sets of 1000 sentences from Mainichi Newspaper '97 (for closed evaluation)
  -5 sets of 1000 sentences from Mainichi Newspaper '98 (for open evaluation)

Table 4. Perplexities for baseline and the model using bunsetsu boundary information. Used 3 words as history.

	Text Closed	Text Open
Baseline model	29.07	97.04
BB model	25.35	89.05
PP Reduction	12.80%	8.23%

Table 5. Perplexities for baseline and the model using bunsetsu boundary information. Used 4 words as history.

	Text Closed	Text Open
Baseline Model	29.07	97.04
BB model	24.48	89.80
PP Reduction	15.79%	7.46%

Table 4 and 5 show the results. In the text closed experiments, the reduction of the perplexity from the baseline model was about 12% when *history=3*, and it was about 15% when *history*=4. These results correspond to Table 3, which says the ability of predicting *bunsetsu* boundary becomes better when the word history gets longer. Also in the text open experiments, 8% of reduction for *history=3* and 7% for *history*=4. Unlike text closed case, the performance when *history=3* was better than when *history*=4. We could say that the results were caused by the fact that the training text data were only 1 year of newspaper texts which are not enough to train the language model of *bunsetsu* boundary prediction. Also from the fact that when we make normal word 3-gram the defacto-standard is *n=3*, even though we increase the quantity of training data, it is possible that the case of *history=3* makes the best result.

4 Conclusions and Future Plans

In this paper, a method including *bunsetsu* boundary information into n-gram language modeling were introduced and compared to the method using accent phrase boundary information. Perplexity reduction from the baseline language model was observed indicating the validity of this method.

As the future plan, we will make use of the deeper syntatic boundaries than *bunsetsu* ones. Also proposed language modeling method, based on the characteristics of Japanese, will be adopted to Korean language, which has very similar grammatical characteristics.

References

[1] K.Hirose; N.Minematsu and M.Terao; N-gram language modeling of japanese using prosodic boundaries. Proc.Speech Prosody 2002, Aix-en-Provence, 2002-4

[2] Japanese morpheme analyser. Version 2.02. http://chasen.aist-nara.ac.jp/

[3] Japanese morpheme analyser. Version 2.0b6. http://www-lab25.kuee.kyoto-u.ac.jp/nl-resource/knp.html

[4] Speech Corpus Set B. http://www.red.atr.co.jp/database_page/digdb.html