N-gram Language Modeling of Japanese Using *Bunsetsu* Boundaries

Sungyup Chung[1], *Keikichi Hirose*[1] *and Nobuaki Minematsu*[2]

[1]Graduate School of Frontier Sciences, University of Tokyo, Japan
[2]Graduate School of Information Science and Technology, University of Tokyo, Japan
{synim, hirose, mine}@gavo.t.u-tokyo.ac.jp

Abstract

A new scheme of N-gram language modeling was proposed for Japanese, where word N-grams were calculated separately for the two cases: crossing and not crossing *bunsetsu* boundaries. Here, *bunsetsu* is a basic grammatical (and pronunciation) unit of Japanese. A similar scheme using accent phrase boundaries instead of *bunsetsu* boundaries has already been proposed by the authors with a certain success, but it suffered from the training data shortage, because assignment of accent phrase boundaries requires a speech corpus. In contrast, *bunsetsu* boundaries can be detected automatically from a written text with a rather high accuracy using a parser. It was shown from the experiment that a perplexity reduction was possible by estimating *bunsetsu* boundaries from the history longer than N-1 words in the case of N-gram modeling and by selecting one from two types of models (crossing and not crossing *bunsetsu* boundaries) according to the estimation. When 1 or 3 years of Mainichi Newspaper corpus was used for the training of tri-grams, the proposed scheme could reduce the perplexity by around 8% from the baseline modeling (without separation). The proposed language modeling was applied to a continuous speech recognition, and it showed that an improvement in word recognition rate was possible especially when the training corpus was small (1 year of newspaper).

1. Introduction

Nowadays, statistical language modeling is the key technology in the continuous speech recognition. N-gram language modeling is widely utilized with a great success, though it is based on a simple idea of representing language features as the probability of N word sequences in a text corpus. The major problem of the model is that it only views a very limited range of sentence (N words) and does not explicitly account for the linguistic structure of the word sequence. Several methods have already been proposed trying to view the relationship of words not adjacent to each other [1, 2], but did not take the linguistic structure into account in a clear way. In this paper we newly propose a scheme to include linguistic structure information into the N-gram language modeling.

In the case of Japanese, an accent phrase is mostly consisted of a content word and following particle(s). Therefore, for instance, word transition from a particle to a noun mostly includes an accent phrase boundary. Also as a pronunciation unit, an accent phrase can be segmented from an utterance by observing prosodic features. These considerations led us to a scheme of N-gram modeling: to separately count the N-grams when crossing and not crossing the accent phrase boundaries [3]. Since the accent phrase boundary is only defined for spoken sentences, it cannot be located in a text corpus directly; to locate, a reading version of the corpus is necessary, which is not obtainable practically. To cope with this problem, we used a speech corpus with around 500 sentences, quite small as compared to the text corpus, to find out how the bi-grams of particles differ when crossing and not crossing accent phrase boundaries. The result was used to separate the word bi-gram counting of the text corpus, and the two types of bi-gram language models were constructed. Through experiments using 1 year of Mainich Newspaper '97 as the text corpus and ATR 503 sentence utterances [4] as the speech corpus, perplexity reduction of 9% to 11% from the baseline modeling (without separation) was realized. Using language models thus obtained, around 2% of improvement was realized in word accuracy rate in continuous speech recognition [3]. In the experiments, morpheme analysis was conducted using a free-ware Chasen [5], and accent phrase boundaries were detected using a method based on the model of transitions of mora fundamental frequency contours [6]. The recognition engine was Julius [7]. Although the scheme showed a certain effect on speech recognition, it still contained the following problems:

- Word N-gram counts are calculated indirectly. (Word N-gram counts of the language corpus are separated according to the accent phrase boundary information obtained from the speech corpus.)
- No method is available for accurate detection of accent phrase boundaries.
- The positioning of accent phrase boundaries may be different among speakers and even among utterances by a speaker. This variation may degrade the performance of the model.
- To apply the scheme to word tri-gram, the size of the speech corpus should be enlarged, which is usually difficult.

To solve these problems while keeping the positive effect of the scheme to the language modeling, in the current paper, we propose to use *bunsetsu* boundary information instead of accent phrase boundary information. *Bunsetsu* is the basic grammatical (and also pronunciation) unit of Japanese and consists of a content word followed by particles. A content word can be a sequence of minor content words, a compound word. Also number of particles can be zero or more than one, though it is one in many cases. From this viewpoint, *bunsetsu* is a unit similar to accent phrase. However, *bunsetsu* can be defined for written texts, while accent phase can only be defined for speech. For a given text, *bunsetsu* boundary can also be located as the position where a speaker can insert pauses without damaging naturalness of speech. This definition sounds ambiguous, but Japanese can do it consistently.

Bunsetsu boundaries are automatically detectable for a given text using a parser with rather high accuracy, and, this fact implies a result better than when using accent phrase boundaries. Instead, when using *bunsetsu* boundaries, a process to predict *bunsetsu* boundaries from the preceding word sequences is required during the decoding process of speech recognition, and its performance largely affects the effect of the proposed language model to the speech recognition. The rest of the paper is constructed as follows; section 2 explains the basic idea of proposed method using *bunsetsu* boundaries. In section 3, the modeling is evaluated in terms of perplexity and word recognition rate. Section 4 concludes the paper.

2. Language modeling using *bunsetsu* boundaries

2.1. Ability of predicting *bunsetsu* boundaries from preceding words

In the proposed modeling, a history longer than N-1 words is used to tell whether the following word boundary is a *bunsetsu* boundary or not. As explained later in section 2.2, if history of N-1 words is used, the proposed modeling is the same with the baseline word N-gram modeling (without separation). The proposed modeling may outperform the baseline one when a better prediction of *bunsetsu* boundary is realized by viewing N or longer word history. We first investigated how the prediction ability changed using 7 years of Mainichi Newspaper corpus from ’91 to ’97. Table 1 shows the result as the predicting ability PA, which is defined as;

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

Where, PB stands for the number of predictable boundaries, and T is the total number of word boundaries. Here, if a word boundary is identified as *bunsetsu* or non-*bunsetsu* boundary with more than 80% accuracy, it is counted as a predictable boundary. In the table, the case ”Word History Length=2” shows the baseline prediction ability for tri-gram. It is clear from the result that better prediction is possible when a longer word history is used.

Table 1: *Ability of predicting bunsetsu boundaries for several lengths of word history. History length of two words corresponds to the tri-gram.*

	Predicting Ability
Word History Length=2	84.62%
Word History Length=3	86.66%
Word History Length=4	91.27%

2.2. Modeling scheme

Figure 1 shows an example for each of two different types of word transitions; transition from noun ”genjitsu (reality)” to particle ”o” includes no *bunsetsu* boundary (intra-*bunsetsu* transition) and transition from particle ”e” to verb ”nejimageta (twisted)” includes a *bunsetsu* boundary (inter-*bunsetsu* transition). The proposed scheme is to construct two types of word N-gram models after separating N-gram counts when transition from N-1th word to Nth word is intra-*bunsetsu* and when it is inter-*bunsetsu*. Henceforth, these two types shall be called intra-*bunsetsu* language model and inter-*bunsetsu* language model. (In the current scheme, *bunsetsu* boundaries, which may occur in the word history, are not taken into account.) In the decoding process of speech recognition, likelihood obtained by the inter-*bunsetsu* model and that by the intra-*bunsetsu* model are interpolated according to the probability of *bunsetsu* boundary occurrence between word history and the current word. The probability of current word can be calculated by:

$$\begin{aligned} P_{total}(w_n|w_{n-N+1}^{n-1}) &= pb * P_{inter}(w_n|w_{n-N+1}^{n-1}) \\ &\quad +(1-pb) * P_{intra}(w_n|w_{n-N+1}^{n-1}) \end{aligned}$$

where, word sequence w_{n-N+1}^{n-1} stands for N-1 word history and w_n stands for the current word. P_{inter} and P_{intra} represent the probabilities of current word calculated from inter-*bunsetsu* model and intra-*bunsetsu* model, respectively. pb is the occurrence probability of *bunsetsu* boundary between word history and the current word.

To calculate pb, we need a predictor for *bunsetsu* boundaries. For this purpose, another kind of language model, which tells the probability of a word boundary being a *bunsetsu* boundary or not from the preceding word history, is constructed. Henceforth, we shall call this

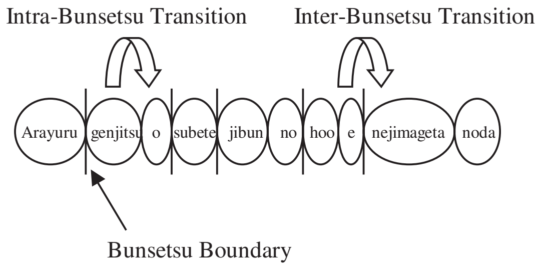

Figure 1: *Two types of word transitions in Japanese sentence ”arayuru geNjitsuo subete jibuNno hooe nejimagetanoda ((He) twisted all the reality to his side.).” Each circle indicates a word.*

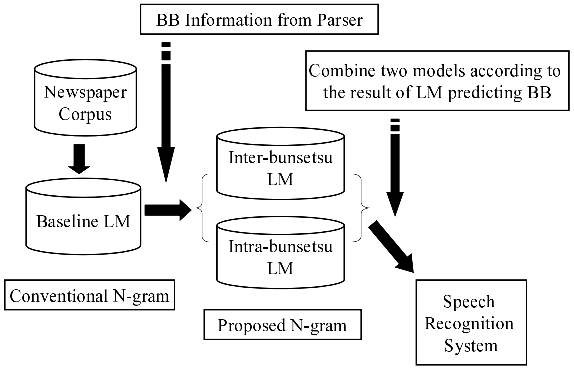
BB Information from Parser
Combine two models according to the result of LM predicting BB
Newspaper Corpus
Inter-bunsetsu LM
Baseline LM
Intra-bunsetsu LM
Conventional N-gram
Proposed N-gram
Speech Recognition System

Figure 2: *Schematic illustration for the proposed scheme of language modeling using bunsetsu boundaries. LM and BB denote language models and bunsetsu boundaries, respectively.*

model as the boundary model. To train a language model with longer history requires a larger text corpus, and this fact makes it difficult to train word N-gram models when N is larger than 3. However, in the case of boundary model, the vocabulary size is limited 2 (*bunsetsu* boundary or non-*bunsetsu* boundary), and the model with a longer word history can be trained. When we represent the event that a *bunsetsu* boundary exists after the word history as b, and the event that does not as $\bar{b}$, the probability of word w_n for the baseline language model can be rewritten as:

$$\begin{aligned} P(w_n|w_{n-N+1}^{n-1}) &= P(w_n, b|w_{n-N+1}^{n-1}) + P(w_n, \bar{b}|w_{n-N+1}^{n-1}) \\ &= P(b|w_{n-N+1}^{n-1}) * P(w_n|w_{n-N+1}^{n-1}, b) \\ &\quad +P(\bar{b}|w_{n-N+1}^{n-1}) * P(w_n|w_{n-N+1}^{n-1}, \bar{b}) \end{aligned}$$

Where, $P(b|*) + P(\bar{b}|*) = 1.0$. If we use $P(b|w_{n-N}^{n-1})$, $P(b|w_{n-N-1}^{n-1})$, $P(b|w_{n-N-2}^{n-1})$, etc., instead of $P(b|w_{n-N+1}^{n-1})$, accuracy in predicting *bunsetsu* boundaries increases, leading to a better prediction of occurrence probability of word w_n. In other words, we can improve Japanese N-gram language model by the higher accuracy in the *bunsetsu* boundary prediction using a longer word history. Figure2 schematically shows the whole idea of the proposed modeling. When an N-1 word history is not found in the modeling, the occurrence probability of word w_n should be predicted from word history shorter than N-1 through the discounting method. Since the discounting method was not yet developed for the proposed modeling, in the current paper, the baseline model was used for such a case. In the following section, we will show the validity of the proposed modeling by evaluating it in terms of perplexity. The section will include the result of continuous speech recognition when using the model.

3. Evaluation Experiments

3.1. Conditions

The conditions of evaluation are as follows:

- sampling frequency: 16kHz
- analysis window: 25ms
- frame shift: 10ms
- feature vector: 12MFCC+12 △ MFCC+△ power
- acoustic model: tri-phone model of 3000 states with 64 mixture provided by IPA Japanese dictation software[8]
- language model: word tri-gram
- number of vocabulary: 20k
- parser: Juman&Knp[9] for boundary detection and Chasen[5] for morpheme analysis
- discounting method: Good Turing discount
- recognition engine: Julius version 3.3 multi-path[7]
- training: 1 year ('97) of Mainichi Newspaper with 1.3M sentences, and 3 years ('95- '97) of Mainichi Newspaper with 3.7M sentences.
- testing (text): 3 sets of 1000 sentences from Mainich Newspaper '98 and text for the speech used in the recognition experiment.
- testing (speech): 344 utterances (172 sentence utterances each from female and male) from JNAS speech corpus [10]. JNAS speech corpus consists of the utterances for sentences selected from Mainichi Newspaper from '91 to '94.

3.2. Perplexity in *bunsetsu* boundary prediction

In order to quantitatively evaluate the prediction ability of *bunsetsu* boundaries in different conditions, perplexity of *bunsetsu* boundary PP_{BB} is defined as an analogy of word perplexity;

$$PP_{BB} = (P(b_1)P(b_2) \cdots P(b_n))^{-\frac{1}{n}}$$

where, $P(b_i)$ denotes the probability of i-th word boundary being a *bunsetsu* (or non-*bunsetsu*) boundary, which is predicted by the word history. Since perplexity is the index of the size of predictable branching after a word history, PP_{BB} should take a value between 1 and 2; 1 when *bunsetsu* boundary is fully predictable from the word history, and 2 when not.

3.3. Results of perplexity calculation

Table2 summarizes the word perplexities and *bunsetsu* boundary perplexities for several cases. Perplexities are calculated for the text for the speech recognition experiment ("transcript") and for 3 sets of 1000 sentences from Mainich Newspaper '98 ("newspaper"). We should note that the above texts are selected from Minichi Newspaper of the years not used for the training of language models. The punctuation marks are treated as words in the calculation. "transcript" does not include punctuation marks, which is the reason of the higher perplexities than those for "newspaper." The table shows the perplexities when one-year ('97) text corpus and three-year ('95-'97) text corpus are used for the tri-gram training. For both cases, perplexity is reduced by around 8% from the baseline model. The *bunsetsu* boundary perplexity is eventually 1.11 for all the cases.

Table 2: *Perplexities for the baseline model and the proposed model (BB model). PP denotes perplexity.*

		transcript	newspaper
one year '97	Baseline model	136.29	56.78
	BB model	126.06	51.79
	PP_{BB}	1.11	1.11
	PP reduction	7.51%	8.79%
three year '95 to '97	Baseline model	104.25	54.63
	BB model	95.60	49.99
	PP_{BB}	1.11	1.11
	PP reduction	8.30%	8.49%

3.4. Results in terms of word accuracy rate (WAR)

Experiment of continuous speech recognition was conducted for the 344 utterances from JNAS speech corpus as shown in section 3.1, using Japanese speech recognizer Julius [7]. The recognizer searches recognition hypotheses in two paths: 1st path for quick and rough search and 2nd path for detailed search. In the current experiment, the proposed language model is applied only to the 2nd path, though it is also applicable for the 1st path. Table3 shows the word accuracy rates (WAR). Recognition was done separately for female (F) and male (M) speakers. 172 sentences were divided into two sets of sentences: "MID" and "LARGE." "MID" included 89 sentences and "LARGE" included 83 sentences. Here, "MID" and "LARGE" mean that the sentences are selected from sentence group with mid-size vocabulary (5k) and large-size vocabulary (20k), respectively [10]. Out-of-vocabulary rates were around 0.5%. When one year of newspaper was used for the training, a clear improvement is observable in WAR when the proposed model was used instead of the baseline model. However, when three years of newspaper was used, the advantage of using the proposed model came unclear. The result may indicate that the boundary information is well (implicitly) included in the baseline modeling when the training corpus size comes large enough.

Table 3: *Word accuracy rates (WAR) when the baseline model is used and when the proposed model is used.*

		Baseline model	BB model	WAR change
one year '97	F-MID	56.93	61.52	+8.06%
	F-LARGE	55.21	60.90	+10.31%
	M-MID	52.02	55.32	+6.34%
	M-LARGE	52.71	57.06	+8.25%
three year '95to97	F-MID	67.14	68.88	+2.59%
	F-LARGE	65.84	65.57	-0.41%
	M-MID	62.50	60.21	-3.66%
	M-LARGE	60.19	60.25	+0.10%

4. Conclusion

A method was proposed to include *bunsetsu* boundary information into N-gram language modeling of Japanese as an attempt to count linguistic structure in the statistical language modeling. Experiments showed a perplexity reduction from the baseline model is possible by the proposed scheme. The advantage of the proposed model over the baseline model was proved as word accuracy rate increases, especially when the size of training text corpus is limited. Although, in the current method, the linguistic information taken into account is limited to the existence of *bunsetsu* boundaries, it can be extended to that with boundary depth. Use of higher-order linguistic information is in the scope of our future research. Since accent phrase boundaries mostly occur at *bunsetsu* boundaries, it will be possible to use the detected accent phrase boundaries as *bunsetsu* boundaries in the decoding process. Accent phrase boundary is detected from prosodic features, and is not predicted from word history. This implies an ability of correcting wrong hypotheses, though it should be proved through the future research. Our preliminary experiment already showed that a further perplexity reduction is possible by introducing accent phrase boundaries. Other languages, having linguistic structure similar to Japanese, like Korean, etc., will be our next research targets.

5. References

[1] R.Rosenfeld, "A maximum entropy approach to adaptive statistical language modeling", Computer Speech and Language, Vol.10, No.3, pp.155-186 (1996).

[2] T.Moriya, K.Hirose, N.Minematsu, and H.Jiang, "Enhansed MAP adaptation of N-gram language models using indirect correlation of distant words", Proceedings ASRU, in CD-ROM (2001).

[3] K.Hirose, N.Minematsu, and M.Terao, "Statistical language modeling with prosodic boundaries and its use for continuous speech recognition", Proceedings ICSLP, Spec7Bo.1, Vol.2, pp.937-940 (2002).

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

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

[6] K.Hirose and K.Iwano, "Detection of prosodic word boundaries by statistical modeling of mora transitions of fundamental frequency contours and its use for continuous speech recognition", Proceedings ICASSP, Vol.3, pp.1763-1766 (2000).

[7] T.Kawahara et. al. "Evaluation of Japanese dictation toolkit -1999 version", Technical Report, Spoken Lanuage Information Processing Group, Information Processing Society of Japan, Vol.54, pp.9-16 (2000).

[8] K.Shikano et.al."Speech recoginition system", Ohmsha (2001).

[9] Japanese morpheme analyser.
Version 3.61(Juman), Version 2.0b6(Knp).
http://www-lab25.kuee.kyoto-u.ac.jp/nl-resource/

[10] JNAS Reading version of newspaper corpus. Acoustic Society of Japan.