N-gram Language Modeling of Japanese Using Prosodic Boundaries

Keikichi Hirose†, Nobuaki Minematsu†† & Makoto Terao†††

†Dept. of Frontier Informatics, School of Frontier Sciences, University of Tokyo
††Dept. of Inf. and Commu. Engineering, School of Inf. Science and Tech., University of Tokyo
†††Dept. of Inf. and Commu. Engineering, School of Engineering, University of Tokyo
{hirose, mine, terao}@gavo.t.u-tokyo.ac.jp

Abstract

A new method was developed to include prosodic boundary information into statistical language modeling. This method is based on counting word transitions separately for the cases crossing accent phrase boundaries and not crossing them. Since direct calculation of the above two types of word transitions requires a large speech corpus which is practically impossible to make, bi-gram counts of part-of-speech (POS) transitions were first calculated for a small speech corpus separately for the two cases instead. Then, word bi-gram counts calculated for a large-scale text corpus were divided into the two cases according to the POS transition feature, and finally, two types of word bi-gram models, one crossing accent phrase boundaries and the other not, were obtained. The method was evaluated through perplexity reduction by the proposed models from the baseline models. When correct boundary position was used, the reduction reached 11%, and when boundaries were extracted using our formerly developed method based on mora-F_0 transition modeling, it was 8%. The reduction around 6% was still observed for speech uttered by a speaker different from the one for the corpus used to calculate the POS bi-gram counts.

1. Introduction

In view of importance of prosodic features in human conversation through speech, many researchers have tried to incorporate them into machine speech recognition processes. There may be roughly two possible ways to use prosodic features in speech recognition process. One is to control acoustic features depending on the prosodic information, and the other is to detect prosodic events (prosodic boundaries, word accent types, speech acts, and so on) and to utilize them to control the speech recognition process. The first way has a major problem in speaker dependency and complexity of the effect of prosodic features on acoustic features [1], but will not be addressed here.

Rather large number of research works have already been conducted along the second way. As for the detection of prosodic events, the performance was improved through introducing statistical framework and using segmental boundary information [2] [3]. As for the use of prosodic information for the speech recognition process, no practical method was not developed yet, though a prosodic module was included rather successfully in the Verbmobil speech recognition system [4]. The probabilistic factor of the prosodic boundary positioning may cause us a hesitation in using prosodic information for speech recognition. However, we should also note that the positioning is not a random process, and humans put boundaries only on possible locations, which correspond to some linguistic boundaries. A possible and good way is to use prosodic boundaries only when they are clearly found. A sophisticated answer to the problem was given as an efficient pruning during the decoding process [5].

In the current paper, we propose another novel method to utilize prosodic boundary information in speech recognition. Although introduction of statistical language modeling realized a significant progress in continuous speech recognition, it includes a problem that the modeling is trained only for written texts. As outputs of human process of sound production, spoken sentences cannot be fully represented only by written language grammars. Prosodic information can be utilized to solve this problem, since it is largely related to the feature of spoken language. This consideration led us to an idea of separately modeling the word transitions for the two cases: one across accent phrase boundaries and the other not. Here, ”accent phrase” is a basic prosodic unit defined as a word or a word chunk corresponding to an accent component, which is also called as ”prosodic word.” The major difficulty along this line will be the collection of enough training corpora with prosodic information. In order to solve this problem, we first calculated part-of-speech bi-ram counts for a small speech corpus to find out the ratio of the two cases (across and not across the boundary), and then applied the result to separate the word bi-gram counts of the text corpus into the two cases. Thus we can obtain two types of word bi-gram models. Validity of the method was checked through experiments on how the model perplexity would be reduced by dividing the original model into the two cases. Although, in the current paper, the method is restricted to bi-grams taking the size of the database into account, it is applicable to n-grams (n; 3 or larger).

The following part of the paper is constructed as follows; in section 2, the proposed scheme of language modeling is explained, followed by evaluation experiments in section 3. Section 4 concludes the paper.

2. Modeling scheme

2.1. Outlines

As shown in Figure 1, the proposed method is based on separately modeling inter-accent-phrase word transitions and intra-accent-phrase word transitions. During the decoding process of speech recognition, the two types of n-gram language models are selected and used according to the existence/absence of accent phrase boundaries. The separate modeling may lead to a reduction of perplexity when the word transition crossing an accent phrase boundary and that not crossing show quite different features to each other. In Japanese, an accent phrase mostly coincide with a ”bunsetsu,” which is defined as a basic unit of grammar and pronunciation, and consists of a content

word (or content words) followed or not followed by a function word (or function words). Also, Japanese speakers sometimes delete particles in a ”bunsetsu,” and in that occasion mostly insert an accent phrase boundary after the ”bunsetsu.” Therefore, accent phrase boundaries usually occur before content words. Tables 1 and 2 show how transitions of part-of-speech differ when crossing and when not crossing accent phrase boundaries. The ATR 503 sentence speech corpus and accent phrase boundary labels for speaker MYI's utterances were used to obtain the tables [6]. The result clearly indicates the differences in part-of-speech transitions and thus implies the differences in word transitions according to the existence/absence of accent phrase boundaries.

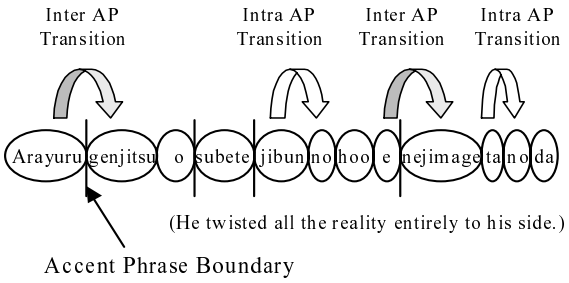

Figure 1: Two types of word transitions. Transitions across accent phrase boundaries and those not across are modeled separately. The word ”accent phrase” is abbreviated as AP in the figure and rest of the tables and figures.

Table 1: Probabilities(in %) of intra-accent-phrase part-of-speech transitions

		Transition to:			
		Noun	Verb	Particle	Adverb
Transition from:	Noun	8.9	5.2	67.5	0.1
	Verb	6.2	12.7	43.8	0.0
	Particle	6.8	47.9	36.9	0.3
	Adverb	2.5	15.0	60.0	0.0

Table 2: Probabilities(in %) of inter-accent-phrase part-of-speech transitions

		Transition to:			
		Noun	Verb	Particle	Adverb
Transition from:	Noun	71.1	13.4	1.4	2.8
	Verb	85.6	5.7	1.1	4.0
	Particle	51.1	34.3	0.2	6.1
	Adverb	59.6	28.8	0.0	1.4

2.2. Problem and solution

When training language models, a large-sized text corpus, such as a newspaper corpus for one or more years, is required. When training the two types of language models of our method, we need a huge speech corpus with accent phrase boundary information, which is mostly impossible to prepare. As pointed out in Tables 1 and 2, differences in word transitions according to the existence/absence of accent phrase boundaries can be well represented as part-of-speech transitions. Therefore, instead of directly constructing the two types of models, we first counted part-of-speech transitions for the two cases for a small speech corpus, and then divided word n-gram counts of the text corpus used for the training of the baseline language models according to the result. Figure 2 schematically illustrates this procedure to construct two types of language models. From now on, the language models before separation shall be called the baseline language models, and those separated using accent phrase boundary information shall be called the proposed models.

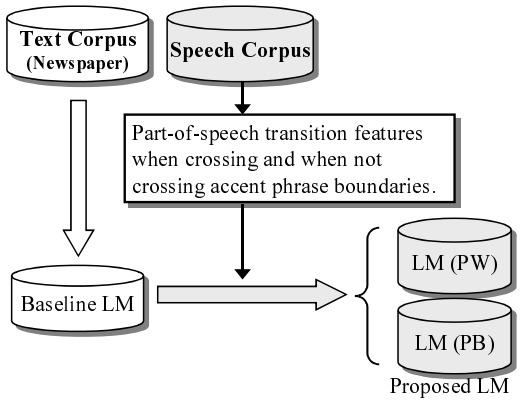

Figure 2: Schematic illustration for the proposed scheme of constructing two types of n-gram language models. In the figure, LM denotes language models, and PB and PW respectively indicate models crossing boundaries and those not crossing.

The concrete procedure currently used can be summarized as follows:

1. Part-of-speech transition counting when crossing accent phrase boundaries and when not crossing them.

   First select a speech corpus, and detect accent phrase boundaries. (Or use accent phrase boundary labels attached to the corpus.) Then parse each sentence of the corpus to obtain a part-of-speech sequence. In the current paper, 26 part-of-speeches were selected such as nouns, pronouns, verbs, adjectives, adverbs, case particles, conjunctions, symbols, punctuation marks, and so on. Finally, we can obtain a chart of part-of-speech transition numbers summarized separately for the two cases: crossing and not crossing accent phrase boundaries. The idea of the chart can be given from Table 3, though it is only for the explanation and does not show actual data.

2. Division of word bi-gram counts of baseline language models.

   In the proposed method, the bi-gram counts of the baseline models are first divided into those crossing accent phrase boundaries and those not crossing. Then two types of bi-gram models are re-constructed for the two cases. The division is done according to the part-of-speech transition feature obtained above. For instance, if the bi-gram count for the sequence ”watashi (I) + wa

Table 3: An imaginary example for the part-of-speech transition counts for the two cases of crossing and not crossing accent phrase boundaries.

Part-of-speech transition	Intra AP	Inter AP
Pronoun → Case practicle	90	10
. . . . .	..	..
Case practicle → Noun	10	40
. . . . .	..	..

(am)" is 1000, it is divided into 900 and 100 for the cases of not crossing and crossing boundaries, respectively. This is because the sequence is the combination of "pronoun + case particle," and from Table 3, the 90 ($= 90/(90+10)$)% of transitions are estimated to occur without accent phrase boundaries and 10% to occur with them. In short, the division is based on the assumption that the part-of-speech bi-gram features obtained from a small corpus also stand for the word bi-gram features of a large text corpus used for the training of the baseline models. Figure 3 shows the above process to construct two types of language models.

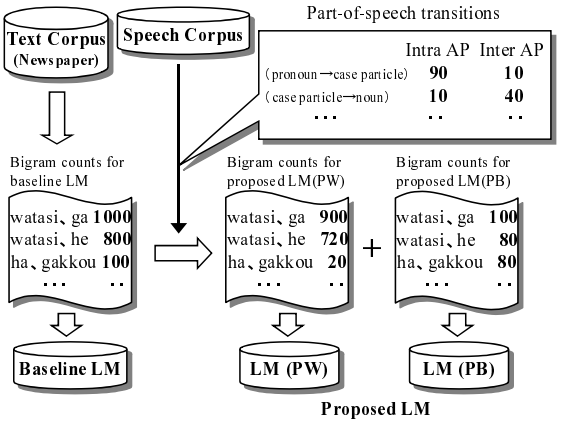

Figure 3: Method of constructing two types of bi-gram models by dividing bi-gram counts of the baseline models. LM, PB and PW are abbreviations for "language models," "models crossing boundaries," and "models not crossing boundaries," respectively.

3. Evaluation through perplexities

3.1. Outlines

The baseline bi-gram models were trained for Mainichi Newspaper Corpus '97. The vocabulary size was 20 k words. Chesen version 2.02 [7] was used for the morpheme analysis and the Good Turing discounting was adopted for the bi-gram calculation. As for the speech corpus for the part-of-speech bi-gram calculation, ATR continuous speech corpus of 503 sentences uttered by speakers MYI and MHT was selected [6]. The perplexity of the proposed models was calculated also for the ATR continuous speech corpus by changing two types of models according to the existence/absence of the accent phrase boundaries. Text closed and open experiments and speaker closed and open experiments were conducted.

3.2. In the case of correct accent phrase boundaries

Experiments were conducted for speaker MYI's utterances using accent phrase boundary labels attached in the speech corpus. The boundary information was used in the training of proposed models and also for the perplexity calculation. Table 4 shows the result when all of 503 sentences were used both for training and perplexity calculation. The total perplexity reduction from the baseline models to the proposed models was reached 11.0%, indicating the validity of the proposed method. When the perplexity reductions were viewed separately for the cases not crossing and crossing accent phrase boundaries, they were about 9% and 15%, respectively. The proposed method is valid for the both cases, but the effect was larger for the bi-gram crossing accent phrase boundaries.

Table 5 shows the result of text-open experiment, where the corpus was divided into 453 and 50 sentences, and used for training and perplexity calculation, respectively. The cross validation scheme (ten combinations of training and evaluation data) was adopted to obtain the reliable result. 9% of perplexity reduction was still observed on average.

Table 4: Perplexities for baseline and proposed models for speaker MYI speech when the accent phrase boundary labels of the corpus are used. Trained and evaluated for all the 503 sentences.

	All	Intra AP	Inter AP
Baseline model	117.0	25.56	2664
Proposed model	104.1	23.32	2253
Reduction rate	11.0%	8.76%	15.4%
Bi-gram hit rate	95.35%	97.65%	90.60%

Table 5: Perplexities for baseline and proposed models for speaker MYI speech when the accent phrase boundary labels of the corpus are used. Trained for 453 sentences and evaluated for the rest 50 sentences. The cross validation scheme is used.

	All	Intra AP	Inter AP
Baseline model	117.4	25.66	2752
Proposed model	107.1	23.93	2408
Reduction rate	8.77%	6.74%	12.5%

3.3. In the case of automatically detected accent phrase boundaries

Experiments were further conducted in the more realistic situation, where the accent phrase boundaries were detected automatically. The detection was done by a method formerly developed by one of the authors. It is based on modeling an accent phrase F_0 contour as an HMM of mora unit F_0 contours [3]. For the 503-sentence speech by speaker MYI, the detection rate and insertion error rate were 57% and 24%, respectively. Here, a morpheme boundary obtained by Chasen is assumed to be an accent phrase boundary when one of the detected accent phrase boundaries drops in a +/- 40 ms period of the morpheme boundary in question.

Tables 6 through 9 show averaged perplexities for 10 sets

of 50 sentences, when the remaining 453 sentences are used for the training in each set (cross validation). Tables 6 and 7 are the results when the training and perplexity calculation were done for the same speaker (speaker closed), while Tables 8 and 9 are those when they were done for different speakers (speaker open). Although the rates of perplexity reduction from the baseline model to the proposed model came smaller as compared to the case when the correct accent boundary information was used, they still indicate the validity of the proposed method. It is interesting the Table 8's results are the best among the results of four tables.

Table 10 shows the result when the accent phrase boundary labels of the corpus were used only for the training. The perplexity increased a lot when the proposed model was used. This result indicates that, in both phases of training and recognition, we should use the accent boundaries detected in the same criterion even if they contain errors.

Table 6: Perplexities for baseline and proposed models when the accent phrase boundaries automatically detected. Training: 453 sentences by speaker MYI, Evaluation: 50 sentences by speaker MYI.

	All	Intra AP	Inter AP
Baseline model	117.4	57.32	1436
Proposed model	111.7	55.12	1331
Reduction rate	4.84%	3.84%	7.26%

Table 7: Perplexities for baseline and proposed models when the accent phrase boundaries automatically detected. Training: 453 sentences by speaker MHT, Evaluation: 50 sentences by speaker MHT.

	All	Intra AP	Inter AP
Baseline model	117.4	46.54	2082
Proposed model	111.4	44.60	1921
Reduction rate	5.07%	4.17%	7.74%

4. Conclusions

A new method was developed to include accent phrase boundary information into n-gram language modeling. Its validity was proved through perplexity reduction from the baseline. Currently, the method comes more promising, since the proposed language model showed a larger reduction rate (increased to 12.6% from 11.0% in Table 4) when the baseline model was trained for the larger text corpus (newspaper corpus of 6 years). Surely, we further should check how accent phrase boundary detection errors influence wrong hypothesis of the decoding process, and so on. A preliminary speech recognition experiment for the ATR continuous speech corpus indicated slight improvements (around 1 point or more in %) in the word recognition rates.

Table 8: Perplexities for baseline and proposed models when the accent phrase boundaries automatically detected. Training: 453 sentences by speaker MYI, Evaluation: 50 sentences by speaker MHT.

	All	Intra AP	Inter AP
Baseline model	117.4	46.54	2082
Proposed model	110.4	44.41	1893
Reduction rate	5.96%	4.58%	9.10%

Table 9: Perplexities for baseline and proposed models when the accent phrase boundaries automatically detected. Training: 453 sentences by speaker MHT, Evaluation: 50 sentences by speaker MYI.

	All	Intra AP	Inter AP
Baseline model	117.4	57.32	1436
Proposed model	112.8	55.34	1364
Reduction rate	3.93%	3.45%	5.02%

Table 10: Perplexities for baseline and proposed models when the accent phrase boundary labels of the corpus are used for training and those automatically detected are used for perplexity calculation. Speaker MYI's 503 sentences are used for both training and perplexity calculation.

	All	Intra AP	Inter AP
Baseline model	117.0	57.13	1344
Proposed model	141.2	69.18	1601
Reduction rate	-20.7%	-21.1%	-19.2%
Bigram hit rate	95.35%	96.42%	91.68%

5. References

[1] Minematsu, N.; Tsuda, K.; Hirose, K., 2001. Quantitative analysis of F_0-induced variations of cepstrum coefficients. *Proc. ISCA Workshop on Prosody in Speech Recognition and Understanding*, Red Bank, 113-117.

[2] Gallwitzs, F.; Batliner, A.; Buckow, J.; Huber, R.; Niemann, H.; Nörth, E., 1998. Integrated recognition of words and phrase boundaries. *Proc. ICSLP'98*, Sydney, Vol.7, 2883-2886.

[3] Hirose, K.; Iwano, K., 2000. Detection of prosodic word boundaries by statistical modeling of mora transitions of fundamental frequency contours and its use for continuous speech recognition. *Proc. IEEE ICASSP'2000*, Istanbul, Vol.3, 1763-1766.

[4] Nörth, E.; Batliner, A.; Kießling, A.; Kompe, R.; Niemann, H., 2000. VERBMOBIL: The use of prosody in linguistic components of a speech understanding system. *IEEE Trans. Speech & Audio Processing*, Vol.8, No.5, 519-532.

[5] Lee, S.; Hirose, K.; Minematsu, N., 2001. Incorporation of prosodic module for large vocabulary continuous speech recognition. *Proc. ISCA Workshop on Prosody in Speech Recognition and Understanding*, Red Bank, 97-101.

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

[7] http://chasen.aist-nara.ac.jp/