STOCHASTIC MODELING OF PAUSE INSERTION USING CONTEXT-FREE GRAMMAR

Shigeru FUJIO Yoshinori SAGISAKA and Norio HIGUCHI

ATR Interpreting Telecommunications Research Labs.,
2-2, Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-02 Japan

ABSTRACT

In this paper, we propose a model for predicting pause insertion using a stochastic context-free grammar (SCFG) for an input part of speech sequence. In this model, word attributes and stochastic phrasing information obtained by a SCFG trained using phrase dependency bracketings and bracketings based on pause locations are used. Using the Inside-Outside algorithm for training, corpora with phrase dependency brackets are first used to train the SCFG from scratch. Next, this SCFG is re-trained using the same corpora with bracketings based on pause locations. Then, the probabilities of each bracketing structure are computed using the SCFG, and these are used as parameters in the prediction of the pause locations. Experiments were carried out to confirm the effectiveness of the stochastic model for the prediction of pause locations. In test with open data, 85.2% of the pause boundaries and 90.9% of the no-pause boundaries were correctly predicted.

1. Introduction

The control of pause insertion is very important to give natural prosody to synthetic speech. Though pause insertion is closely related to phrase dependency structure and has stochastic characteristics[1], no computational model has been proposed for the statistical training directly from pause insertion characteristics. In this paper, a stochastic modeling of pause insertion is proposed for Japanese and the experimental results of pause location prediction using this model are shown.

2. A model for the prediction of pause locations

Phrase dependency structures and part of speech (POS) before and after boundaries have commonly been used for predicting pause locations. The automatic decision of a phrase dependency structure is difficult as phrase dependency structures are decided not only by syntax but also by semantics. In this paper, stochastic information of the phrase dependency structure is used in a model for the prediction of pause locations. To obtain this structural information, phrase dependency structures and pause locations are first learned using a SCFG. Next, the parameters representing stochastic information about phrase dependency structures are computed using the probability of production rules in the SCFG. Then, these parameters and POS before and after the boundaries are used in the model for the prediction of pause locations. This model predicts pause locations using a feed-forward type neural network, which will be described in section 2.3..

2.1. Learning pause locations using SCFG

For training of the SCFG, the Inside-Outside algorithm was used[2], applied to partially bracketed text corpora[3]. To predict pause locations using a SCFG, the information of pause location must be captured using SCFG. We have applied this method to learn a prosodic phrase dependency structure using this algorithm[4][5]. Using this method for training the SCFG, the sentential speech data set are given the following two types of bracketing information.

- Phrase dependency bracketing
  The phrase dependency structure and part of speech sequence were hand-tagged by trained transcribers according to conventional Japanese school grammar. As the hand-tagged phrase dependency structure is determined using both syntactic and semantic relations, we only expect that syntactic information implicitly manifested in this bracketing will be captured through SCFG training.

- Pause bracketing
  Only accent phrase boundaries and pause boundaries were considered in this second level of bracketing. By listening to speech data and observing the analyzed F_0 contour, accent phrase-sized units were manually bracketed. Prosodic bracketing was obtained by grouping all the constituents segmented by pause boundaries.

Corpora with phrase dependency brackets are first used to train the SCFG from scratch. Next, this SCFG is

0-7803-2431-5/95 $4.00 © 1995 IEEE

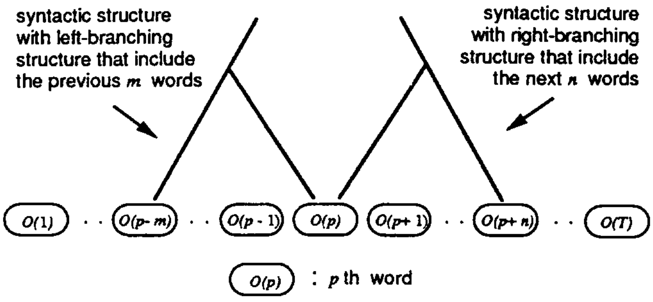
syntactic structure with left-branching structure that include the previous m words
syntactic structure with right-branching structure that include the next n words
O(1)
O(p-m)
O(p-1)
O(p)
O(p+1)
O(p+n)
O(T)
O(p) : p th word

Figure 1. syntactic structure

re-trained using the same corpora with pause brackets.

2.2. Parameters for the prediction of pause locations

For the prediction of the pause locations, the parameters which represent phrase dependency structures are needed. As shown in Figure 1, the probability that the word is part of a left-branching structure which includes the previous m words [left-branching probability P_m] and the probability that the word is part of a right-branching structure which includes the next n words [right-branching probability Q_n] are required. The probabilities represent phrase dependency structures, and can be calculated using a SCFG in the following fashion. Let $a[i,j,k]$ be the probability that the non-terminal symbol i will generate the pair of non-terminal symbols j and k. Let $b[i,m]$ be the probability that the non-terminal symbol i will generate a single terminal symbol m. In the Inside-Outside algorithm, the inner probability $e(s,t,i)$ is defined as the probability of the non-terminal symbol i generating the observation $O(s),\cdot\cdot,O(t)$ and can be expressed as follows:

CASE 1: $s = t$

$$e(s,s,i) = b[i,O(s)]$$

CASE 2: $s \neq t$

$$e(s,t,i) = \sum_{j,k}\sum_{r=s}^{t-1} a[i,j,k]e(s,r,j)e(r+1,t,k)$$

And the outer probability $f(s,t,i)$ is the probability that, in the rewrite process, i is generated and that the strings not dominated by it are $O(1),\cdot\cdot,O(s-1)$ to the left and $O(t+1),\cdot\cdot,O(T)$ to the right. Hence:

$$\begin{aligned} f(s,t,i) &= \sum_{j,k}\Big[\sum_{r=1}^{s-1} f(r,t,j)a[j,k,i]e(r,s-1,k) \\ &\quad + \sum_{r=t+1}^{T} f(s,r,j)a[j,i,k]e(t+1,r,k)\Big] \end{aligned}$$

$$and \;\; f(1,T,i) = \begin{cases} 1 & \text{if } i = S(\text{start symbol}) \\ 0 & \text{otherwise} \end{cases}$$

The non-terminal symbol i can have two possible rules $j \rightarrow ik$ or $j \rightarrow ki$. $f_l(s,t,i)$ is the probability when only rules $j \rightarrow ik$ are considered and $f_r(s,t,i)$ is the probability when only rules $j \rightarrow ki$ are considered. These are expressed as follows:

$$f_l(s,t,i) = \sum_{j,k}\sum_{r=t+1}^{T} f(s,r,j)a[j,i,k]e(t+1,r,k)$$

$$f_r(s,t,i) = \sum_{j,k}\sum_{r=1}^{s-1} f(r,t,j)a[j,k,i]e(r,s-1,k)$$

The probability that the observation $O(1),\cdot\cdot,O(T)$ has a left-branching structure which includes the observation $O(s),\cdot\cdot,O(t)$, and the probability that the observation $O(1),\cdot\cdot,O(T)$ has a right-branching structure which includes the observation $O(s),\cdot\cdot,O(t)$ are respectively given as follows:

$$\sum_i e(s,t,i)f_l(s,t,i)$$

$$\sum_i e(s,t,i)f_r(s,t,i)$$

The probability generated for the entire observation $O(1),\cdot\cdot,O(s),\cdot\cdot,O(t),\cdot\cdot,O(T)$ is given $e(1,T,S)$. Therefore, P_m and Q_n at the pth word are given as follows:

$$P_m = \frac{\sum_i e(p-m,p,i)f_l(p-m,p,i)}{e(1,T,S)}$$

$$Q_n = \frac{\sum_i e(p,p+n,i)f_r(p,p+n,i)}{e(1,T,S)}$$

2.3. Pause prediction model using a neural network

A model for the prediction of pause locations is represented by a feed-forward type neural network. The structure of this neural network has four layers: an input layer with 50 units, two hidden layers with 25 units, and an output layer with 2 units. The input parameters consist of left-branching probabilities P_m and right-branching probabilities Q_n for each input unit (based on POS labels after clustering particles with content words). The terminal symbol of words around the boundary as follows.

- P_m and Q_n at the following words (m,n=1,2,3,4 and sum of more than 5)
  - The content word preceding the word before the boundary

Table 1: Comparison of compatibility scores to brackets based on pause locations between SCFGs obtained from different learning procedure

corpus	SCFG learning using	
	linguistic brackets only	linguistic brackets & brackets based on pause locations
training corpus	89.27 %	91.33 %
test corpus	86.99 %	88.36 %

  - The word before the boundary
  - The word after the boundary
  - The content word following the word after the boundary

- The terminal symbols of the 5 words preceding the boundary
- The terminal symbols of the 5 words following the boundary

3. Experiments

To evaluate the effectiveness of this model, three kinds of experiments were conducted. One is the comparison of compatibility scores to brackets based on pause locations to measure the consistency between the pause locations derived from the trained SCFG and the pause locations of natural utterances. Another is an analysis of the relation between pause insertion and parameters which are obtained using the SCFG. The other is the prediction accuracy of pause location using natural utterances.

3.1. Effect of learning pause locations using SCFG

A sentential speech data set consisting of 503 sentences[6] was used as the corpora for training the SCFG. The SCFG used 20 non-terminal symbols and POS with seven categories of case particles as terminal symbols. To confirm the effectiveness of training using two types of bracketing, we compared a SCFG learned using corpora with phrase dependency brackets only with a SCFG learned using corpora with phrase dependency brackets and brackets based on pause locations. For the comparison, compatibility scores[3] were calculated to measure the consistency. As shown in Table 1, the results, indicate that the possibility of learning pause locations using SCFG. Phrase dependency brackets give reasonable prosodic structuring, and the additional prosodic bracketing information increases the adequacy of SCFG for the prosodic phrase structure reflecting pausing characteristics.

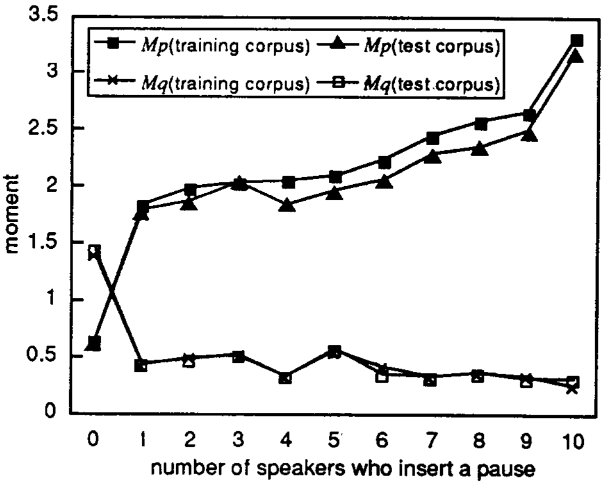

Figure 2: Linear moment of P_m, Q_n at each number of speakers who insert a pause

3.2. Analysis of the relation between pause insertion and P_m, Q_n

As P_m (Q_n) is the probability that the word is part of a left(right)-branching structure which include the previous m (next n) words, the linear moment of P_m (Q_n) [$Mp(Mq)$] represent left(right)-side phrase characteristics. As the pause insertion relate to left(right)-side phrase structure, a simple pause insertion model can be represent by this linear moment. Though it is estimated that the proposed model using a neural network are represent similar and more complex model, it is difficult to accurately represent the structure of this model. Therefore, the linear moment of P_m, Q_n was used in this analysis of the relation between pause insertion and P_m, Q_n. To investigate the relation between Mp, Mq and pause insertion, boundaries were classified into eleven categories by the number of speakers who insert pause, and the average of these moments is calculated for each categories. As shown in Figure 2, the results show Mp increases in proportion to the the number of speakers who insert a pause. In particular, Mp of boundaries at which all speakers insert a pause has very large score, Mp of boundaries where no speaker inserts a pause has very small score. And Mq of boundaries at which someone inserts a pause is smaller than Mq of boundaries where no speaker inserts a pause. These results show that the boundaries will can be classified pause boundary and no pause boundary by difference of Mp, Mq. And the pause prediction model using P_m, Q_n will be effective.

Table 2. Prediction results

used parameter	corpus	percentage of correct prediction	
		A+B	A
			B
P_m, Q_n POS	training corpus	95.0% (5351/5633)	99.7% (371/372)
			94.7% (4980/5261)
	test corpus	89.8% (5061/5633)	85.2% (317/372)
			90.9% (4744/5261)
POS only	training corpus	82.5% (4645/5633)	93.3% (346/372)
			81.7% (4299/5261)
	test corpus	81.5% (4590/5633)	86.8% (322/372)
			81.1% (4268/5261)

A: boundaries where all speakers insert a pause
B: boundaries where no speaker inserts a pause

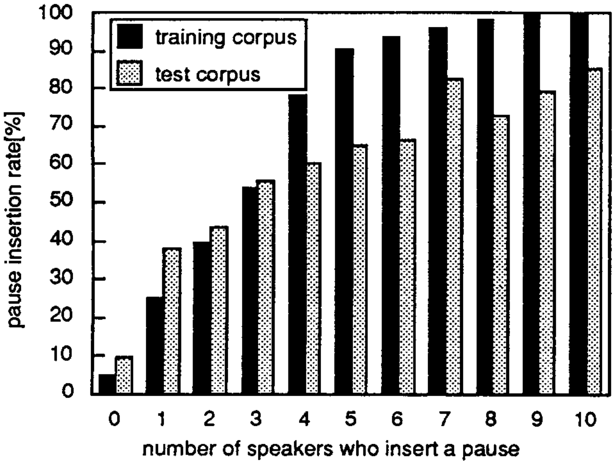

Figure 3: Pause insertion rate at each number of speakers who insert a pause

3.3. Accuracy of the prediction of pause locations

Experiments using the model for the prediction of pause locations were carried out to confirm the effectiveness of the model. The model was trained using 7002 samples. As the pause location is different among speakers, ten speaker's utterances were analyzed. And the boundaries are classified to eleven categories by the number of speakers who insert pause. As the it is important to predict boundaries where all speakers insert a pause and boundaries where no speaker inserts a pause, the evaluation of prediction results was carried out for the 371 boundaries where all speakers insert a pause and the 5261 boundaries where none of them insert a pause. For the calculation of prediction score, all word boundaries are used as boundary candidates. The result is shown in Table 2. To evaluate the effect of the use of P_m, Q_n, experimental results of the prediction using POS only are shown in same table. Table 2 shows that pause locations were predicted successfully using the model which was controlled by the parameters obtained by a SCFG, and this model will be effective for the prediction of pause locations. The prediction results for each number of speakers who insert pause are shown in Figure 3. Figure 3 shows that the pause insertion rate increases in proportion to the number of speakers who insert a pause, the prediction using this proposed model is reasonable in the result for boundaries where someone insert a pause. Though the prediction of no-pausing boundaries look quite successful in percentage terms, the false alarm rates calculated over possible phrase boundaries are still high (about 24%).

4. Conclusion

We propose a model for predicting pause insertion using a stochastic context-free grammar (SCFG) for an input word sequence without any information on phrase structure. Our results show that the proposed model will be effective for the prediction of pause locations. In the future, we will carry out experiments for predicting some kinds of pause locations. As the false alarm rate is high, we will consider the reduction of this rate.

Acknowledgments

We would like to thank Dr. Y. Schabes and Dr. F. Pereira for providing the program for Inside-Outside training.

References

[1] N.Kaiki,Y.Sagisaka :"Pause characteristics and local phrase dependency structure in Japanese", Proc. ICSLP, Vol.1 pp.357-360,1992

[2] K.Lari,S.J.Young :"The estimation of stochastic context-free grammars using the Inside-Outside algorithm", J.Computer Speech and Language, Vol.4 pp.35-56,1990

[3] F.Pereira,Y.Schabes :"Inside-outside reestimation from partially bracketed corpora", Proc. ACL, pp.128-135,1992

[4] Y.Sagisaka,F.Pereira :"Inductive learning of prosodic phrasing characteristics using stochastic context-free grammar", The Acoustic Society of Japan Spring Meeting Proc., 2-8-10(1994,3)

[5] S.Fujio,Y.Sagisaka,N.Higuchi :"Prediction of prosodic phrase boundaries using stochastic context-free grammar", Proc. ICSLP94, Vol.2 pp.839-842,1994

[6] Y.Sagisaka, K.Takeda, M.Abe, S.Katagiri, T.Umeda, H.Kuwabara : "A Large-Scale Japanese Speech Database", Proc. ICSLP90, Vol.2 pp.1089-1092,1990