Recovery of Japanese Dependency Structure Using Multiple Pause Information

Meirong Lu Kazuyuki Takagi Kazuhiko Ozeki

The University of Electro-Communications, Tokyo 182-8585, Japan
{lumeirong,takagi,ozeki}@ice.uec.ac.jp

Abstract

This paper is concerned with the problem of exploiting pause information for recovering dependency structures of read Japanese sentences. In our past work, two kinds of pauses were employed: post-phrase pause that immediately succeeds a phrase in question, and post-post-phrase pause which immediately succeeds the phrase that follows a phrase in question. It was found that simultaneous use of two kinds of pause information improves the parsing accuracy compared to the case where only the post-phrase pause is used. In this paper, we employed yet another kind of pause (pre-phrase pause), which immediately precedes a phrase in question. By combining the three kinds of pause information appropriately, the parsing accuracy was further improved compared to the case where the post-phrase pause and the post-post-phrase pause were used as in our previous work.

1. Introduction

It is well known that prosody contains certain amount of syntactic information [1∼5]. In the field of speech synthesis, many research works have been conducted to control prosody so that it conforms to the syntactic structure of the sentence [6]. We have been working for several years on the inverse problem: recovery of syntactic structure with the help of prosody. In our past work, we have tried various prosodic features and found that the duration of pause which immediately succeeds a phrase in question (post-phrase pause) provides very effective information for parsing [7∼10]. Encouraged by this fact, we tried another kind of pause: post-post-phrase pause which immediately succeeds the phrase that follows a phrase in question. It was confirmed that simultaneous use of post-phrase pause information and post-post-phrase information improves the parsing accuracy compared to the case where only the post-phrase pause information is used [10]. In this paper, we further widen the window to look at pauses, and employ yet another kind of pause: pre-phrase pause which precedes a phrase in question. In the following sections, we will discuss how to appropriately combine the three different kinds of pause information, and how much improvement in parsing accuracy is achieved by such information.

2. Minimum Penalty Parser

Although we use the same parsing method as used in our past works, a brief overview is given here for self-containedness of the paper.

2.1. Dependency structure and parsing

A Japanese sentence is a sequence of phrases, where a phrase is a syntactic unit called *bunsetsu* (hereafter simply referred to as "phrase") in Japanese, consisting of a content word, or a string of content words, followed by (possibly zero) function words such as particles and auxiliary verbs.

From a dependency grammatical point of view, the structure of a Japanese sentence can be described by specifying which phrase modifies which phrase in the sentence. Thus the syntactic structure of a sentence $w_1 w_2 \dots w_m$, represented as a sequence of phrases, is described by specifying a function S that maps a modifier phrase to its modified phrase:

$$S : \{1, 2, \dots, m-1\} \longrightarrow \{2, 3, \dots, m\}.$$

The function S must satisfy certain conditions, reflecting the syntactic properties of Japanese. A function that satisfies those constraints is referred to as a *dependency structure* on $w_1 w_2 \dots w_m$. For phrases w_i, w_j, the number $j - i$ is the *distance* between them. Under a dependency structure S, $S(i) - i$ is called the *dependency distance* between w_i and $w_{S(i)}$, or simply dependency distance of w_i.

In our method, linguistic knowledge concerning a modifier phrase w_i and a modified phrase w_j is represented by a function $F(w_i, w_j)$ that measures the amount of penalty when a phrase w_i is to modify a phrase w_j. The parser finds a dependency structure S that minimizes the total penalty $\sum_{i=1}^{m-1} F(w_i, w_{S(i)})$ [7].

2.2. Penalty function based on prosodic features

A dependency structure S is determined by specifying the dependency distance $S(i)-i$ of each phrase w_i. Thus any

information related to the dependency distance must be useful for parsing. For this reason, the penalty function $F(w_i, w_j)$ is defined on the basis of statistical knowledge about the prosodic features associated with w_i and its dependency distance.

Let d be the dependency distance of a phrase in a sentence, and $\boldsymbol{p} = (p_1, \ldots, p_n)$ the prosodic feature vector associated with the phrase. The conditional probability of d given $\boldsymbol{p}$ is denoted by $P(d \mid \boldsymbol{p})$, which is represented by the Bayes theorem as

$$P(d \mid \boldsymbol{p}) = \frac{P(\boldsymbol{p} \mid d)P(d)}{\sum_d P(\boldsymbol{p} \mid d)P(d)}.$$

Thus, $P(d \mid \boldsymbol{p})$ can be calculated from $P(\boldsymbol{p} \mid d)$ and $P(d)$. $P(d)$ is estimated as $P(d) = N_d / \sum_d N_d$, where N_d is the number of phrases with dependency distance d. Then the penalty function $F(w_i, w_j)$ is defined as

$$F(w_i, w_j) = \begin{cases} -\log P(j - i \mid \boldsymbol{p}), & \text{if } (w_i, w_j) \in DR \\ \infty, & \text{otherwise} \end{cases}$$

where $\boldsymbol{p}$ is the prosodic feature vector associated with w_i, and $(w_i, w_j) \in DR$ signifies that w_i is allowed to modify w_j by local syntactic constraints, or a *dependency rule*, which is constructed on the basis of the morphological structure of phrases [7].

3. Dependency Distance and Pause Duration

3.1. Multiple pauses

Fig. 1 illustrates three kinds of pauses used in this paper: the pre-phrase pause, the post-phrase pause and the post-post-phrase pause. The duration of the pre-phrase pause, the post-phrase pause and the post-post-phrase pause are denoted by p_0, p_1 and p_2, respectively. As mentioned before, p_1 and p_2 were used in our previous work, while p_0 is newly introduced this time.

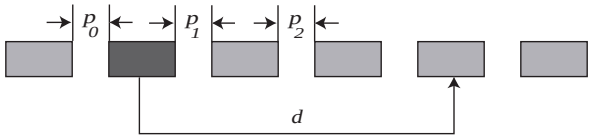

Figure 1: *The pre-phrase pause duration p_0, the post-phrase pause duration p_1, the post-post-phrase pause duration p_2. The dark box is the phrase in question, which modifies the phrase pointed by the arrow. The dependency distance between the two phrases is d.*

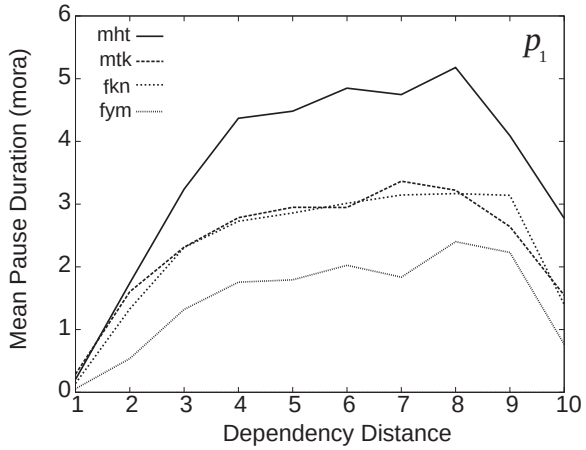

Figure 2: *Mean duration of the post-phrase pause as a function of dependency distance d for 4 speakers.*

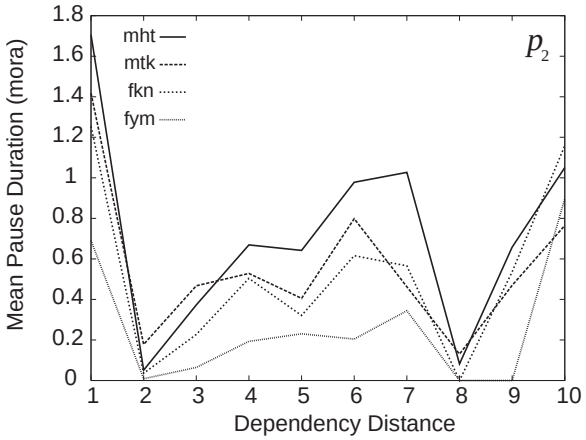

Figure 3: *Mean duration of the post-post-phrase pause as a function of dependency distance d for 4 speakers.*

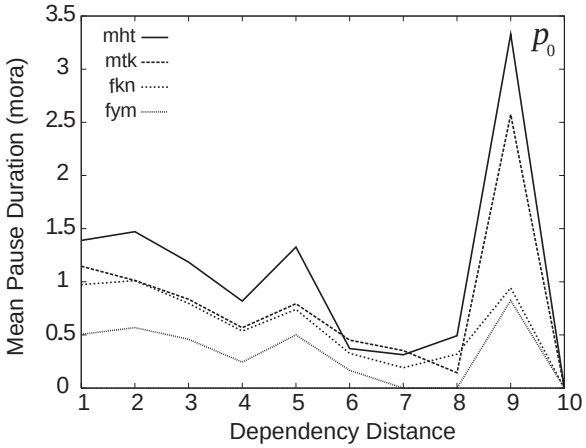

Figure 4: *Mean duration of the pre-phrase pause as a function of dependency distance d for 4 speakers.*

Fig. 2 shows the mean value of p_1 as a function of the dependency distance d of the phrase in question. The unit of measurement for the pause duration is the mean mora duration of the utterance in which the pause appears. The curves clearly show a close statistical relationship between d and p_1. It is particularly noted that

the mean value of p_1 increases almost linearly with d up to $d = 4$ for every speaker, though the slope differs from speaker to speaker. For $d \geq 5$ the curves saturate, and for $d \geq 7$, the mean pause duration is probably unreliable because there are only very small number of tokens for long dependency distances.

Similarly, Fig. 3 illustrates the mean value of p_2. We note some remarkable phenomena on the graphs. There is a clear dip at $d = 2$, which is explained as follows. Let $\cdots w_i w_{i+1} w_{i+2} \cdots$ be a sentence represented as a sequence of phrases, and let w_i be the phrase in question. Then $d = 2$ means that w_i modifies w_{i+2}. In that case w_{i+1} must modify w_{i+2} to satisfy the non-crossing constraint of dependency. So the dependency distance of w_{i+1} equals 1, and according to Fig. 2, the pause between w_{i+1} and w_{i+2}, i.e. p_2, must be very short. There is another dip at $d = 8$. This is accidental and unreliable due to inadequate amount of tokens. An important fact is that the mean value of p_2 has a systematic tendency to increase for $2 \leq d \leq 6$, showing a statistical relationship between p_2 and d.

Fig. 4 is for the mean value of p_0. On the whole, p_0 exhibits moderate variations compared with p_1 and p_2. It also shows a decreasing tendency with d up to $d = 8$. As to data for $d \geq 8$, the result will lack statistical reliability.

3.2. Penalty function based on multiple pause information

One-dimensional Gaussian p.d.f.'s $P(p_0 \mid d)$, $P(p_1 \mid d)$ and $P(p_2 \mid d)$ were fitted to p_0, p_1 and p_2 data respectively for each value of d. Then $P(d \mid p_0)$, $P(d \mid p_1)$ and $P(d \mid p_2)$ were calculated by using Eq.(2.2), and then linearly combined to define $F(w_i, w_j)$. Thus Eq.(2.2) is modified as follows:

$$F(w_i, w_j) = \begin{cases} -\{\alpha_0 \log P(d \mid p_0) \\ \qquad +\alpha_1 \log P(d \mid p_1) \\ \qquad\quad +\alpha_2 \log P(d \mid p_2)\}, \\ \qquad\qquad\qquad \text{if } (w_i, w_j) \in DR \\ \infty, \qquad\qquad\qquad \text{otherwise} \end{cases}$$

where $d = j - i$, and $\alpha_0, \alpha_1, \alpha_2$ (α_0+α_1+α_2=1) are weighting factors to adjust the contribution of p_0, p_1 and p_2. In this work, the values of the weighting factors were determined experimentally. That is, with a small step, every combination of values of α_0, α_1 and α_2 were tried, and the ones which give the highest parsing accuracy were finally selected.

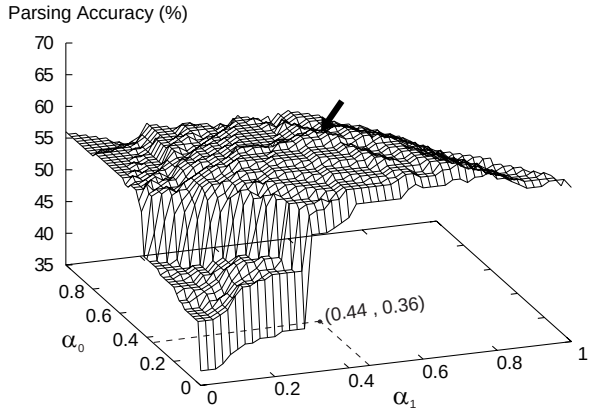

Figure 5: *Parsing accuracy as a function of weighting factors α_0 and α_1 for speaker mht. Data used is Set(i).*

4. Dependency Analysis Experiments

4.1. Training data and test data

An ATR speech database (Set B) [11] was used in this work. This database contains 503 Japanese sentences, grouped into A – J. In the following experiments, those groups were divided into training data and test data as in Table 1. All the experiments in this paper are speaker-dependent. Results of analysis were evaluated by *parsing accuracy*, i.e., the percentage of test sentences whose dependency structures determined by parsing are the same as those described in the database, and also by *dependency accuracy*, i.e., the percentage of phrase pairs in the test sentences whose dependency relations determined by parsing are the same as those described in the database.

4.2. Results of dependency analysis

Fig. 5 shows how the parsing accuracy changes with α_0 and α_1 for speaker mht. Since $\alpha_0 + \alpha_1 + \alpha_2 = 1$, and $\alpha_2 \geq 0$, only the region $\alpha_0 + \alpha_1 \leq 1$ exists on the graph. The point (α_0, α_1)=(1,0) corresponds to the case where only the pre-phrase pause is used. Similarly, the points (α_0, α_1)=(0,1) and (α_0, α_1)=(0,0) correspond to the cases where only the post-phrase pause and the post-post-phrase pause are used, respectively. There appears a maximum of the parsing accuracy for α_0, α_1 between 0 and 1. The arrow on the surface indicates the position of highest parsing accuracy. The broken lines and the dot on (α_0, α_1) plane illustrates the corresponding values of α_0 and α_1, which are the best weighting factors determined experimentally in this work.

Table 2 shows parsing accuracy, averaged over Set(i), Set(ii) and Set(iii), for $\alpha_1 = 1$ (p_1 only), $\alpha_0 = 0$ (p_1 and p_2) and for optimum α_0, α_1 and α_2 (p_0, p_1 and p_2). "Dist" means a case where only the distribution of dependency distance is used to define the penalty function: $P(j - i \mid \boldsymbol{p})$ in Eq.(2.2) is replaced with $P(j - i)$. Thus by the use of p_1, the parsing accuracy was improved by

Table 1: *Training data and test data*

Set	training data	test data
i	D–J (353 snt.)	A–C (150 snt.)
ii	A–G (350 snt.)	H–J (153 snt.)
iii	A–C, G–J (353 snt.)	D–F (150 snt.)

Table 2: *Parsing accuracy* $(\%)$ *using pause information. "Dist" means a case where only dependency distance distribution is used.*

Cond.	mht	mtk	fkn	fym	Av.
p_1	58.0	57.8	55.4	54.5	56.4
$p_1\ p_2$	60.7	59.2	57.6	56.5	58.5
$p_0\ p_1\ p_2$	61.4	59.6	58.9	57.1	59.3
Dist					54.5

Table 3: *Dependency accuracy* $(\%)$ *for each dependency distance. Data used is Set(i).*

d	Cond.	mht	mtk	fkn	fym
1	p_1	94.5	94.5	94.9	96.3
	$p_1\ p_2$	94.3	94.9	95.1	96.5
	$p_0\ p_1\ p_2$	94.9	95.2	95.4	96.5
2	p_1	82.6	86.3	81.4	74.5
	$p_1\ p_2$	87.0	88.2	85.7	83.2
	$p_0\ p_1\ p_2$	87.6	88.2	86.3	81.4
3	p_1	89.2	86.5	86.5	83.8
	$p_1\ p_2$	89.2	83.8	86.5	81.1
	$p_0\ p_1\ p_2$	89.2	85.1	87.8	83.8
4	p_1	80.0	80.0	73.3	76.7
	$p_1\ p_2$	76.7	73.3	76.7	70.0
	$p_0\ p_1\ p_2$	76.7	76.7	76.7	73.3

1.9 points compared to Dist case, and another 2.1 points improvement was achieved by adding p_2. The parsing accuracy was further improved by 0.8 points on average by using p_0, which was newly employed in this work.

Table 3 shows the dependency accuracy for each dependency distance when Set(i) was used. There are improvements mainly at both $d = 1$ and $d = 2$ by adding p_0 information, which shows that the pre-phrase pause does have the information about the dependency distance of the phrase in question. The effective range is, however, limited to comparatively short dependency distance. Note that for dependency distances higher than 2, the use of extra pauses reduced the dependency accuracy. The reason has not been clarified yet.

5. Conclusion

This paper introduced a new kind of pause and used it in dependency analysis of spoken Japanese sentences. It was shown that by using the pre-phrase pause, post-phrase pause and the post-post-phrase pause simultaneously, the parsing accuracy was improved further compared to the case where only the post-phrase pause was used, or the case where the post-phrase pause and the post-post-phrase pause were used as in our past work. It was proved that the pre-phrase pause duration has the information about the dependency distance of the phrase in question, in spite of the fact that the pre-phrase pause comes before the phrase in question. Improvement of dependency accuracy was attained mainly when dependency distance is 1 or 2, which showed a limited effective distance range of the pre-phrase pause.

Our future work will focus on finding better statistical models of the distribution of pause duration and more effective methods of combining multiple pause information. Also, automatic determination of optimum weighting factors for combining multiple pause information is our another challenge.

6. References

[1] T. Uyeno, H. Hayashibe, K. Imai, H. Imagawa, and S. Kiritani, 1981. "Syntactic structure and prosody in Japanese: a study on pitch contours and the pauses at phrase boundaries," Annual Bulletin of Research Institute of Logopedics and Phoniatrics, University of Tokyo, Vol.15, 91–108.

[2] A. Komatsu, E. Ohira, and A. Ichikawa, 1988. "Conversational speech understanding based on sentence structure inference using prosodics, and word spotting," IEICE Trans., Vol.J71-D, No.7, 1218–1228.

[3] N. M. Veilleux and M. Ostendorf, 1993. "Probabilistic parse scoring with prosodic information," Proc. ICASSP'93, Vol.II, 51–54.

[4] Y. Sekiguchi, Y. Suzuki, T. Kikukawa, Y. Takahashi, and M. Shigenaga, 1995. "Existential judgement of modifying relation between successively spoken phrases by using prosodic information," IEICE Trans., Vol.J78-D-II, No.11, 1581–1588.

[5] T. Ohsuga, Y. Horiuchi, T. Umeda, A. Ichikawa, 2003. "Estimating Syntactic Structure from Prosody in Japanese Speech," IEICE TRANS, Vol.E86-D, NO.3, 558–564.

[6] N. Kaiki and Y. Sagisaka, 1996. "Study of pause insertion rules based on local phrase dependency structure," IEICE Trans., Vol.J79-D-II, No.9, 1455–1463.

[7] N. Eguchi and K. Ozeki, 1996. "Dependency analysis of Japanese sentences using prosodic information," J. Acoust. Soc. of Japan, Vol.52, No.12, 973–978.

[8] K. Ozeki, K. Kousaka, and Y. Zhang, 1997. "Syntactic information contained in prosodic features of Japanese utterances," Proc. Eurospeech'97, Vol.3, 1471–1474.

[9] Y. Hirose, K. Ozeki, and K. Takagi, 2000. "Effectiveness of prosodic features in syntactic analysis of read Japanese sentences," Proc. ICSLP2000, Vol.3, 215–218.

[10] M. Lu, K. Takagi, and K. Ozeki, 2003. "The use of multiple pause information in dependency structure analysis of spoken Japanese sentences," Proc. Eurospeech2003, 3173–3176.

[11] M. Abe, Y. Sagisaka, T. Umeda, and H. Kuwabara, 1990. "Manual of Japanese Speech Database," ATR.