Dependency Analysis of Read Japanese Sentences using Pause Information: A Speaker Independent Case

Kazuyuki Takagi Kazuhiko Ozeki

The University of Electro-Communications, Tokyo 182-8585, Japan

{takagi,ozeki}@ice.uec.ac.jp

Abstract

This paper deals with the problem of recovering syntactic structures of sentences by using the prosodic information extracted from spoken versions of the sentences. Prosodic information has proven to be effective to disambiguate syntactic structures, which is not utilized in a conventional rule-based parser. In our previous works, the duration of pauses at phrase boundaries has been found to be consistently and dominantly effective for improving parsing accuracy of read Japanese sentences, although the evaluation was limited to a small set of test speakers. In this paper, dependency analysis using pause information was conducted in a speaker-independent manner by using larger amount of speech data read by 316 speakers. Effects of pause duration normalization were observed, although the parsing accuracy was lower than that in speaker-dependent case.

1. Introduction

Prosody and syntax are closely related to each other. In the field of speech synthesis, many works have been published on prosody control based on the syntactic structure of a sentence [1, 2]. This paper is concerned with the inverse problem of recovering syntactic structure based on prosodic information. Researches related to this problem can be found in the literature [3, 4, 5, 6]. However very little work has been done to incorporate prosodic information directly into a parser as linguistic knowledge, and exploit it in the search process.

Eguchi and Ozeki presented a method of incorporating prosodic information into a Japanese dependency structure parser [7]. The parser can handle both symbolic information such as syntactic rule and numerical information such as probability of dependency distance in a unified way as linguistic information. The work has been further extended by increasing the number of prosodic features [8, 9]. As results of our previous work in which an optimal combination of these features was sought for [9], the duration of pauses at phrase boundaries has been found to be most effective. So the use of pause information has been being pursued [10]. However the evaluation of the method and effectiveness of various prosodic features were limited to a speaker-dependent case that uses up to 10 speakers. In this paper, dependency analysis was conducted speaker-independently, with the use of much larger amount of speech data.

2. Inter-phrase dependency distance

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 followed by (possibly zero) function words. Let $w_1 w_2 \ldots w_m$ be a sentence represented as a sequence of phrases. If w_i modifies w_j, then $j - i$ is referred to as the *dependency distance* of w_i. From a dependency grammatical point of view, the structure of a Japanese sentence can be determined by specifying the dependency distance of each phrase in the sentence except for the last phrase in the sentence. Thus any information related to the dependency distance is expected to be useful for dependency structure analysis.

3. Minimum penalty parsing

Although the basic framework of dependency structure parsing is the same as in our previous papers [7, 8, 9, 10], a brief overview is given here for self-containedness of the paper.

3.1. Parser

The dependency structure of a sentence $w_1 w_2 \ldots w_m$, represented as a sequence of phrases, is determined by specifying a function S that maps a modifier phrase to the modified phrase:

$$S : \{1, 2, \ldots, m-1\} \rightarrow \{2, 3, \ldots, m\}.$$

Reflecting syntactic properties of the Japanese language, the function S must satisfy the following constraints:

- $\forall i \in \{1, 2, \ldots, m-1\} : i < S(i)$
- $\forall i, j \in \{1, 2, \ldots, m-1\} :$
  $i < j \Rightarrow (S(i) \leq j \ or \ S(j) \leq S(i))$.

A function that satisfies these constraints is referred to as a *dependency structure* on $w_1 w_2 \ldots w_m$.

In our parser, linguistic knowledge 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 then searches for a dependency structure S that minimizes the total penalty

$$\sum_{i=1}^{m-1} F(w_i, w_{S(i)})$$

given a sentence $w_1 w_2 \ldots w_m$ [7].

3.2. Penalty function

The penalty function $F(w_i, w_j)$ is defined on the basis of conditional probability of the dependency distance given the prosodic features [7]:

$$F(w_i, w_j) = \begin{cases} -\log P(d \mid \boldsymbol{p}), & \text{if } (w_i, w_j) \in DR \\ \infty, & \text{otherwise} \end{cases} \quad (1)$$

where $d = j - i$, $\boldsymbol{p}$ 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 the local syntactic constraints, or *dependency rule*, which is based on the morphological structure of the phrases.

4. Syntactic information in pause

4.1. Pause duration and dependency distance

Given an utterance, prosodic features associated with a phrase in question w_i are defined on the basis of pause duration, log-power contour, log-F_0, speaking rate, etc. Many of the features are defined relative to the immediately succeeding phrase w_{i+1}. The duration of pauses at phrase boundaries has been found to be consistently and dominantly effective for improving parsing accuracy in our previous work [9].

The pause duration of a phrase in question is defined as the time interval between the ending point of the phrase and the starting point of the immediately succeeding phrase. Fig. 1 illustrates the mean pause durations for 10 speakers in the ATR 503 Phonetically Balanced Sentences [11], as functions of the dependency distance. The mean pause duration grows linearly with the dependency distance up to $d = 4$, though the slope depends on the speaker. This shows that the duration of pause contains information about dependency distance.

4.2. Normalization of pause duration

The use of pause information has been pursued [10], but the experiments were limited to speaker-dependent cases that use up to 10 speakers in ATR 503 Phonetically Balanced Sentences [11]. In this paper, effectiveness of pause information is examined in speaker-independent cases: the prosodic model $P(d \mid \boldsymbol{p})$ is trained speaker-independently with the use of much larger amount of speech data.

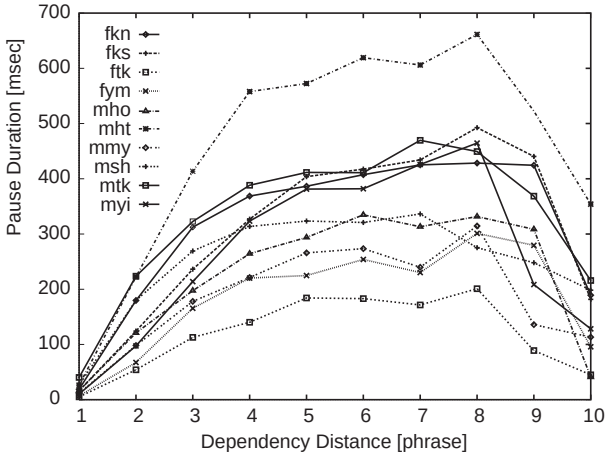

Figure 1: *Mean pause duration for 10 speakers as functions of the dependency distance.*

Speakers in JNAS corpus [12] exhibit a diverse range of speech rate from 4.47 to 9.73 [mora/sec], while the speech rate of professional speakers in ATR database ranges between 6.04 and 7.94 [mora/sec]. So in the following experiments, normalization of pause duration by average speech rate is tested. The average speech rate [mora/sec] was first measured over one sentence, then duration of pauses in the sentence is divided by the speech rate to give the normalized pause duration.

5. Speaker independent analysis

5.1. JNAS corpus

In our previous works [7, 8, 9, 10], the experimental data used was limited to the ATR 503 Phonetically Balanced (ATR 503PB) sentences spoken by 10 speakers [11], because no other databases provide a complete set of acoustic labels, linguistic information, syntactic structure labels, hand-corrected F_0 values, and speech waveforms. Although the quality of the database is high, the insufficiency of sentence structure types and the number of speakers may limit the reliability of evaluation of our result. So databases that have wider variety of sentence structures and larger population of speakers are needed in order for a further study. JNAS corpus [12] has a much larger set of 306 speakers, although the set of sentences is still limited to the ATR 503PB sentences. Each speaker read aloud one of the 10 sentence groups of ATR 503PB sentences. In this paper, all speech data of ATR 503PB sentences in JNAS corpus is used in addition to the speech data of the ATR database. The reading text and linguistic information in ATR database can also be used for analysis of ATR 503PB sentences in JNAS corpus.

5.2. Measuring pause duration

There is no phoneme label information in JNAS corpus, by which duration of pauses, phrases, and sentence are measured. Automatic phoneme labeling was performed by forced alignment using HTK[13]. Acoustic models were gender-dependent triphones created in the previous work[14]: 1294 triphones for male, 1177 triphones for female. The training data of the models contains the same speaker set of the experiment of this paper. The number of Gaussian mixture components for each state is 16. The performance of the HTK decoder using this set of triphone models is 84.58% for male speakers and 87.88% for female speakers in word accuracy for ATR 503PB sentences.

Phrase boundaries were determined and then duration of pauses between phrases were measured, by matching the aligned phoneme labels with the linguistic information labels of ATR database. Utterance duration of a sentence was also measured for calculation of average speech rate of a sentence.

6. Experiments

The ATR 503PB database contains 503 sentences extracted from newspapers, journals, novels, letters, textbooks, etc., which are divided into 10 groups A – J. The sentences have labels that indicate their dependency structures. It also contains the speech waveforms for all the sentences read by professional announcers or narrators. JNAS corpus also contains speech waveforms for ATR 503PB sentences, but each speaker read 50 to 53 sentences of only one group. Then the total number of sentences is 15,372 as in Table 1. For each sentence group, there are 24 to 34 speakers.

Experiments were conducted on various conditions concerning the training and test dataset combination, kind of phoneme labels, speaker dependence as in Table 2. As a baseline, speaker dependent analysis was conducted using only ATR database with its manually corrected phoneme label (SD-ATR), i.e., $P(d \mid \boldsymbol{p})$ was estimated for each speaker. SI-ATR is the condition in which $P(d \mid \boldsymbol{p})$ was estimated from all the speakers' training data. For the conditions using both ATR and JNAS database, automatically generated phoneme labels were used. When ATR data was used as training data, JNAS was used as test data in SI-ATR-JNAS; and vice versa in SI-JNAS-ATR.

In the conditions, SD-ATR and SI-ATR, the sentence groups A – J were divided into training data and test data as in Table 2. Results were averaged over Set(i) and Set(ii). For the other conditions, all the sentences were used for both training and test.

Results of parsing were evaluated by parsing accuracy, i.e., the percentage of test sentences whose dependency structures determined by parsing are exactly the same as those described in the database.

Table 1: *ATR 503 PB sentences data*

ATR 503 PB sentences data		
Speaker	6 male (mxx), 4 female (fxx)	
Sentence	503 sentences, 10 groups: A – J	
Phrase	3426 phrases	
Dataset	Training data	Test data
Set(i)	D – J (353 snt.)	A – C (150 snt.)
Set(ii)	A – G (350 snt.)	H – J (153 snt.)
JNAS's ATR 503 PB sentence data		
Speaker	153 male, 153 female	
Utterances	1 sentence group per speaker 15,372 ($\simeq$ 50 sent. $\times$ 306 speakers)	

Table 2: *Experimental conditions*

Training	Test	Label	SD / SI	Sent.
ATR	ATR	ATR	SD-ATR	open
			SI-ATR	open
ATR	JNAS	HTK	SI-ATR-JNAS	closed
JNAS	ATR	HTK	SI-JNAS-ATR	closed

7. Results

Table 3 shows the parsing accuracy obtained on the various conditions. The parsing accuracy was 49.5 % by the deterministic analysis method [15] or DET in which no prosodic information is used. The parsing accuracy was improved from 49.5% in DET to 54.5% in DIST condition where $P(d \mid \boldsymbol{p})$ is replaced with $P(d)$ in Eq. 1. With the use of pause information in speaker dependent manner in ATR database, the performance was improved to 56.0% (SD-ATR).

In speaker independent cases the performance was lower than that in the speaker dependent case. However, effects of pause duration normalization were observed. In fact, although the parsing accuracy of SI-ATR case was 55.6%, it was improved to 55.9% by normalization, which is comparable to SD-ATR case. The performance did not change by normalization in SI-JNAS-ATR cases. When the new test data from JNAS corpus was used as the test data with no normalization (SI-ATR-JNAS) the results were the worst. This is mainly due to the mismatches of the speech rate between the training (ATR) and the test (JNAS) data. The parsing performance was, however, improved from 51.6% (SI-ATR-JNAS) to 51.9% (SI-ATR-JNAS-N) by normalization of pause duration. Table 4 shows the dependency accuracy, that is, the percentage of dependent phrase pairs that were correctly estimated by the parser. The improvement of parsing accuracy on SI-ATR-JNAS-N was due to the improvements in the analysis of short distances. No correlation between the average speech rate and the parsing accuracy was observed.

Table 3: *Parsing accuracy (%) – DET: deterministic method, DIST: DET+distance distribution information, SD: speaker dependent, SI: speaker independent, ATR: ATR database, JNAS: JNAS database, N: pause normalization by speech rate*

Condition	Parsing accuracy (%)
DET (no prosody)	49.5
DIST (no prosody)	54.5
SD-ATR	56.0
SI-ATR	55.6
SI-ATR-N	55.9
SI-JNAS-ATR	54.2
SI-JNAS-ATR-N	54.2
SI-ATR-JNAS	51.6
SI-ATR-JNAS-N	51.9

Table 4: *Dependency accuracy (%) for d= 1 $\sim$ 4*

Distance (d)	1	2	3	4
SI-ATR-JNAS	95.5	84.8	76.9	61.4
SI-ATR-JNAS-N	95.5	85.0	77.4	61.6

8. Conclusion

In this paper, speaker-independent experiments of dependency analysis using pause information were conducted, by using large amount of speech data read by 316 speakers, with automatically annotated phoneme label and pause durations. Although in speaker-independent case the parsing accuracy was lower than that in the speaker dependent case, effects of pause duration normalization were observed. The overall performance of the experiments was not as good as in our previous works. This should be partly because of the inaccuracy of alignment in automatic labeling stage. So the future work includes elaboration of the phoneme label, better method of preprocessing prosodic features, as well as the use of F_0 information in speaker-independent analysis.

9. References

[1] 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.

[2] N. Kaiki and Y. Sagisaka, 2000, "F_0 control based on local phrase dependency structure," *IEICE Trans.*, Vol. J83-D-II, No. 9, 1853–1860.

[3] 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.

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

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

[6] J. Venditti, S. Jun and M. Beckman, 1996, "Prosodic cues to syntactic and other linguistic structures in Japanese, Korean and English,", J. L. Morgan and K. Demuth (eds.), *Signal to syntax: bootstrapping from speech to grammar in early acquisition* (Hillsdale, NJ: Lawrence Erlbaum), 287–311.

[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, 2001, "Effectiveness of prosodic features in dependency analysis of read Japanese sentences," *Natual Language Processing*, Vol. 8, No. 4, 71–89.

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

[11] Y. Sagisaka and N. Uratani, 1992, "ATR spoken language database," *Journal of The Acoustical Society of Japan*, Vol. 48, No. 12, 878–882.

[12] K. Itou, M. Yamamoto, K. Takeda, and T.Takezawa, T.Matsuoka, T. Kobayashi, K. Shikano, and S. Itahashi, 1998, "The design of the newspaper-based Japanese large vocabulary continuous speech recognition corpus," *Proc. ICSLP98*, Vol. 7, 3261–3264.

[13] "HTK: Hidden Markov Model Toolkit," Version 3.1, http://htk.eng.cam.ac.uk/

[14] K. Takagi, R. Oguro, and K. Ozeki, 2002, "Effectiveness of word string language models on noisy broadcast news speech recognition," *IEICE Trans. Inf. & Syst.*, Vol. E85-D, No. 7, 1130–1137.

[15] S. Kurohashi and M. Nagao, 1994, "A syntactic analysis method of long Japanese sentences based on coordinate structures' detection," *Journal of Natural Language Processing*, Vol. 1, No. 1, 35–57.