A Method for Morphological Analysis of Transcripts from Spontaneously Spoken Dialogue based on Inter Pausal Utterance

Takuya KANEKO and Shun ISHIZAKI

Graduate School of Media and Governance
Keio University
5322 Endo, Fujisawa, Kanagawa, 252-0816 JAPAN
takuya@sfc.keio.ac.jp ishizaki@sfc.keio.ac.jp

We present a morphological analysis method for a dialogue text understanding system which progressively understands utterrances with robustness. We developed an analysis method for spontaneous dialogue texts based on inter pausal utterrance. This method is effective to make its analysis results much better. The results shows that: (1) connectivity cost that is assigned to each bi-gram of part-of-speeches is differ from each speaker in particular cases, and therefore it is effective in increasing recall to extract connectivity cost separately from the other speaker's utterances; (2) if the system is trained with enough pausal data, additional trainings by pausal infomation for new utterance sets are not nessesary; (3) it is effective in increasing recall to use classified pause information by using their duration; and (4) the increase in the recall of pronunciation level analysis results is bigger than that of reading level analysis results, and that of reading is bigger than that of surface form level analysis results. one of the reasons for (4) is because Japanese ideograph present in surface form provides much infomation about japanese contain words than Japanese syllabary in reading and pronunciation.

KeyWords: Corpus, Inter Pausal Utterance, Part-of-speech Tagging, Spontaneously Spoken Dialogue

1 Introduction

Our aim is to bild a spoken dialogue text understanding system which incrementaly understands utterrances with robustness. The initial stage of text analysis for any NLP task usually involves the tokenization of the input into words (Sproat, Shin, Gale, and Chang 1996). In many languages the punctuation mark that indicates the end-of-sentence boundary is ambiguous, and most tokenizers of writings must be equipped with special sentence boundary recognition rules (Palmer and Hearst 1997). Forthermore, it is more complicated problem than in writing to recognize sentence boundaries in speech, because explicit punctuation is absent in speech and speech is not necessarily composed of sentences. Similary, both manual or automatic recognition of discourse segment boundaries is also complicated, because the discourse structural infomation can be inferred from orthographic cues in text, such as paragraphing and punctuation; from linguistic cues in text or speech, such as discourse markers; from variation in referring expressions, tense, and aspect; from prosodic/accoustic cues, such as pitch range, pausal duration, intonational variation (Hirschberg and Nakatani 1996). Meanwhile, the recognition of pauses as utterance boundaries is less complicated than that of sentence

boundaries or discourse segment boundaries, thus we present a morphological analysis method based on inter pausal utterance.

2 Theory

2.1 A Stochastic Tagging Model

Chasen (NAIST 1999), a free Japanese Morphological analyser which segments sentences into morphemes and tags them with their parts of speech, is used as a morphological analysis engine of our experiment. This analyzer is based on the minimum connection cost method, which estimates the optimal tag sequence $s_{0,n+1} = \{s_0, s_1, \ldots, s_n, s_{n+1}\}$ for given observation morpheme sequence $w_{1,n} = \{w_1, w_2, \ldots, w_n\}$ by summing up morpheme cost $Cb(w_i)$ and connectivity cost $C(s_i, s_{i+1})$ of two parts of speech, and outputs the result with the minimum cost:

$$\arg \min_{s_{0,n+1}} \sum_{i=0}^{n} (Cb(w_i) + C(s_i, s_{i+1})), \tag{1}$$

where s_0 and s_{n+1} are the special parts of speech which denote beginning-of-sentence and end-of-sentence, respectively.

Connectivity and morpheme costs are automatically extracted from a part of speech tagged corpus. In order to tune these costs, part of speech bi-gram Markov model is employed, and the probability parameters of maximum likelihood estimate (MLE) model are transformed into its connectivity costs:

$$C(s_i, s_{i+1}) = \log \frac{1}{P(s_{i+1} \mid s_i)} = \log \frac{F(s_i)}{F(s_i, s_{i+1})}, \tag{2}$$

where the $F(s_i)$ is the frequency of the part of speech s_i, and $F(s_i, s_{i+1})$ is the frequency of the part of speech state transition from s_i to s_{i+1}.

Similarly, morpheme costs are also obtained from the MLE model:

$$Cb(w_i) = \log \frac{1}{P(w_i \mid s_i)} = \log \frac{F(s_i)}{F(w_i, s_i)}, \tag{3}$$

where $F(w_i, s_{i+1})$ is the frequency of the morpheme w_i produced by the part of speech state s_i, and the Chasen's standard maximum costs of connectivity and morpheme are 4000.

In this study, We used our own connectivity consts and Chasen's default morpheme costs.

2.2 A criterion of difference between connectivity costs of speakers

Connectivity rules $R_k(s_i, s_{i+1})$ denote that transitions from part of speech states s_i to part of speech states s_{i+1} are possible for each i. Figure 1 shows a diagram associated with $(p-1)$-regular graph G that is a set $V(G)$ of those speakers $v_1, \ldots, v_p$ with a certain connectivity rule $R(s_{i_0}, s_{i_0+1})$ and a set $E(G)$ of pairs of speakers, where each label on the speakers is his v_j own connectivity cost $C_j(s_{i_0}, s_{i_0+1})$, and each label on the arcs is the difference between the costs of adjacent speakers.

5, and 6:

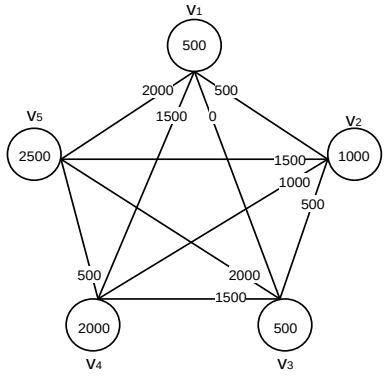


Figure1: Difference between connectivity costs of speakers. A criterion for the measure of difference between connectivity costs of speakers is given by the following equations 4

$$q = \sum_{l=1}^{p-1} l, \tag{4}$$

$$t = \sum_{l=1}^{p-1} \sum_{m=l+1}^{p} \mid (C_l(s_{i_0}, s_{j_0}) - C_m(s_{i_0}, s_{j_0}) \mid, \tag{5}$$

$$\delta = \log qt, \tag{6}$$

where p and q are the order and size of graph G, respectively, t is total difference between connectivity costs of speakers, δ is a criterion of difference between connectivity costs of speakers, and the maximum δ is 1000 in this study.

2.3 Recall and Precision

The performance of a set of morphological analyses is measured by recall and precision. Suppose that strings are yielded incrementaly based on a certain unit, such as inter pausal utternce, discourse segment, sentence, etc., and that a system analyze each string in all the yields $\vec{u} = (u_1, \ldots, u_n)$, where u_i is a string which forms a certain unit. Then let W_i be the set of all morphemes identified by the correct analysis of s_i, and W_i' be the set of all morphemes identified by system analysis of s_i. The intersection of W_i and W_i' (written $W_i \cap W_i'$) is the set of all morphemes which are in both W_i and W_i', and recall is a proportion between the number of points in $W_i \cap W_i'$ and the number of points in W_i:

$$Recall = \frac{\sum_{i=1}^{n} \mid W_i \cap W_i' \mid}{\sum_{i=1}^{n} \mid W_i \mid}, \tag{7}$$

where $\mid W_i \mid$, $\mid W_i' \mid$, and $\mid W_i \cap W_i' \mid$ are the numbers of points in W_i, W_i', and $W_i \cap W_i'$, respectively. On the other hand, Precision is a proportion between $\mid W_i \cap W_i' \mid$ and $\mid W_i' \mid$:

$$Precision = \frac{\sum_{i=1}^{n} \mid W_i \cap W_i' \mid}{\sum_{i=1}^{n} \mid W_i' \mid}. \tag{8}$$

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
4.089	S	BEGIN							
4.954	E	e-Qto	eto	eto	eto	eto	filler		
6.894	E	kono'	kono	kono	kono	kono	rentaishi		
7.079	E	hito	hito	hito	[hito]	[hito]	noun-general		
7.469	E	wa-	wa	ha	ha	ha	pp-theme		
8.176	S	0.707							
8.619	E	kuro'me	kurome	kurome	[kurome]	[kurome]	noun-general		
8.833	E	qa	ga	ga	ga	ga	pp-case-general		
9.398	E	hikakuteki	hikakuteki	hikakuteki	[hikakuteki]	[hikakuteki]	adv-general		
9.750	E	ni-	ni	ni	ni	ni	pp-adv		
10.053	S	0.303							
10.352	E	o'-ku	o-ku	ooku	[oo]ku	[oo]i	adj-content	adj-auo	ren'yo-te-connective
10.771	E	te-	te	te	te	te	pp-connective		
11.571	S	0.800							

Figure2: Example of the part of speech tagged dialogue corpus: (1) acoustic time, (2) start or end flag of morphemes, (3) phonological transcription including pausal duraion, (4) pronunciation, (5) reading, (6) surface form, (7) surface base form, (8) parts of speech, (9) conjugation type, (10) conjugated form. Japanese ideograph is bracketed by [and].

3 Experiment

3.1 Corpus

The Multi-modal Dialogue Corpus (Kaneko and Ishizaki 1999) is a spontaneously spoken and task-oriented dialogue corpus of colloquial Japanese, which consists of nine dialogues with eleven participants (approximately 80 minutes dialogues). This corpus includes Face Task. Face Task is a cooperative task involving two participants. One speaker, the Explainer, has a portrait printed on his/her sheet of material, while the other, the Answerer, has 16 portraits on his/her sheet of material, including the same one as Explainer's. Explainer describes the facial feature of the target subject. Their goal is that Answerer detects the target sugject. We transcribed four of these dialogues with seven participants verbatim at four orthographic levels: phonological, pronunciation, reading, and surface-form levels. Although the corpus has not enough size, it manipulates as much as possible the following variables: task, familiarity and sexuality of speakers, eye contact between speakers. A feature of present corpus is collection of colloquial Japanese spoken by University students. We tagged these transcripts manually with a morphological structure of a forme; <surface base form, part of speech, conjugation type, conjugated form>. A example of the part of speech tagged dialogue corpus is shown in Figure 2.

3.2 Classification of pauses

An inter pausal utterance is extracted from corpus, where pauses are classified into two categories by their duration and they are denoted by two characters: if a pausal duration is shorter than a threshold T of time, then denoted by ○, else by □, as shown in Figure 3. Then the inter pausal utterance and categorized pauses are input to the morphological analyzer and analyzed, where we can see classification of pauses by threshold of time causes the recall of morphological analysis either of increase or decrease, as is shown in Figure 4. In this case, the recall increases from 84.39% (T = 0*sec.*) to 84.86% (T=0.316sec), then decrease to 84.03% (T=2.512sec).

Surface form $T = 0msec.$	Reading $T = 0.316msec.$	Pronunciation $T = 0.794msec.$
□ etokono[hito]ha □	□ etokonohitoha □	□ etokonohitowa ○
□ [kurome]ga[hikakuteki]ni □	□ kuromegahikakutekini ○	○ kuromegahikakutekini ○
□ [oo]kute □	○ ookute □	○ o-kute □
□ de[kami]no[ke]ga □	□ dekaminokega □	□ dekaminokega ○
(a)	(b)	(c)

Figure3: Examples of inter pausal utterances, where utterances are transcribed at three authographic levels: (a) surface form level, (b) reading level, and (c) pronunciation level, and pauses are denoted by two characters: if a pausal duration is shorter than threshold T, then by ○, or else by □. Japanese ideograph is bracketed by [and].

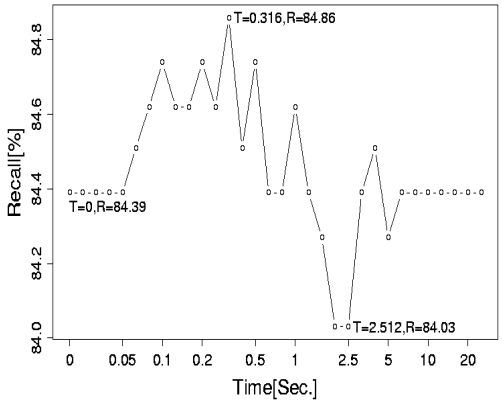


Figure4: Relation between the threshold of time and the recall of the system analysis with using examples of pronunciation level transcripts of speaker FKO.

3.3 Analysis conditions

A dialogue corpus $\mathcal{H}$ is a collection of utterance sets of speakers: $\mathcal{H} = \{H_1, \ldots, H_n\}$, where H_i is an utterance set of the speaker h_i. We use eight analysis conditions by combinations of training sets, test sets, and the number of thresholds as is shown in Table 1, where H_k^* is an utterance set which substanciates dummies for the pausal elements of the intersection of H_k and 20% of H_k. The test set and the lerning set are generally disjoint, but most of our test sets are contained by their lerning set, because our courpus does not have enough size to extract test sets from it.

4 Results and Discussion

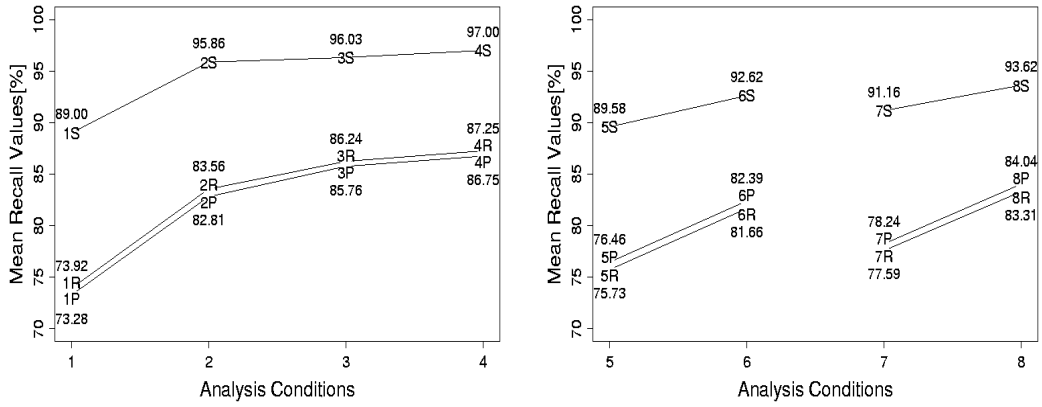
Table 2 shows each speaker's recall value of the analysis results at the surface form level. Figure 5 shows the mean recall values of the analysis results.

Table 1: Analysis Conditions.

Analysis Conditions	Number of Thresholds	Training Set	Test Set
1	0	$\mathcal{H} - H_k$	H_k
2	0	$\mathcal{H}$	H_k
3	0	H_k	H_k
4	1	H_k	H_k
5	0	H_k^*	20% of H_k
6	0	H_k	20% of H_k
7	1	H_k^*	20% of H_k
8	1	H_k	20% of H_k

Table 2: Each speaker's recall and precision at the surface form level

Condition	Speaker	FNO	FMA	FTS	FKO	MNN	MYS	MKS	Mean
C1	RECALL	94.17	94.12	90.21	87.01	81.59	90.00	85.91	89.00
	PRECISION	93.27	94.12	88.83	87.32	80.50	88.67	84.18	88.13
C2	RECALL	99.03	97.06	95.77	94.40	93.02	94.50	97.25	95.86
	PRECISION	99.03	97.06	95.24	95.08	91.25	94.50	97.25	95.63
C3	RECALL	97.09	97.06	96.11	95.59	94.57	94.50	99.31	96.32
	PRECISION	97.09	97.06	95.58	96.39	92.25	94.50	99.31	96.03
C4	THRESH	0.158	0	0.398	0.063	0.794	3.162	0	
	RECALL	98.06	97.06	96.77	95.95	95.35	96.50	99.31	97.00
	PRECISION	98.06	97.06	96.13	96.64	92.83	96.50	99.31	96.65
C5	RECALL	90.91	71.43	95.56	94.94	89.42	88.10	96.67	89.58
	PRECISION	90.91	83.33	95.03	98.83	89.42	90.24	98.31	92.30
C6	RECALL	90.91	85.71	96.67	96.07	89.42	92.86	96.67	92.62
	PRECISION	90.91	100.00	96.67	99.42	89.42	95.12	98.31	95.69
C7	THRESH	0.158	0	0.794	0.794	0.079	1.585	0	
	RECALL	95.45	71.43	96.67	96.07	91.35	90.48	96.67	91.16
	PRECISION	95.45	83.33	95.60	99.42	90.48	92.68	98.31	93.61
C8	THRESH	0.158	0	0.063	0	0.079	0	0	
	RECALL	95.45	85.71	97.22	96.07	91.35	92.86	96.67	93.62
	PRECISION	95.45	100.00	97.22	99.42	90.48	95.12	98.31	96.57



(a) (b)

Figure5: The relation between mean recall values and analysis conditions, where S, R, and P are Surface form, Reading, and Pronunciation - three authographic levels of the transcription, respectively. The analysis conditions are shown in Table1.

Table 3: Difference between the connectivity costs of speakers: the difference between the costs of part of speech state transition from s_i to s_{i+1}, where q is a number of relations between the speakers, δ is the difference between connectivity costs of speakers which is given by equation 4, 5, and 6 of section 2.2, and $\frac{\delta}{q}$ is mean difference of costs. The part of speech pp denotes Japanese postposition.

	s_i	s_{i+1}	q	$\frac{\delta}{q}$	δ
1	pp-theme	pause	21	972	1000
2	pause	adv-connective_to_pp	21	958	999
3	noun-general	pause	21	797	985
4	pause	filler	21	751	980
5	pause	interjection	21	722	977
6	pause	noun-general	21	684	973
7	pp-connective	aux	15	1111	958
8	pp-case-general	pause	15	1001	950
9	adv-general	verb-content	15	938	945
10	aux	aux	21	456	942

4.1 Difference between connectivity costs of speakers

As is shown in Figure 5 (a), the mean recall values of the analysis condition 2 are smaller than those of the condition 3 at all authographic levels. Since the difference between condition 2 and condition 3 is only that the training set H_k of condition 3 does not involve those utterance collection $\mathcal{H} - H_k$ of the other speakers, then this fact indicates the connectivity costs of speakers are different from one another, and we substantiated, as is shown in Table 3, that the costs of some part of speech state transitions of speakers are considerably different from one another, eg., the transition from pause to adverb, filler, interjection, and noun; from postposition to pause, and auxiliary verb; from auxiliary verb to auxiliary verb, etc..

4.2 Effect of the additional training in pausal infomation

Figure 5 (b) shows the effect of additional training in pausal infomation. The mean recalls of the analysis conditions 5 and 7 are smaller than those of analysis conditions 6 and 8 at all authographic levels, respectively. Although the difference between the former conditions and the latter is only that the trainig sets of conditions 5 and 7 do not involve the 20% of pausal bi-grams in the trainig sets of conditions 6 and 8, respectively. The above facts does not necessarily mean the additional training of pausal infomation is essential for analyzing every new set of utterances, because recalls hardly increase in case training sets are not small. For instance, making a comparison between condition 5 and 6 at pronunciation transcript level, the recalls of the speaker FNO and FMA increase from 72.73% and 71.43% to 86.36% and 85.71%, respectively; on the other hand the recalls of the speaker FTS and FKO increase from 81.67% and 83.33% to 82.02% and 83.71%, respectively.

4.3 Effect of the Classification of pauses

As is shown in Figure 5 (a), the mean recalls of the analysis condition 4 is bigger than those of the analysis condition 3 at all authographic levels. Since the difference between conditions 3 and 4 is only that the training sets of the latter involves pauses classified into two categories by their duration, then this fact indicates the classification of pauses is effective.

4.4 Relation between Authographic levels and recall

As is shown in Figure 5 (a) and (b), making a comparison between adjacent conditions, the increase in the mean recalls of pronunciation transcritpts is bigger than that of reading transcripts, and that of reading transcripts is bigger than that of surface form transcripts.

5 Conclusion

We studied a method for morphological analysis of transcripts from spontaneously spoken dialogue based on inter pausal utterance. The resuts shows that: (1) connectivity cost that is assigned to each bi-gram of part-of-speeches is differ from each speaker in particular cases shown in Table 3, and therefore it is effective in increasing recall to extract connectivity cost separately from the other speaker's utterances; (2) if the system is trained with enough pausal data, additional trainings by pausal infomation for new utterance sets are not nessesary; (3) it is effective in increasing recall to use classified pausal information by using their duration; and (4) the increase in the recall rate of pronunciation is bigger than that of reading, and that of reading is bigger than that of surface form, one of the reasons for these results is because Japanese ideograph present in surface form provides much infomation about japanese contain words than Japanese syllabary in reading and pronunciation.

Reference

Hirschberg, J., and Nakatani, C. H. (1996). "A Prosodic Analysis of Discourse Segments in Direction-Giving Monologues." In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, pp. 286–293.

Kaneko, T., and Ishizaki, S. (1999). "The Multi-Modal Dialogue Corpus." In *Proceedings of the 2nd International Conference on Cognitive Science*, pp. 1010–1013.

NAIST (1999). "Chasen2.02.". http://cl.aist-nara.ac.jp/lab/nlt/chasen.html.

Palmer, D. D., and Hearst, M. A. (1997). "Adaptive Multilingual Sentence Boundary Disambiguation." *Computational Linguistics*, *23*(2), 241–267.

Sproat, R., Shin, C., Gale, W., and Chang, N. (1996). "A Stochastic Finite-State Word-Segmentation Algorithm for Chinese." *Computational Linguistics*, *22*(3), 377–404.