IDENTIFICATION OF “SENTENCES” IN SPONTANEOUS JAPANESE — DETECTION AND MODIFICATION OF CLAUSE BOUNDARIES —

Katsuya Takanashi [†] *Takehiko Maruyama* [‡] [*] *Kiyotaka Uchimoto* [†] *Hitoshi Isahara* [†]

[†] Communications Research Laboratory
[‡] The National Institute for Japanese Language
[*] ATR Spoken Language Translation Research Laboratories

ABSTRACT

The identification of basic units in spontaneous Japanese is an indispensable issue for spoken language processing. This paper describes a method for the semi-automatic detection of “sentences” from the *Corpus of Spontaneous Japanese* (**CSJ**).

1. INTRODUCTION

The *Corpus of Spontaneous Japanese* (**CSJ**) [3], a large scale spontaneous speech corpus of common Japanese, is now being compiled under a project entitled “Spontaneous Speech: Corpus and Processing Technology.” This corpus consists mainly of monologues, including ‘academic presentation speech (APS)’ and ‘simulated public speech (SPS)’, which will be a fruitful resource for the research of spontaneous speech.

In written language processing, a sentence is generally used as a basic unit for syntactic parsing, translation, text summarization, and so on. In spoken language processing, however, it is difficult to use a sentence as a basic unit because a spoken language corpus often contains no punctuation. Therefore, it is necessary to find reasonable “sentences” in spoken language, instead of sentences in written language, which will be useful for automatic text summarization, parsing the dependencies between *bunsetsu*s (Japanese phrasal units) [1], and analyzing discourse structure [7].

In this paper, we describe a semi-automatic method of detecting “sentence” boundaries from the CSJ. We first extracted clause boundaries automatically as candidates (Chapter 3), then manually modified the results along previously defined criteria (Chapter 4). The modification was applied to various characteristic phenomena in spoken language, such as noun final clauses, shared topics, quotations, insertions, and inversions.

2. “SENTENCES” IN SPONTANEOUS JAPANESE

In general, a sentence is a very standard unit for natural language processing, syntactic analysis in linguistics, and ordinary human language activities. In dealing with the spontaneous speech, however, a sentence is not necessarily appropriate for processing or analysis because spontaneous speech basically contains no periods to mark the sentence boundaries. Moreover, it is fundamentally difficult to find obvious sentence boundaries from spontaneous utterances, which usually contain utterance errors, utterance stops, and other characteristic phenomena. Thus, we need to define and detect some reasonable segmented unit for processing as a “sentence” in spontaneous speech.

Here, we will introduce clauses as candidates for these basic units, “sentences”, in spontaneous speech. Clauses, whose boundaries are marked mainly by verb phrases in Japanese, are meaningful constituents for a variety of purposes. The boundaries between each clause can be detected more easily than those of sentences, considering the conjugated verb forms or conjunctive particles put at the end of clauses. Accordingly, we divide utterances into small units by detecting clause boundaries.

3. AUTOMATIC SEGMENTATION

Japanese is a SOV language, and verb phrases are placed at the end of clauses [6]. Clause boundaries are marked by conjugated forms of verb phrases or conjunctive particles. We can extract various types of boundaries quite precisely by referring to part-of-speech (POS) tags. We developed a program that segments the transcription of the CSJ into clauses automatically by referring to POS tags [4].

The program we developed to detect clause boundaries from the CSJ is in reality a set of conversion rules which finds particular patterns of concatenations of one to three morphemes and inserts labels after the boundaries. When a particular concatenation of morphemes is accepted as an input, the program compares it with the boundary patterns prepared manually. Each morpheme is formed by four tags, i.e., surface form, POS, conjugation form, and conjugation type [8]. If the input matches a certain boundary pattern, boundary labels are inserted into the text. Examples of the rules are shown below.

```
1.  s/(*_動詞_*_連用形) /$1 <連用節> /g;
2.  s/(が_接続助詞__) /$1 \/並列節ガ\/ /g;
```

Rule 1 accepts an infinitive form (連用形) of any verb (動詞) as an input and puts a boundary label **<**連用節**>** (infinite clause) after the verb. Rule 2 accepts a conjunctive particle *ga* (が), and puts a boundary label **/** 並列節ガ **/** (*ga*-clause). Applying these rules, we can obtain the labeled text, as in (1).

(1) 成田から飛行機に乗り <**連用節**> ロスで乗り換えて <**テ節**> オーランドに着きます **[文末]** 八泊十日とちょっと長めだったのですが **/並列節ガ/** お金を頑張って <**テ節**> 溜めて行きました **[文末]** . . .

The inserted labels are bracketed like <**連用節**>. The program consists of 142 conversion rules, and extracts 33 types of clause boundaries that can be classified into three levels as follows:

- AB: Absolute boundary (labeled by **[***]**)
- SB: Strong boundary (labeled by **/ *** /**)
- WB: Weak boundary (labeled by < *** >)

These three levels differ in terms of their degree of completeness as a syntactic and semantic unit, and their independencies of the subsequent clauses. Absolute boundaries correspond to sentence boundaries in the usual meaning. Strong boundaries are the points which can be regarded as major breaks in utterances, and proper points for segmentation. Weak boundaries are also clause boundaries; however, they are not regarded as proper points for segmentation because they are strongly dependent to other clauses.

These three levels were set up manually according to the classification of Japanese subordinate clauses by Minami(1974) [5], and its empirical revision. Preparing these three levels makes it possible to estimate the syntactic and semantic relations among clauses in advance, e.g., the scope of auxiliary verbs, the sharing of arguments and topics.

We applied this program to 338 CSJ monologues (containing 804,983 morphemes). Table 1 shows the result of segmentation.

Table 1. Clause boundaries in the CSJ

Boundary	Frequency	Rate
Absolute	21,693	(25.00%)
Strong	14,350	(16.53%)
Weak	50,770	(58.48%)
Total	86,813	(100 %)

We adopt absolute boundaries and strong boundaries as default boundary candidates because we considered that both are useful and meaningful for the many aspects of processing. We left the labels of weak boundaries in the text for later manual modification (See Chapter 4). We also have to deal with the problem of utterance errors, utterance stops, and a few kinds of clause boundaries that cannot be extracted by the program. We need to check and modify the candidates manually along the criteria described below.

4. MANUAL MODIFICATION

As described above, the automatic segmentation rule determines the boundary by referring only to the local concatenation of morphemes. However, there are some characteristic phenomena which neither be extracted nor treated appropriately by the local segmentation rule, i.e., noun final clauses, shared topics, quotations, inserted clauses. In these cases, the default boundaries must be manually modified by referring to the recorded speech to construct processing units that are syntactically well-formed as well as semantically adequate. Manual modifications comprise three kinds of operations:

Connecting two or more default units by `+`
Separating a default unit by `-`
Enclosing some elements by `(***)`, {***} or 《***》.

Annotators are required to modify the results of automatic segmentation based on several prescribed criteria, which consist of definitions of phenomena and the operations required. Some of these definitions of phenomena and examples of manual operation are explained below.

4.1. Noun final clauses

Nouns sometimes constitute independent clauses by themselves. The most typical pattern of this is the title of a lecture, which is independent of both the preceding and following clauses. Since the automatic segmentation rule cannot detect these points because of a lack of verb phrases, the noun must be separated from the following part.

AAA BBB*
→ AAA -# BBB* [1]

Operation: If AAA constitutes a noun final clause which is syntactically independent of BBB, AAA must be separated from BBB.

(2) タイトル `- ;Noun final clause` [2]
夢の 国 ディズニーワールド `- ;Noun final clause`
私は旅行が大好きで <**並列節デ**>
“Title. Dream land, Disney World. I like traveling very much,”

4.2. Shared topics

In Japanese, topicalized elements are marked by the topic-marking particles *wa* (は) or *mo* (も). Topicalized elements

[1] “*” means the default boundary ([AB] or /SB/), “#” means a new boundary after modification. `-`, `+`, ( ), { } and 《 》 are operation symbols described below.

[2] We put comments after ‘;’ to show the type of operation.

tend to take wide scopes over the default boundaries. In these cases, two or more default units must be connected as one unit.

XXX*wa* AAA /SB/ BBB *
→ XXX*wa* AAA /SB/ + BBB *

Operation: If the topicalized element 'XXX*wa*' depends not only on AAA but also on BBB, i.e., it is shared by the verb phrases of both AAA and BBB, the default boundary must be removed and AAA must be connected with BBB.

(3) 私は旅行が大好きで **/並列節デ/** + 今まてもあちこち行きましたけれども **/並列節ケレドモ/**;`Shared topic`
"I like traveling very much, + and have been to many places,"

4.3. Embedded clauses

Default units ending in absolute or strong boundaries may be embedded in the quotation. Since the local segmentation rule cannot parse the full structure of utterances, it segments all of the default boundaries even if they are embedded in the quotation, and thus the main clause is not completed. In such cases, embedded clauses must be bracketed by ｛ ｝, and the default boundaries replaced by "：".

AAA BBB* CCC DDD*
→ AAA { BBB : CCC } DDD*

Operation: If AAA depends on DDD, and BBB and CCC are embedded in them, BBB and CCC must be enclosed by ｛ ｝. Default boundaries between the embedded clauses must be replaced by ：.

(4) もう 何度も これは テレビで 見て <**テ節**> ｛ いいな **/文末候補/** ： いつか 行きたいな ｝と <**引用節**> もう ずっと 思っていた ところで <**並列節デ**> ;`Quotation`
"I have been feeling for a long time that ｛ the place must be very nice ： and I hope to go there someday ｝ every time I watched it on TV."

4.4. Phenomena Characteristic to Spontaneous Speech

In the production of spontaneous speech, speech plans constructed beforehand are sometimes changed during the utterance due to phonological, lexical, syntactic or ordering problems. In particular, long spontaneous monologues impose heavy linearization problem on speakers, such as deciding what to say first, and what to say next [2]. This causes various disfluencies, such as insertions, inversions, utterance stops, and so on. If these disfluencies bring about indesirable default units, they must be marked.

Inserted clauses

In spontaneous speech, it can be observed that speakers insert clauses in the middle of other clauses. This occurs when speakers change their speech plans while producing utterances, which results in supplements, annotations, or paraphrases of main clauses.

AAA BBB /SB/ CCC*
→ AAA （ BBB /SB/ ） + CCC*

Operation: If BBB can be judged as an inserted clause, and, in addition, AAA depends on CCC, BBB must be enclosed by （ ）, and the default boundary after BBB must be removed and connected with CCC.

(5) 色んな パターンを （ ここに 書いてある 数字は 頻度ですが **/並列節ガ/** ） + たくさん 集めてみました **[文末]** ;`Inserted clause`
"I collected, （ the number written here presents frequency, ） various patterns."

Inversion

*Bunsetsu*s are sometimes placed immediately after the verb phrase on which they depend. This can be regarded as an inversion from the viewpoint of canonical word order of Japanese, which may be caused by various reasons in the production of speech, such as supplements or afterthoughts.

AAA BBB* CCC DDD*
→ AAA BBB + 《CCC 》 -# DDD*

Operation: If *bunsetsu* CCC can be judged as an inverted element which depends on the verb phrase in BBB, the default boundary after BBB must be moved after CCC, and CCC must be separated from DDD.

(6) 家に いる 時間に 必ず 掛かってくるんですね **[文末]** + 《一日に 三四回 》 - ;`Inversion`
で 何か 今 よく テレビで ストーカーの 話とか やってるけれども **/並列節ケレドモ/**. . .
"I have been called whenever I am at home, 《several times a day 》 - And news about stalkers is frequently shown on TV recently,. . . "

Cut off — giving up the utterance

Speakers sometimes give up uttering following parts in the middle of a clause when they change the speech plans they constructed beforehand during the utterance. The abandoned part is left alone, and cannot be treated as a meaningful unit in the subsequent processing.

AAA BBB*
→ AAA -# BBB*

Operation: If some elements AAA can be judged as a fragment of a clause left alone by giving up the utterance that does not constitute a meaningful unit, AAA must be separated from BBB.

(7) 今回の 実験で - ;`cut-off`
次ページの 表に 示しましたのは 実験条件と 実験結果の 一覧です **[文末]**
"In this experiment - The table on the next page shows the list of conditions and results of the experiment."

Table 2. Manual modification results.

	Default	Modified	MM rate	+	-	( )	{ }	《 》
APS	102.1	94.4	16.9%	10.8	4.8	3.5	1.7	0.1
SPS	89.6	84.5	24.4%	12.4	8.5	4.4	1.5	0.5
Total	94.0	88.0	21.6%	11.9	7.2	4.1	1.6	0.4

5. DISCUSSION

5.1. Results of manual modification

Based on the criteria described in the previous chapter, 43 monologues, comprising 15 academic presentation speeches (APS) and 28 simulated public speeches (SPS), were modified manually and "sentences" in the CSJ were extracted. Two or more annotators tagged the automatic segmentation results of each lecture. The modification results presented by each annotator were compared and complied.

In table 2, "Default" means the average number of default units through automatic segmentation within a lecture, "Modified" means the average number of "sentences" after the manual modification, and "MM rate" means the ratio of the total number of operations[3] to the number of default units. The average number of times that each operation was performed is also shown.

The point to notice here is the significant difference in the MM rate between APS and SPS. It can be considered that this is caused by the difference in spontaneity between APS, academic presentations that are generally spoken in accordance with prepared manuscripts, and SPS, personal narratives that are spoken less formally. This tendency roughly corresponds to the fact that SPS contains more phonological or morphological disfluencies than APS [3].

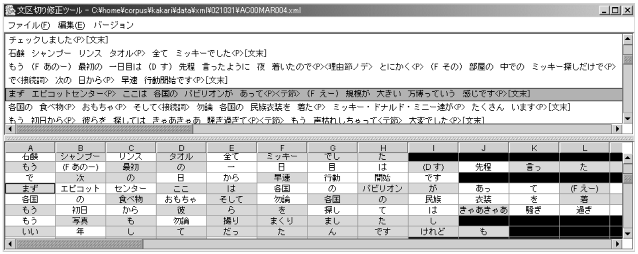
5.2. Necessity of supporting tool for the modification

The total average of the MM rate was 21.6%. This result shows that manual modification is not an exceptional task if phenomena as described in Chapter 4 are significant. Therefore the work should be supported by annotation tools that can decrease the number of inevitable errors as well as the strain on annotators in the modification process.

We developed an annotation tool for manual modification, as shown in Figure 1. This tool restricts annotators to the prescribed uses of symbols, and ensures the correspondences between symbols and obligatory comments showing the kinds of operations. This tool also enables easy comparison of the results of annotators and efficient data management.

6. CONCLUSION AND PROSPECTS

The identification of useful units in spontaneous speech is a necessary but difficult task. This paper has proposed a method for semi-automatically detecting "sentences" from the CSJ.


文区切り修正ツール - C:¥home¥corpus¥kakari¥data¥xml¥021031¥AC00MAR004.xml
ファイル(F)
編集(E)
バージョン

チェックしました<P>[文末]
石鹸　シャンプー　リンス　タオル<P>　全て　ミッキーでした<P>[文末]
もう　(F あのー)　最初の　一日目は　(D す)　先程　言ったように　夜　着いたので<P><理由節ノデ>　とにかく<P>　(F その)　部屋の　中での　ミッキー探しだけで<P>
で<接続詞>　次の　日から<P>　早速　行動開始です<P>[文末]
まず　エピコットセンター<P>　ここは　各国の　パビリオンが　あって<P><テ節>　(F えー)　規模が　大きい　万博っていう　感じです<P>[文末]
各国の　食べ物<P>　おもちゃ<P>　そして<接続詞>　勿論　各国の　民族衣装を　着た<P>　ミッキー・ドナルド・ミニー達が<P>　たくさん　います<P>[文末]
もう　初日から<P>　彼らを　探しては　きゃあきゃあ　騒ぎ過ぎて<P><テ節>　もう　声枯れしちゃって<テ節>　大変でした<P>[文末]

A	B	C	D	E	F	G	H	I	J	K	L
石鹸	シャンプー	リンス	タオル	全て	ミッキー	でし	た				
もう	(F あのー)	最初	の	一	日	目	は	(D す)	先程	言っ	た
で	次	の	日	から	早速	行動	開始	です			
まず	エピコット	センター	ここ	は	各国	の	パビリオン	が	あっ	て	(F えー)
各国	の	食べ物	おもちゃ	そして	勿論	各国	の	民族	衣装	を	着
もう	初日	から	彼	ら	を	探し	て	は	きゃあきゃあ	騒ぎ	過ぎ
もう	写真	も	勿論	撮り	まくり	まし	た	し			
いい	年	し	て	だっ	た	ん	です	けれど	も		

Fig. 1. Annotation tool for manual modification.

We are tagging the "sentences" to a subset of the CSJ (183 monologues), whose size will be 500K words. The CSJ tagged with these "sentences" will be used for automatic text summarization, parsing the dependencies between *bunsetsu*s, and analyzing discourse structure.

7. REFERENCES

[1] Kurohashi, S., Nagao, M. A Syntactic Analysis Method of Long Japanese Sentences based on the Detection of Conjunctive Structures. *Computational Linguistics*, **20**(4), 507–534. 1994.

[2] Levelt, W.J.M. *Speaking: From Intention to Articulation.* The MIT Press. 1989.

[3] Maekawa, K. Corpus of Spontaneous Japanese: its design and evaluation. This volume.

[4] Maruyama, T., Kashioka, H., Kumano, T. and Tanaka, H. Setsu kyoukai jidou kenshutu ru-ru no sakusei to hyouka. *Proc. of The Nineth Annual Meeting of The Association for Natural Language Processing*. 2003.

[5] Minami, F. *Gendai Nihongo-no Kouzou*. Taishukan-Shoten. 1974.

[6] Shibatani, M. *The Languages of Japan*. Cambridge University Press.

[7] Takeuchi, K., Takanashi, K., Morimoto, I., Koiso, H. and Isahara, H. Committee-based Discourse Purpose Assignment: Discourse Structure Annotations of Spontaneous Japanese Monologue. This volume.

[8] Uchimoto, K., Nobata, C., Yamada, A., Sekine, S., and Isahara, H. Morphological Analysis of The Corpus of Spontaneous Japanese. This volume.

[3] Regular correlations of symbols such that each ( ) is followed by + are taken into consideration.