TR-703

A Discourse Structure Analyzer for Japanese Text

by
K. Sumita, K. Ono, T. Chino, T. Okita &
A. Amano (Toshiba)

October, 1991

ICOT
Mita Kokusai Bldg. 21F
4-28 Mita 1-Chome
Minato-ku Tokyo 108 Japan
(03)3456-3191～5
Telex ICOT J32964

Institute for New Generation Computer Technology

A Discourse Structure Analyzer for Japanese Text *

K. Sumita, K. Ono, T. Chino, T. Ukita, S. Amano
Toshiba Corp. R&D Center
Komukai-Toshiba-cho 1, Saiwai-ku, Kawasaki 210, Japan

Abstract

This paper presents a practical procedure for analyzing discourse structures for Japanese text, where the structures are represented by binary trees. In order to construct discourse structures for Japanese argumentative articles, the procedure uses local *thinking-flow* restrictions, *segmentation rules*, and *topic flow* preference. The thinking-flow restrictions restrict the consecutive combination of relationships detected by connective expressions. Whereas the thinking-flow restrictions restrict the discourse structures locally, the segmentation rules constrain them globally, based on rhetorical dependencies between distant sentences. In addition, the topic flow preference, which is the information concerning the linkage of topic expressions and normal noun phrases, chooses preferable structures. Using these restrictions, the procedure can recognize the scope of relationships between blocks of sentences, which no other discourse structure analysis methods can handle. The procedure has been applied to 18 Japanese articles. Results show that this approach is promising for extracting discourse information.

1 Introduction

A computational theory for analyzing linguistic discourse structure and its practical procedure are necessary to develop machine systems dealing with plural sentences; e.g., systems for text summarization and for knowledge extraction from a text corpus.

*This work was supported by ICOT (Institute for New Generation Computer Technology), and was carried out as a part of the Fifth Generation Computer Systems research.

Hobbs develops a theory in which he arranges three kinds of relationships between sentences from the text coherency viewpoint [Hobb 79]. Grosz and Sidner propose a theory which accounts for interactions between three notions on discourse: linguistic structure, intention, and attention [Gros 86]. Litman and Allen describe a model in which a discourse structure of conversation is built by recognizing a participant's plans [Litm 87]. These theories all depend on extra-linguistic knowledge, which requires an extensive effort to accumulate for realizing a practical analyzer. The authors aim to build a practical analyzer which dispenses with such extra-linguistic knowledge as depends on topic areas of articles to be analyzed.

Mann and Thompson propose a linguistic structure of text describing relationships between sentences and their relative importance [Mann 87]. However, no method for extracting the relationships from superficial linguistic expressions is described in their paper. Cohen proposes a framework for analyzing the structure of argumentative discourse [Cohe 87]. Yet he dose not provide a concrete identification procedure for 'evidence' relationships between sentences, where no linguistic clues state the relationships. Also, since relationships only between successive sentences are considered, the scope which the relationships cover cannot be analyzed, even if explicit connectives are detected.

This paper discusses a practical procedure for analyzing the discourse structure for Japanese text. The authors present a machine analyzer for extracting such structure, whose main component is a structure analysis using thinking-flow restrictions for argumentation. These restrictions, which examine possible sequences of relationships extracted from connective expressions in sentences, indicate which sentences should be

grouped together.

2 Discourse structure of Japanese text

2.1 Discourse structure

This paper focuses on analyzing discourse structure, representing relationships between sentences. In text, various rhetorical patterns are used to clarify its principle of argument. Among them connective expressions, which state inter-sentential relationships, are the most significant. They can be divided into several categories.

Here, connective expressions include not only normal connectives such as "therefore", but also idiomatic expressions stating relations to the other part of text such as "in addition" and "here ... is described." The authors have extracted 800 connective expressions from a preliminary analysis of more than 1,000 sentences [Ono 89]. Then, connective relationships were classified into around 20 categories as shown in Table 1. Using these relationships, linguistic structures of articles can be captured.

Sentences of similar content may be grouped together into a block. Just as each sentence in a block serves certain roles, e.g., "serial", "parallel", and "contrast", each block in text serves similar functions. Thus, the discourse structure must be able to represent hierarchical structures, as well as individual relationships between sentences. In this paper, a discourse structure is represented as a binary tree, whose terminal nodes are sentences; sub-trees in the whole tree correspond to local blocks of sentences in text.

Figure 1 shows a paragraph from an article on "a zero-crossing rate which estimates the frequency of a speech signal," where underlined words indicate connective

expressions. Figure 2 shows an example of its discourse structure. Extension relationships are set to sentences without any explicit connective expressions. Although the fourth and the fifth sentences are the exemplifications of the first three sentences clearly, the sixth is not. Thus, the first five can be grouped into a block, and the structure shown in Figure 2 is more natural than others

(e.g., (((P <EX> Q) <EX> R) <EG> ((S <EX> T) <SR> U))).

Table 1: Connective relationships.

RELATION		EXAMPLES and EXPLANATION
serial connection	<SR>	だから(thus, therefore), よって(then)
negative connection	<NG>	だが(but), しかし(though)
reason	<RS>	なぜなら(because), その訳は(the reason is ...)
parallel	<PA>	同時に(at the same time), さらに(in addition)
contrast	<CT>	一方(however), 反面(on the contrary)
exemplification	<EG>	例えば(for example), ... 等である(and so on)
repetition	<RP>	というのは(in other words), それは(it is ...)
supplementation	<SP>	もちろん(of course)
rephrase	<RH>	つまり, すなわち(that is ...)
summarization	<SM>	結局(after all), まとめると(in sum)
extension	<EX>	これは(this is)
definition	<DF>	ここで ... とする(... is defined as ...)
rhetorical question	<RQ>	なぜ ... なのだろうか(Why is it ...)
direction	<DI>	ここでは ... を述べる(here ... is described)
reference	<RF>	図Xに ... を述べる(Fig.X shows ...)
topic shift	<TS>	さて, ところで(well, now)
background	<BG>	従来(hitherto)
enumeration	<EN>	第一に(in the first place), 第二に(in the second place)

2.2 Local constraint for consecutive relationships

For analyzing a discourse structure, a local constraint on consecutive relationships between blocks of sentences is introduced. The example shown in Figure 1 and Figure 2 suggests that the sequence of connective relationships can constrain the discourse structure from a natural argumentation viewpoint. Consider the sequence

(P <EG> Q <SR> R), where P, Q, R are arbitrary (blocks of) sentences. The premise

of R is obviously not only Q but a block of P and Q. Since the argument in P and Q is considered to close locally, the two should be grouped into a block. This is a local constraint on natural argumentation.

Thinking-flow is defined by a sequence of connective relationships and the allowable structure versus the sequence. The authors have investigated all 324 (18×18) pairs of connective relationships and derived possible local structures for thinking-flow restrictions. They can be classified into the following four major groups, where the relations r1 and r2 are arbitrary connective relationships.

(1) POP-type : permitting ((P r1 Q) r2 R) (eliminating (P r1 (Q r2 R)))

ex. ((P <EG> Q) <SR> R), <EG> : exemplification, <SR> : serial.

(2) PUSH-type : permitting (P r1 (Q r2 R))

ex. (P <RS> (Q <SR> R)), <RS> : reason.

(3) NEUTRAL-type : permitting both (1) and (2)

ex. ((P <PA> Q) <EG> R), (P <PA> (Q <EG> R)), <PA> : parallel.

(4) NON-type : permitting non-structure (P r1 Q r2 R)

ex. (P <PA> Q <PA> R).

The relationship sequence of POP-type means that the local structure for the first two blocks should be popped up, because the local argument is closed. On the other hand, the relationship sequence of PUSH-type means that the local structure should be pushed down.

The thinking-flow restrictions can be used to eliminate structures expressing unnatural argumentative extensions, by examining their local structures. For example,

the structure (P <EG> (Q <SR> R)) is eliminated, because it violates the above POP-type restriction. Although the thinking-flow indicates local constraints on relationships to neighbors, the scope of relationships is analyzed by recursively checking all local structures of a discourse structure.

P : In the context of discrete-time signals, zero-crossing is said to occur if successive samples have different algebraic signs.
Q : The rate at which zero crossings occur is a simple measure of the frequency content of a signal.
R : This is particularly true of narrowband signals.
S : For example, a sinusoidal signal of frequency F0, sampled at a rate Fs, has Fs/F0 samples per cycle of the sine wave.
T : Each cycle has two zero crossings so that the long-term average rate of zero-crossings is Z = 2F0/Fs.
U : Thus, the average zero-crossing rate gives a reasonable way to estimate the frequency of a sine wave.

Figure 1: Text example 1.

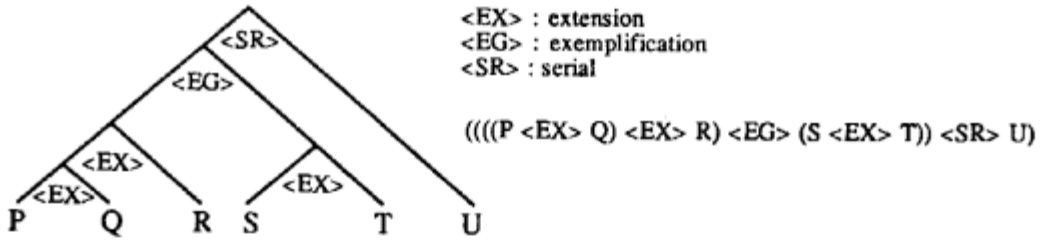


Figure 2: Discourse structure for the text example 1.

2.3 Distant dependencies

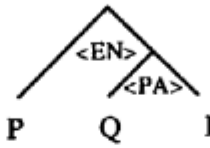
The greater part of text can be appropriately analyzed, using the above local constraints on connective relationships to neighbors, if the relationships are extracted correctly. However, in real text, there are rhetorical dependencies concerning distant sentences, which cannot be detected only within the normal relationships to neighbors. Two kinds of linguistic clues to distant dependencies must be considered to embody a precise discourse analyzer: rhetorical expressions which cover distant sentences, and referential relations of words, in particular, *topics*.

(a) Rhetorical expressions stating global structure First, rhetorical expressions which relate to an entire article play an important role. Their examples are : "... ? ... ? The reason is, ...", "... as follows. ...(TENSE=present). ...(TENSE=present).", "... is not an exceptional case. ... ...". Consider a text example in Figure 3, in which unnecessary minute words are omitted for expositional clarity. In this text the rhetorical expressions which relate to the entire paragraph affect its discourse structure. The expressions "first" and "second" in the last two sentences correspond to the expression "two pieces" in the first sentence. Therefore, the second and the third sentences must be connected by parallel relationships, because they have similar relations with the first sentence. Thus, in this case, the discourse structure in Figure 4 is natural.

While, in real text, there are a wide variety of rhetorical expressions of this type, those that are often used in argumentative articles can be given in advance. A practical discourse analysis system must detect these rhetorical expressions to restrict discourse structures.

P : Two pieces of X are relevant.
Q : First, ...
R : Second, ...

Figure 3: Text example 2. (X is a noun phrase.)



<EN> : enumeration
<PA> : parallel

Figure 4: Discourse structure for the text example 2.

(b) Topic flow The other significant phenomenon concerning the distant dependencies is *reference*. In Japanese, usually the same or partially the same noun phrases are used for referring to some other part of the text, while in English pronouns and definite noun phrases are used. By analyzing the appearance of the same expressions,

P : AはBとCからなる。
A consists of B and C.
Q : Cは ... DとEに分けられる。
The C is divided into D and E.
R : Dは ... Fを持つ。
The D has ... F.
S : Fは ... 。
The F is ... .

Figure 5: Text example 3.
(A - F are noun phrases.)

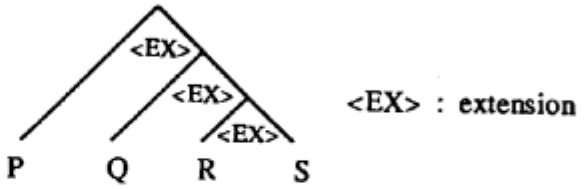


Figure 6: Discourse structure for the text example 3.

a restriction or a preference for building discourse structures can be acquired. However, the same expressions tend to scatter in a text, and it is difficult to determine the referent for a referer without task-dependent knowledge [Sumi 91]. This is contrary to the authors' aim in this paper. Thus, the appearance of the same expressions is to used as a preference for structure determination.

Figure 5 shows a text example in Japanese, where the underlined words are the same expressions. Note that many underlined words are followed by the character " は (*wa*)", and they refer to other previous words. This character is a postpositional word topicalizing the preceding noun in a sentence.

A *topic* of a sentence is an object indicating what the sentence is about; it can localize the reader's attention in the area that the object relates to. In contrast to topic processing for English (cf. [Shan 77], [Sidn 83]), we can use a linguistic device to extract topics for Japanese; some postpositional words are said to indicate a topic of a sentence [Naga 86].

In this paper, topic information is used for preference judgment of discourse structures, but not as an element of the structures. To simplify explanation, let us denote a topic of the sentence Q by T^{Q}, and a case where T^{Q} refers to a word in the previous

sentence P by $T^{Q} \Rightarrow$ P. In the case of the text shown in Figure 5, $T^{Q} \Rightarrow$ P, $T^{R} \Rightarrow$ Q, and $T^{S} \Rightarrow$ R hold. If a topic in a sentence refers to a word in the previous sentence, it is regarded as an elaboration of the earlier sentence. Thus, these sentences must be kept close together in their discourse structure, and therefore a structure depicted in Figure 6 is appropriate for this text.

In addition, relative importance of sentences connected by a relationship, which depends on the relationship, must be considered for the topic flow analysis. Connective relationships can be classified into three categories according to their relative importance: left-hand, right-hand, and neutral type. For example, exemplification relationship is a left-hand type; i.e., for (P <EG> Q), P strongly relates to the global flow of argumentation beyond the outside of this block, and in this sense P is more important than Q. Also, serial relationship is a right-hand type, and parallel relationship is a neutral type, at least in Japanese.

Consider the structure ((P r1 Q) r2 R), where 'r1' is a left-hand type relationship, and 'r2' can be any relationship. If $T^{R} \Rightarrow$ P, the above structure is natural, even if there is the same word as T^{R} in Q. However, if $T^{R} \Rightarrow$ Q, this structure is unnatural, in the sense of coherency. In this case, the structure (P r1 (Q r2 R)) is preferable to ((P r1 Q) r2 R).

On the contrary, in the case where 'r1' is a right-hand type, ((P r1 Q) r2 R) is a natural structure, even if $T^{R} \Rightarrow$ Q. In short, the naturalness of a discourse structure closely depends on the appearance position of topics and their referents, and the relative importance of the referred nodes.

3 Discourse structure analyzer

3.1 System configuration

The discourse structure analyzer consists of five parts, as shown in Figure 7.

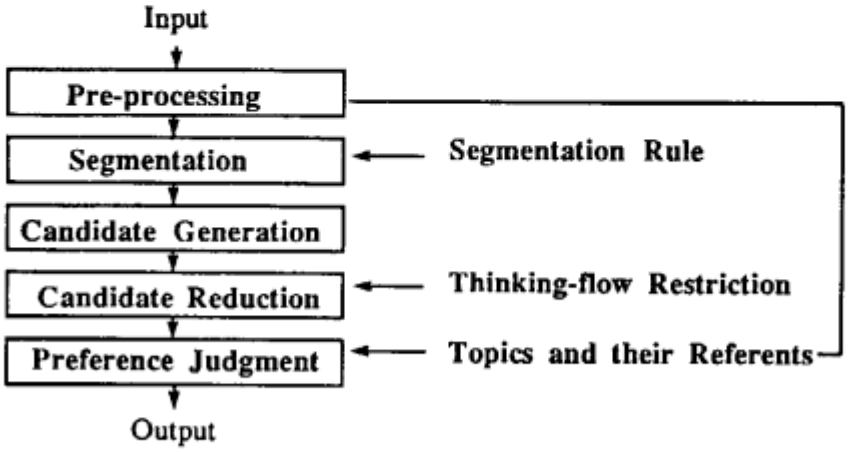


Figure 7: System overview.

(a) Pre-processing In this stage, input sentences are analyzed, character strings are divided into words, and the dependency structure for each sentence is constructed. The stage consists of the following sub-processes :

(1) Extracting the body of an article from chapters or sections.

(2) Accomplishing morphological and syntactic analysis.

(3) Extracting topic expressions and the appearance of the same expression.

(4) Detecting connective relationships and constructing their sequence.

In Step (1), the title of an article is eliminated, and the body is extracted. Next, in Step (2), sentences in the body of the article, extracted in Step (1), are morphologically and syntactically analyzed. In Step (3), topic expressions are extracted, according to a table of topic denotation expressions. The following are the examples of topic

expressions.

... *wa* (as for ...), ... *niwa*, ... *dewa*, ... *nioitewa* (in ...),

In Step (4), a connective expression is detected based on an expression table consisting of a word and its part of speech for individual connective relationships. In this step, *connection sequence* is acquired, which is a sequence of sentence identifiers and connective relationships. For example, a connection sequence is of the form `(P <EN> Q <EX> R <EN> S)`.

(b) Segmentation In this stage, rhetorical expressions between distant sentences, which restrict a discourse structure, are detected. They form restrictions on segmentation.

This stage is implemented as a rule-based procedure [Ono 91]. *If-then* rules, called *segmentation rules*, have been formulated in advance. The *if*-part of a segmentation rule corresponds to linguistic surface patterns to detect inter-sentence rhetorical expressions, e.g. "as follows. ... First ... ... Second ...". On the other hand, the *then*-part represents a connection sequence embedded with control operators such as '{' and '}', which direct a start or an end position of a block of sentences. Also, the *then*-part can indicate an exchange of connective relationships. Thus, an output example of this stage is of the form `(P <EN> {Q <EX> R <PA> S})`. At present, approximately 100 rules are available for the system.

(c) Candidate generation All possible discourse structures, described by binary-trees which do not violate segmentation restrictions, are generated as discourse structure candidates. The generation is performed in a bottom-up manner of parsing a

sentence by the well-known CYK algorithm. After the generation of sub-trees for blocks directed by segmentation restrictions, the whole trees are generated based on these sub-trees.

(d) Candidate reduction Local structures of generated structure candidates are checked by inspecting thinking-flow restrictions. The candidates including a local structure violating the restrictions are discarded. Only permitted candidates are passed on to the next stage.

The thinking-flow restrictions are represented as a table of the applicable pairs of consecutive relationships and the acceptable local structure for them.

(e) Preference judgment After a penalty calculation based on topics and their referents, a structure candidate with the lowest penalty is selected as a final result of discourse structure.

A penalty is set against each arc of path on a discourse structure, which leads from a sentence containing a topic to a sentence referred to by the topic. The concrete arc of a discourse structure, on which a penalty is imposed, is either an arc from or to an unimportant node or an arc to an equally important node. For example, for the structure ((P <EG> Q) <EX> R) where $T^{\mathrm{R}} \Rightarrow$ Q, a penalty is imposed on the arc from the parent node of P and Q to Q because the left node for exemplification relationship is unimportant.

The penalty of a discourse structure is defined as a sum of penalties for all paths concerning all topics in the paragraph. Finally, by selecting a structure candidate with the lowest penalty, the most coherent discourse structure is obtained.

3.2 Experiment

To evaluate the discourse structure analyzer, 18 journal articles, different from the data used for algorithm development, have been analyzed. The journal used is "Toshiba Review", which publishes technical short papers of a couple of pages. An experiment has been carried out on every paragraph.

Table 2 shows analysis results. There are a total of 554 paragraphs. Nealy 50% of them consist of only one sentence, and they are excluded from consideration. For 114 paragraphs consisting of more than three sentences, a correct analysis has been obtained for approximately 74% of them.

There have been 15 errors for all of the processed paragraphs. Most of the errors are due to an incorrect detection of relationships (60%), or an incorrect candidate reduction using thinking-flow restrictions (27%). For the former, the procedure has failed to detect explicit connective expressions because of insufficient dictionary data, which can be improved by refining the dictionary data. Most of the latter errors occur in a paragraph, in which the first or last sentences is related with the outside of the paragraph by such phrases as "as shown above" or "as follows." This suggests that the procedure should also take into account relationships to the outer paragraph as an important factor.

The segmentation stage activate segmentation rules for 35 paragraphs, and 85% of the rules have correctly used; 65% have contributed to structure determination for itemized parts of text, and 20% to relationship determination. In addition, the preference judgment stage has increased the accuracy of the analysis by 3%. Except for the effects of these contributions, correct relationships have been detected in 73

paragraphs, and correct results have been obtained for 55 paragraphs. Thus, if correct connective relationships are detected, 73% of discourse structures can be appropriately analyzed using thinking-flow restrictions only.

Table 2: Analysis results

paragraph size (number of sentences)	correct* (unique)	correct* (other candidate)	incorrect*	Total*
1	-	-	-	293
2	-	-	-	147
3	53	8	6	67
4	12	5	7	24
5	7	1	2	10
6	3	0	0	3
7	5	0	0	6
8	2	0	0	2
9	2	0	0	2
Total	84	14	15	114^{+}(554)

* Numbers indicate counts of paragraphs, except for #sentence.
\+ Total number of paragraphs consisting of more than 3 sentences.

4 Concluding remarks

A practical analyzer has been described for building discourse structures for Japanese argumentative or explanatory articles. To analyze structures, three types of knowledge, thinking-flow restrictions, segmentation rules, and topic-flow preference, are used. They represent relative constraints between connective relationships or structural restrictions spanning a paragraph, unlike the relative importance between consecutive sentences, on which other researchers on discourse structure analysis depend. Using linguistic knowledge, global structures or the scope of relationships can be determined appropriately.

In addition, the above knowledge on which the procedure is based is detected from superficial linguistic clues, independent of topic areas in analyzed articles. The authors are convinced that the method is applicable to any articles whose essential aim lies in

persuasion of assertion.

It should be noted that the relative importance of sentences can be evaluated, using the extracted discourse structure. For example, a left-hand node of a structure linked by exemplification relationship is more important than the right-hand node, as discussed in Section 2.3(b). By a recursive application of relative importance judgment from the top node of discourse structure analyzed from a paragraph, the key-sentence in the paragraph can be extracted.

Besides the key-sentence extraction as shown above, the extracted structure can be a promising clue to other various NL processings, such as topic estimation, and knowledge extraction. The authors intend to polish up the presented restrictions and rules, and refine the procedure toward these NL processings.

Reference

[Cohe 87] Cohen, R.: "Analyzing the Structure of Argumentative Discourse", *Computational Linguistics*, Vol.13, pp.11–24, 1987.

[Gros 86] Grosz, B.J. and Sidner, C.L.: "Attention, Intentions and the Structure of Discourse", *Computational Linguistics*, Vol.12, pp.175–204, 1986.

[Hobb 79] Hobbs, J.R.: "Coherence and Coreference", *Cognitive Science*, Vol.3, pp.67–90, 1979.

[Litm 87] Litman, D.J. and Allen, J.F.: "A Plan Recognition Model for Subdialogues in Conversations", *Cognitive Science*, Vol.11, pp.163–200, 1987.

[Mann 87] Mann, W.C. and Thompson, S.A.: "Rhetorical Structure Theory: A Framework for the Analysis of Texts", *USC/Information Science Institute Research Report* RR–87–190, 1987.

[Naga 86] Nagano, K.: *Bunshouron Sousetsu —Bunpouron-teki Kousatsu—* (An Introduction to Theory of Texts —Syntactic Consideration—), Asakusa Shoten, 1986, (in Japanese).

[Ono 89] Ono, K., Ukita, T., and Amano, S.: "An Analysis of Rhetorical Structure", *IPS Japan Technical Report NL* 70–2, 1989, (in Japanese).

[Ono 91] Ono, K., Sumita, K., Ukita, T., and Amano, S.: "Text Segmentation and Discourse Analysis", *Proc. IPS Japan '91* October, 4E–2, 1991, (in Japanese).

[Shan 77] Schank, R.C.: "Rules and Topics in Conversation", *Cognitive Science*, Vol.1, pp.421–441, 1977.

[Sidn 83] Sidner, C.L.: "Focusing in Comprehension of Definite Anaphora", M.Brady and R.C.Berwick (Eds.), *Computational Models of Discourse*, MIT Press, pp.267–330, 1983.

[Sumi 91] Sumita, K., Ukita, T., and Amano, S.: "Disambiguation in Natural Language Interpretation Based on Amount of Information", *IEICE Trans.*, E74, 6, pp.1735–1746, 1991.