TM-0657

Computational Analysis of Linguistic Discourse Structure for Japanese Text

by
T. Ukita, K. Ono & S. Amano (Toshiba)

December, 1989

ICOT

Mita Kokusai Bldg. 21F
4-28 Mita 1-Chome
Minato-ku Tokyo 108 Japan

(03) 456-3191～5
Telex ICOT J32964

Institute for New Generation Computer Technology

Computational Analysis of Linguistic Discourse Structure for Japanese Text*

Teruhiko Ukita, Kenji Ono, and Shin'ya Amano
Toshiba Corp. R&D Center
Komukai-Toshiba-Cho 1, Saiwai-ku, Kawasaki, 210 Japan
tel. (044) 549-2240 Japan

Summary:

This paper discusses a computational analyzer for linguistic discourse structure, defined as relationships between sentences, in a manner similar to the Rhetorical Structure proposed by Mann and Thompson [Mann and Thompson 87]. A practical procedure to extract discourse structure, using conjunctive expressions and topic presenting expressions, is presented and applied in analyzing journal articles in Japanese.

* 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 in Japan.

1. Introduction

A computational theory for use in analyzing linguistic discourse structure and its practical procedure are required to develop a machine system to deal with plural sentences; e.g., systems for text summarization and for appropriate conversation flow management. Hobbs developed a theory in which he arranges three kinds of relationships between sentences from the viewpoint of text coherency [Hobbs 79]. Grosz and Sidner proposed a theory which accounts for relationships between three notions (linguistic, intention, and attention) [Grosz and Sidner 85]. These theories, though, require extensive efforts for us to realize practical procedures, because precise identification of various relationships require an inference function with a knowledge base. On the other hand, linguistic structure of text was proposed which describes relationships between sentences and the relative importance of the related sentences [Mann and Thompson 87], which was then applied to text generation [Hovy 88; Tabuchi, et al. 88].

This paper discusses linguistic discourse structure which describes relationships between sentences and the possibility of building a machine parser for extracting such structure; a parser using both conjuntive expressions and topic presenting expressions, whose subject material is composed of journal articles written in Japanese.

2. Japanese text discourse structure

The structure for a chunk of text must represent the relationships between sentences which are stated by explicit conjunctive expressions as well as implicit "causal chains". Based on the preceding researches on the linguistic structure [Mann and Thompson 87; Tokoro 86], this paper focuses on the linguistic phenomenon of conjunctions between sentences, although the complete extraction of the text structure might require human inference ability.

Based on a preliminary analysis carried out on more than 1000 sentences, the authors extracted about 800 conjunctive expressions used in Japanese language sentences (throughout the paper, a character string between periods is called a sentence). Table 1 shows examples of such conjunctive expressions. Since several some different kinds or levels of expressions can be observed, such as those for explaining the preceding sentence and those for logical connectivities, two connectivity levels are distinguished in this paper. They are;

Table 1 Conjunctive expression examples

RELATION	EXAMPLES
(1) exemplification	*tatoeba* (for example), ... *nadode aru* (... and so on)
(2) repetition	*to iunowa* (in other words), *sorewa* (it is...)
(3) reason	*nazenara* (because), *sono wakewa* (the reason is...)
(4) supplementation	*mochiron* (of course), *kokode XXXtowa* ... *dearu* (here, XXX means ...)
(5) parallel category	
(5-1) parallel	*doujini* (at the same time), *sarani* (in addition)
(5-2) contrast	*ippou* (however), *hanmen* (to the contrary)
(6) serial category	
(6-1) affirmative connection	*dakara* (thus), *soshite* (and), *yotte* (then)
(6-2) negative connection	*daga* (but), *shikashi* (though)
(7) rephrase	*tsumari, sunawachi* (that is)
(8) meta categories	
(8-1) direction	*kokodewa ...wo noberu* (here described ...) *zu X-ni ...wo shimesu* (Fig. X shows ...)
(8-2) topic shift	*sate, tokorode* (well, now)
(8-3) summarization	*kekkyoku* (anyway), *matomeruto* (in summary)

(a) Statement level,

(b) Thinking level.

The first level is designed so that sentences which describe "one thing" are gathered into one group with a structure, representing a modifying relationship between sentences. This structure level roughly corresponds to a "paragraph" in a text, and consists of one central statement and supporting sentences for explanations and examples. Basic relations for this level and the relative importance of sentences are defined as follows.

(1) ***exemplification*** : This modification explains the central statement by describing examples; the central statement according to this modification is the preceding part of the current sentence.

(2) ***repetition*** : A modification stressing almost the same contents as the preceding sentence; the central part is either the preceding and the current sentences (for a summarizing task, the procedure should accomplish abstraction from these sentences).

(3) ***reason*** : A modification presenting the cause or reason for the central statement; the central statement is the preceding part of the current sentence

(4) ***supplementation*** : A modification supplying additional information; the central statement is the preceding part of the current sentence.

The second level is for describing the thinking flow. Though this level seems to be specific regarding the charcteristics for the text materials, texts whose main aim is transmitting some information to the reader can be considered to have a thinking flow, except materials such as poems, whose major value might exist in their character strings themselves. This level of structure represents a thinking flow and has a central statement (≈ paragraph) as a unit, thus described as a node of the structure tree. In this level, four different relationships are distinguished.

(5) ***parallel category***: This category is classified into the subsequent two categories; the central part is both preceding and current statements.

(5-1) ***parallel*** : This relationship enumerates similar contents.

(5-2) ***contrast*** : This adds a different sort of contents.

(6) ***serial category***: This is classified into two subsequent categories and describes thinking development; the central statement is the preceding part of the current sentence.

(5-1) ***affirmative connection***: This is an ordinary connection with positive sense.

(5-2) ***negative connection***: This is an ordinary connection with negative sense.

(6) ***rephrase***: This shows interruption in thinking flow and rephrases a similar statement as in the preceding context; the central part is either the preceding and the current statements.

(7) ***meta categories***: This includes direct indications of some information conveyed from a writer to a reader.

(7-1) ***direction***: This includes "direction" and "references" conveyed directly to a reader; the central part is the preceding part of the current statement.

(7-2) ***topic shift***: This indicates change in text topics; the central part is any of the preceding and current statements.

(7-3) ***summarization***: This indicates the current statement (and its subsequent sentences) is a summary of the preceding context; the central part is the succeeding part of the current sentence.

The structure built from these relationships is a binary tree whose nodes are statements (or sentences). An example of the structure is shown in Section 5.

3. Extraction of connectivities from the text

Based on the list of conjunctive expressions, as shown in Table 1, a procedure is built which detects relationships between sentences, when they exist, and builds discourse structure. Using the relations and some heuristic rules and considering all possible combinations of statements, the procedure analyzes the discourse structure in a bottom-up manner and produces a binary tree, as a result. The following assumption and rules are used in this procedure;

[Assumption] Relationships between sentences are non-crossing, i.e., a sub-structure, consisting of consecutive sentences, is closed, independently from their context.

[RULE1] The following four combinations are prohibited ("X" is a sentence or a sub-structure and "((..(" means one or more parentheses):

(....... serial relation ((..(X serial relation ....)..)),
(....... rephrase relation ((..(X serial relation ....)..)),
(....... direction relation ((..(X serial relation ....)..)),
(....... direction relation ((..(X parallel relation ....)..)).

The prohibition for the first combination, for example, means that a string of statements combined by serial relation like "thus" should be formalized into ((A thus B) thus C), where A,B, and C are statements and the original string is "A. Thus B. Thus C.".

[RULE2] The following three combinations are less prefered to other combinations:

(....... serial relation ((..(X parallel relation ....)..)),
(....... rephrase relation ((..(X parallel relation ....)..)),
(....... parallel relation ((..(X same rank relation ....)..)).

[RULE3] When no conjuntive expressions are found for a sentence, the procedure manages the sentence as though it has relationships which belong to ***parallel*** or ***serial*** categories.

4. Using topic presenting expression for discourse structure

4.1 Topic presenting expression in Japanese text

The "topic" of a sentence is an object which the sentence describes. A topic can localize a reader's attention to the area that the object relates to. Thus, this implies that an economical expression can be obtained in the sense of Grice's maxim of quantity [Grice 75]. Concerning linguistic text structure, topics and a trace of topics can be viewed as auxiliary information, which indicates the relationships between sentences.

Japanese is an agglutinative language and a phrase consists of content words (nouns, verbs, adjectives, adverbs, etc.) and (optional) accompanying functional words, which denote "case", "tense" and so on (auxiliary verb, postpositonal words, etc.). Some postpositional words are said to express the topic for the sentence [e.g., Nagano 87]. A pertinent representative is the postpositional word "*wa*", which denotes the topic of a sentence that functions as a subject in a sentence (e.g., "*kore-wa*" ≈ "this is") or an adverbial phrase in combination with other postpositional words ("*kono ronbun de wa*" ≈ "in this paper"). This topic presenting expression has been used in supplementing zero pronouns in Japanese [Kameyama 86; Yoshimoto 88].

As a preliminary analysis of topic presenting expressions, about 200 sentences in Japanese were analyzed for the postpositional word "*wa*". The word "*wa*" was picked up from the character strings and the preceding nouns, including compound words, were detected as "topic". At the same time, repetition of the "topic" word is searched for in the preceding context inside the sections. As major results of the analysis, for the 202 sentences, 63 sentences do not have "*wa*" expressions and have no other topic representation expressions. Fifty occurrences have no repetition of words in the preceding context and all of them are considered to be apparent notions to the readers (e.g., "Fig. 3 is ..."). The next categories are full or partial repetitions of the topic words in the preceding context (48 occurrences). In these cases, the expression of the topic are not so familiar to a reader, so that the reader must determine concepts which have already appeared. From the view point of detecting relationships between sentences, they can be used as a cue to state the strong relationships between sentences. The last major category of "*wa*" expressions is pronouns and definite expressions, which are apparently considered to be the repetitions of the already presented concept. Though their precise correspondence is difficult to identify, it can be said that a referent for a pronoun or a definite noun in a topic expression should appear in a sentence after the previous topic presentation.

4.2 Using topics for discourse structure

In building the discourse structure, topic presenting expressions with word repetitions and topicalized pronouns can be used as follows ([RULE4]):

(a) If a topic word was already presented in the preceding context with the same character string, then the current sentence has a direct relationship* to the previous sentence having the same word (*direct relationship is a relation that connects two nodes with one branch in a structure tree).

(b) If a topic expression is in the form of "anaphora indicator + *wa*", such as "*sono, konoyouna* (the,these) + compound noun + *wa*", then the current sentence has a direct relationship to some of the sentences after the one having the topic expression (here cataphora possibility is ignored).

(c) If the current sentence has no topic expression, then there is no cue as to what the current sentence can be related to.

These heuristic rules can be used to make a structure tree in combination with rules described in Section3.

5. A discourse structure analysis example

To evaluate the procedure involved in building discourse structure proposed in this paper, two articles of a journal were analyzed by hand. To illustrate the process, let's investigate a text chunk from a complete section. In this analysis, it is assumed that morphological analysis is correctly performed and that the structure of the statement level can be built for a paragraph. In the example, words with underline indicate conjunctive expressions and words with bold font represent topics denoted by "*wa*" in the original Japanese text .

(1-1) **Thermoelecric power plants** are required to operate in various modes and perform load-following duty. (1-2) Under such severe conditions, it is absolutely essential that each piece of equipment in the plant operate without problems, providing a high-level of long-term plant reliability.

(2-1) On the other hand, life of each piece and the time between periodical examinations have been invetigated to lengthen them by rationalizing maintenance affairs and inspection for operation and preventing maintenance. (2-2) **To accomplish this**, it is necessary to develop a diagnostic system with such funtions as early malfunction detection of equipment in operation, assumption of the cause of malfunction while providing data for corrective-action decisions, verification of historical tendency, maintenance support and life consumption estimation.

(3-1) **These judgements** have been arrived at by an operator based on his experience with general principles. (3-2) At present, though, the quantity of information which requires judgement is growing rapidly because of growing diverse applicatons and higher functions for a system.

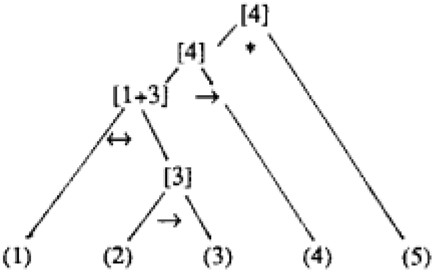
(4-1) For these reasons, rapid development of a diagnostic system which can manage information for operation and maintenance, is expected.

(5-1) An outline is introduced below for a diagnostic system for thermoelectric power plant and its practical examples.

Based on conjunctive expressions, the relationships between sentences and relationships between central statements (paragraphs) are as follow (the preference rule [RULE 2] was used as the same way as the prohibition rule):

(1-1) → (1-2) ↔ (2-1) → (2-2) → (3-1) → (3-2) → (4-1) * (5-1)
(1) ↔ (2) → (3) → (4) * (5)
(symbols: → affirmative or negative conn., ↔ contrast, * direction)

Since a paragraph contains at most two sentences and structures inside paragraphs do not suffer from ambiguous combinations, let us see the relationships between paragraphs, for the simplicity. There are 14 possible combinations of paragraphs (central statements) as discourse structures which satisfy the assumption of non-crossing. The topic presenting expressions indicates that there is a direct relationship between paragraphs (2) and (3). Applying rules of conjunctive expressions and topics, we obtain only the structure which has the following form for the paragraphs:



((((1 ↔ (2 →3)) → 4) * 5)

[N] shows that central part is statement N.
[M+N] means that central part is the combination of sentences M and N.

Table 2 Analysis results for sentences inside sections

	no. sentences or paragraphs	possible structures	structures (Case1)	structures (Case2)
1*	5	14	2	1
2	3	2	2	2
3	3	2	2	2
4	5	14	10	5
5*	6	42	4	4
6	4	5	5	5
7	4	5	1	1
8	3	2	1	1
9	5	14	9	9
10*	4	5	2	1
11*	3	2	1	1
12	3	2	2	2
13*	3	2	1	1
15	3	2	2	2
16	4	5	5	2
17*	4	5	2	2
18	3	2	2	1
19*	3	2	2	2
20	3	2	2	2
21	4	5	5	5
23	5	14	1	1
24	5	14	14	14
25*	3	2	2	2
26	3	2	2	1
27	3	2	1	1
28	8	429	260	10
29	8	429	58	28
30	10	4862	1358	264
31	9	1430	1430	1430
32	5	14	5	2
33	5	14	11	5
34	6	42	42	42
35	8	429	359	76

* Results for inter-paragraph structures.
Possible structures are those which satisfy the non-crossing assumption.
Case 1: Results for [RULE1]~[RULE3]
Case 2: Results for [RULE1]~[RULE4]

For other data consisting of two articles from Japanese language journal, analyzed results are summaried in Table 2. For the total of 150 sentences, the procedure described in this paper was applied manually using a preprocessor for screening character strings, where it was assumed that morphological analysis was correctly performed. Some ambiguities remain in discourse structures, especially for sentences having few conjuntive expressions. However, overall performance is indicative that it is possible to obtain discourse structures mechanically from the real corpus of data.

6. Concluding remarks

The computational approach toward building an automatic extraction of a discourse structure is described. The discourse structure is defined by relationships between sentences and built from conjunctive expressions between sentences as well as topic presenting postpositional words. Some manual analyses were accomplished for Japanese language journal articles and some positive prospects were obtained. Though materials managed in this paper are those written in Japanese, the basic framework is applicable to other languages, since the level of thinking can be considered to share a common tendency for different languages, whereas a list of conjunctive expressions and topic extraction algorithm should be modified for a different language (for English, [Webber 80; Joshi and Weinstein 81; Sidner 83; Grosz et al. 83]).

As an application of this discourse structure, introducing the content selection of a binary operator, such as selecting the latter sentence for an affirmative connection, a rough summary of the text can be easily obtained from the discourse structure. For example, in the example of Section 5, the fourth paragraph is correctly selected as the most important statement.

The current problem in the procedure is that it can do nothing for a sentence which has neither conjunctive expressions nor topic presenting expressions (e.g., sample 31 in Table2). Appropriate management of such a sentence requires an inference mechanism with a knowledge-base, which can find a causal relationship between sentences. The structure and the procedure proposed in this paper can be positioned as an elementary step for this ultimate stage.

REFERENCES

[Grice 75] Grice, H.P.: "Logic and Conversation", Syntax and Semantics, Vol.3, Speech Act, Seminar Press, pp.41-58, 1975.

[Grosz, et al. 83] Grosz, B.J., Joshi, A.K., and Weinstein, S.: "Providing a Unified Account of Definite Noun Phrases in Discourse", Proc. 21st Annual Meeting of the Association for Computational Linguistics, pp.44-50, 1983.

[Grosz and Sidner 85] Grosz, B.J. and Sidner, C.L.: "Discourse Structure and the Proper Treatment of Interruptions", Proc. IJCAI-85, pp.832-839, 1985.

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

[Hovy 88] Hovy, E. H.: "Planning Coherent Multisentental Text", Proc. 26th Annual Meeting of the Association for Computational Linguistics, pp.163-169, 1988.

[Joshi and Weistein 81] Joshi,A.K. and Weinstein,S.: "Control of Inference : Role of Some Aspects of Discourse Structure -Centering-", Proc. IJCAI-81, pp.385-387,1981.

[Kameyama 86] Kameyama, M.: "A Property-sharing Constraint in Centering", Proc. 24th Annual Meeting of the Association for Computational Linguistics, pp.200-206, 1986.

[Mann and Thompson 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.

[Nagano 86] Nagano, K.: "Bunshouron Sousetsu --Bunpouron-teki Kousatsu--(An Introduction to theory of Texts--Syntactic Consideration--)", Asakura Shoten,1986 (in Japanese).

[Sidner 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.

[Tabuchi, et al. 88] Tabuchi,A, Tsujii,J, and Nagao,M: "Text Generation System Considering Context", Rep. Meeting Natural Language Processing, Information Processing Society of Japan, No. 88-NL-65, 1988 (in Japanese).

[Tokoro 86] Tokoro,K. : "Fundamentals of Rhetorical Grammar", Chap.3, Tokoro, How to Read Japanese, Takumi-Shuppan, 1986 (in Japanese).

[Webber 80] Webber,B.L.: "Syntax beyond the Sentence: Anaphora", in R.J.Spiro,et al.(Eds.), Theoretical Issues in Reading Comprehension, Lawrence erlbaum associates,Inc., 1980.

[Yoshimoto 88] Yoshimoto, K.: "Identifying Zero Pronouns in Japanese Dialogue", Proc. COLING-88, pp.779-784, 1988.