ABSTRACT GENERATION BASED ON RHETORICAL STRUCTURE EXTRACTION

Kenji Ono, Kazuo Sumita, Seiji Miike

Research and Development Center
Toshiba Corporation
Komukai-Toshiba-cho 1, Saiwai-ku, Kawasaki, 210, Japan
ono@isl.rdc.toshiba.co.jp

1 ABSTRACT

We have developed an automatic abstract generation system for Japanese expository writings based on rhetorical structure extraction. The system first extracts the rhetorical structure, the compound of the rhetorical relations between sentences, and then cuts out less important parts in the extracted structure to generate an abstract of the desired length. Evaluation of the generated abstract showed that it contains at maximum 74% of the most important sentences of the original text. The system is now utilized as a text browser for a prototypical interactive document retrieval system.

2 INTRODUCTION

Abstract generation is, like Machine Translation, one of the ultimate goal of Natural Language Processing. However, since conventional word-frequency-based abstract generation systems(e.g. [Kuhn 58]) are lacking in inter-sentential or discourse-structural analysis, they are liable to generate incoherent abstracts. On the other hand, conventional knowledge or script-based abstract generation systems(e.g. [Lehnert 80], [Fum 86]), owe their success to the limitation of the domain, and cannot be applied to document with varied subjects, such as popular scientific magazine. To realize a domain-independent abstract generation system, a computational theory for analyzing linguistic discourse structure and its practical procedure must be established.

Hobbs developed a theory in which he arranged three kinds of relationships between sentences from the text coherency viewpoint [Hobbs 79].

Grosz and Sidner proposed a theory which accounted for interactions between three notions on discourse: linguistic structure, intention, and attention [Grosz et al. 86].

Litman and Allen described a model in which a discourse structure of conversation was built by recognizing a participant's plans [Litman et al. 87]. These theories all depend on extra-linguistic knowledge, the accumulation of which presents a problem in the realization of a practical analyzer.

Cohen proposed a framework for analyzing the structure of argumentative discourse [Cohen 87], yet did not provide a concrete identification procedure for 'evidence' relationships between sentences, where no linguistic clues indicate the relationships. Also, since only relationships between successive sentences were considered, the scope which the relationships cover cannot be analyzed, even if explicit connectives are detected.

Mann and Thompson proposed a linguistic structure of text describing relationships between sentences and their relative importance [Mann et al. 87]. However, no method for extracting the relationships from superficial linguistic expressions was described in their paper.

We have developed a computational model of discourse for Japanese expository writings, and implemented a practical procedure for extracting discourse structure[Sumita 92]. In our model, discourse structure is defined as the rhetorical structure, i.e., the compound of rhetorical relations between sentences in text. Abstract generation is realized as a suitable application of the extracted rhetorical structure. In this paper we describe briefly our discourse model and discuss the abstract generation system based on it.

3 RHETORICAL STRUCTURE

Rhetorical structure represents relations between various chunks of sentences in the body of each section. In this paper, the rhetorical structure is represented by two layers: intra-paragraph and inter-paragraph structures. An intra-paragraph structure is a structure whose representation units are sentences, and an inter-paragraph structure is a structure whose representation units are paragraphs.

In text, various rhetorical patterns are used to clarify the principle of argument. Among them, connective expressions, which state inter-sentence relationships, are the most significant. The typical grammatical categories of the connective expressions are connectives and sentence predicates. They can be divided into the thirty four categories which are exemplified in Table 1.

Table 1: Example of rhetorical relations

Relation	Expressions
serial (<SR>)	*dakara* (thus)
summarization (<SM>)	*kekkyoku* (after all)
negative (<NG>)	*shikashi* (but)
example (<EG>)	*tatoeba* (for example)
especial(<ES>)	*tokuni* (particularly)
reason (<RS>)	*nazenara* (because)
supplement (<SP>)	*mochiron* (of course)
background (<BI>)	*juurai* (hitherto)
parallel (<PA>)	*mata* (and)
extension (<EX>)	*kore wa* (this is)
rephrase (<RF>)	*tsumari* (that is to say)
direction (<DI>)	*kokode wa ... wo noberu* (here ... is described)

The rhetorical relation of a sentence, which is the relationship to the preceding part of the text, can be extracted in accordance with the connective expression in the sentence. For a sentence without any explicit connective expressions, extension relation is set to the sentence. The relations exemplified in Table 1 are used for representing the rhetorical structure.

Fig. 1 shows a paragraph from an article titled "A Zero-Crossing Rate Which Estimates the Frequency of a Speech Signal," where underlined words indicate connective expressions. Although the fourth and fifth sentences are clearly the exemplification of the first three sentences, the sixth is not. Also the sixth sentence is the concluding sentence for the first five. Thus, the rhetorical structure for this text can be represented by a binary-tree as shown in Fig. 2.This structure is also represented as follows:

```
[[[1 <EX> 2] <ES> [3 <EG> [4 <EX> 5]]] <SR> 6]
```

1: In the context of discrete-time signals, zero-crossing is said to occur if successive samples have different algebraic signs.
2: The rate at which zero crossings occur is a simple measure of the frequency content of a signal.
3: This is particularly true of narrow band signals.
4: For example, a sinusoidal signal of frequency F_0, sampled at a rate F_s, has F_ϕ/F_s samples per cycle of the sine wave.
5: Each cycle has two zero crossings so that the long-term average rate of zero-crossings is $Z = 2F_0/F_s$.
6: Thus, the average zero-crossing rate gives a reasonable way to estimate the frequency of a sine wave.

(L.R.Rabiner and R.W.Schafer, *Digital Processing of Speech Signals*, Prentice-Hall, 1978, p.127.)

Figure 1: Text example

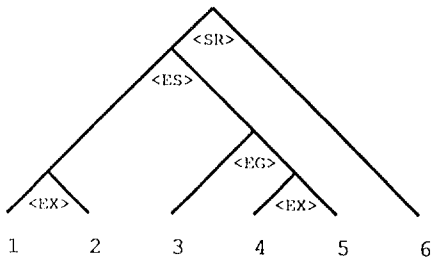

Figure 2: Rhetorical structure for the text in Fig.1

The rhetorical structure is represented by a binary tree on the analogy of a syntactic tree of a natural language sentence. Each sub tree of the rhetorical structure forms an argumentative constituent, just as each sub-tree of the syntactic tree forms a grammatical constituent. Also, a sub-tree of the rhetorical structure is sub-categorized by a relation of its parent node as well as a syntactic tree.

5 Implementation Note

The current version of TECHDOC is running on Sun Sparc stations with LUCID CommonLISP 1.4 and LOOM 1.41 (a port to LOOM 2.1 is underway), and a PENMAN version from 1991. The user interface is based on the CommonLISP Motif interface package CLM and the application building tool GINA [Spenke *et al.*, 1992].

Acknowledgements

The success of the TECHDOC project depended heavily on contributions from a number of student interns, in alphabetical order: Brigitte Grote, Sandra Kübler, Haihua Pan, Jochen Schoepp, Alexander Sigel, Ralf Wagner, and Uta Weis. They all have contributed to grammar or lexicon coverage in one way or another. Gerhard Peter has implemented TECHDOC-I, an interactive version giving car maintainance assistance. Thorsten Liebig has implemented TECHDOC's user interface for workstations using CLM and GINA, Hartmut Feuchtmüller has added multimedia facilities and mouse-sensitive text output. We also have to thank the PENMAN and LOOM groups at USC/ISI and the KOMET project at GMD Darmstadt, who gave us invaluable help.

References

[Bateman, 1990] John A. Bateman. Upper modeling: A level of semantics for natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Generation*, Pittsburgh, PA., 3 - 6 June 1990.

[Grote *et al.*, 1993] Brigitte Grote, Dietmar Rösner, and Manfred Stede. Representation levels in multilingual text generation. In Brigitte Grote, Dietmar Rösner, Manfred Stede, and Uta Weis, editors, *From Knowledge to Language – Three Papers on Multilingual text Generation*. FAW Ulm, FAW-TR-93017, 1993.

[LOOM, 1991] The LOOM Knowledge Representation System. Documentation package, USC/Information Sciences Institute, Marina Del Rey, CA., 1991.

[Mann and Thompson, 1987] William C. Mann and Sandra A. Thompson. Rhetorical structure theory: A theory of text organization. In L.Polanyi, editor, *The Structure of Discourse*. Ablex, Norwood, N.J., 1987. Also as USC/Information Sciences Institute Research Report RS-87-190.

[Rösner and Stede, 1992a] Dietmar Rösner and Manfred Stede. Customizing RST for the automatic production of technical manuals. In R. Dale, E. Hovy, D. Rösner, and O. Stock, editors, *Aspects of Automated Natural Language Generation - Proceedings of the 6th International WS on Natural Language Generation*, Lecture Notes in Artificial Intelligence 587. Springer, Berlin/Heidelberg, 1992.

[Rösner and Stede, 1992b] Dietmar Rösner and Manfred Stede. TECHDOC : A system for the automatic production of multilingual technical documents. In G. Görz, editor, *KONVENS 92*, Reihe Informatik aktuell. Springer, Berlin/Heidelberg, 1992.

[Spenke *et al.*, 1992] Michael Spenke, Christian Beilken, Thomas Berlage, Andreas Bäcker, and Andreas Grau. *GINA Reference Manual Version 2.1*. German National Research Center for Computer Science, Sankt Augustin, Germany, 1992.

that case the system cuts out terminal nodes from the last sentences, which are given the same penalty score.

If the text is written loosely, the rhetorical structure generally contains many *BothNucleus* relations (e.g., parallel(*mata*(and, also)), and the system cannot gradate the penalties and cannot reduce sentences smoothly.

After sentences of each paragraph are reduced, inter-paragraph structure reduction is carried out in the same way based on the relative importance judgement on the inter-paragraph rhetorical structure.

If the penalty calculation mentioned above is accomplished for the rhetorical structure shown in Fig. 2, each penalty score is calculated as shown in Fig. 3. In Fig. 3 italic numbers are the penalties the system imposed on each node of the structure, and broken lines are the boundary between the nodes imposed different penalty scores. The figure shows that sentence four and five have penalty score three, that sentence three has two , that sentence one and two have one, and that sentence six has no penalty score. In this case, the system selects sentence one, two, three and six for the longest abstract, and and also could select sentence one, two and six as a shorter abstract, and also could select sentence six as a still more shorter abstract.

After the sentences to be included in the abstract are determined, the system alternately arranges the sentences and the connectives from which the relations were extracted, and realizes the text of the abstract.

The important feature of the generated abstracts is that since they are composed of the rhetoricaly consistent units which consist of several sentences and form a rhetorical substructure, the abstract does not contain fragmentary sentences which cannot be understood alone. For example, in the abstract generation mentioned above, sentence two does not appear solely in the abstract, but appears always with sentence one. If sentence two appeared alone in the abstract without sentence one, it would be difficult to understand the text.

6 EVALUATION

The generated abstracts were evaluated from the point of view of key sentence coverage. 30 editorial articles of "Asahi Shinbun", a Japanese newspaper, and 42 technical papers of "Toshiba Review", a journal of Toshiba Corp. which publishes short expository papers of three or four pages, were selected and three subjects judged the key sentences and the most important key sentence of each text. As for the editorial articles, The average correspondence rates of the key sentence and the most important key sentence among the subjects were 60% and 60% respectively. As for the technical papers, they were 60% and 80 % respectively.

Table 2: Relative importance of rhetorical relations

Relation Type	Relation	Import. Node
RightNucleus	serial, summarization, negative, ...	right node
LeftNucleus	example, reason, especial, supplement, ...	left node
BothNucleus	parallel, extension, rephrase, ...	both nodes

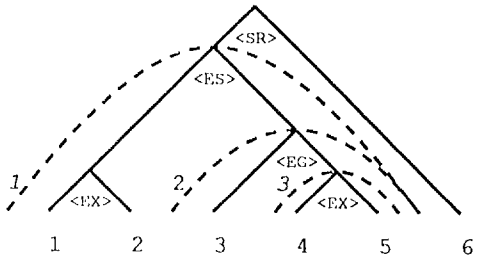

Figure 3: Penalties on relative importance for the rhetorical structure in Fig.2

Then the abstracts were generated and were compared with the selected key sentences. The result is shown in Table 3. As for the technical papers, the average length ratio(abstract/original) was 24 %, and the coverage of the key sentence and the most important key sentence were 51% and 74% respectively. Whereas, as for the editorials, the average length ratio(abstract/original) was 30 %, and the coverage of the key sentence and the most important key sentence were 41% and 60% respectively.

The reason why the compression rate and the key sentence coverage of the technical papers were higher than that of the editorials is considered as follows. The technical papers contains so many rhetorical expressions in general as to be expository.

That is, they provide many linguistic clues and the system can extract the rhetorical structure exactly. Accordingly, the structure can be reduced further and the length of the abstract gets shorter, without omitting key sentences. On the other hand, in the editorials most of the relations between sentences are supposed to be understood semantically, and are not expressed rhetorically. Therefore, they lack linguistic clues and the system cannot extract the rhetorical structure exactly.

Table 3: Key sentence coverage of the abstracts

Material	total num.	length ratio	cover ratio	
			key sentence	most important sentence
editorial *(Asahi Shinbun)*	30	0.3	0.41	0.60
tech. journal *(Toshiba Review)*	42	0.24	0.51	0.74

7 CONCLUSION

We have developed an automatic abstract generation system for Japanese expository writings based on rhetorical structure extraction.

The rhetorical structure provides a natural order of importance among sentences in the text, and can be used to determine which sentence should be extracted in the abstract, according to the desired length of the abstract. The rhetorical structure also provides the rhetorical relation between the extracted sentences, and can be used to generate appropriate connectives between them.

Abstract generation based on rhetorical structure extraction has four merits. First, unlike conventional word-frequency-based abstract generation systems(e.g. [Kuhn 58]), the generated abstract is consistent with the original text in that the connectives between sentences in the abstract reflect their relation in the original text. Second, once the rhetorical structure is obtained, various lengths of generated abstracts can be generated easily. This can be done by simply repeating the reduction process until one gets the desired length of abstract. Third, unlike conventional knowledge or script-based abstract generation systems(e.g. [Lehnert 80], [Fum 86]), the rhetorical structure extraction does not need prepared knowledge or scripts related to the original text , and can be used for texts of any domain , so long as they contain enough rhetorical expressions to be expository writings. Fourth, the generated abstract is composed of rhetoricaly consistent units which consist of several sentences and form a rhetorical substructure. so the abstract does not contain fragmentary sentences which cannot be understood alone.

The limitations of the system are mainly due to errors in the rhetorical structure analysis and the sentence-selection-type abstract generation. the evaluation of the accuracy of the rhetorical structure analysis carried out previously([Sumita 92]) showed 74%. Also, to make the length of the abstract shorter, It is necessary to utilize an inner-sentence analysis and to realize a phrase-selection-type abstract generation based on it. The anaphora-resolution and the topic-supplementation must also be realized in the analysis.

The system is now utilized as a text browser for a prototypical interactive document retrieval system.

References

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

[Fum 86] Fum, D. : "Tailoring Importance Evaluation to Reader's Goals: A Contribution to Descriptive Text Summarization", *Proc. of Coling*, pp.252-259, 1986.

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

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

[Kuhn 58] Kuhn, H.P.: "The Automatic Creation of Literature Abstracts", *IBM Journal*, Apr. 1958, pp.159-165.

[Lehnert 80] Lehnert, W. : "Narrative Text Summarization", *Proc. of AAAI*, pp.337-339, 1980.

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

[Mann et al. 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.

[Sumita 92] Sumita, K., et al. : "A Discourse Structure Analyzer for Japanese Text", *Proc. Int. Conf. Fifth Generation Computer Systems 1992 (FGCS'92)*, pp.1133-1140, 1992.