Text Analysis and Knowledge Extraction

Fujio Nishida, Shinobu Takamatsu,
Tadaaki Tani and Hiroji Kusaka

Department of Electrical Engineering,
Faculty of Engineering, University of Osaka Prefecture,
Sakai, Osaka, 591 JAPAN

1. Introduction

The study of text understanding and knowlegde extraction has been actively done by many researchers. The authors also studied a method of structured information extraction from texts without a global text analysis. The method is available for a comparatively short text such as a patent claim clause and an abstract of a technical paper.

This paper describes the outline of a method of knowledge extraction from a longer text which needs a global text analysis. The kinds of texts are expository texts[2] or explanation texts[3]. Expository texts described here mean those which have various hierarchical headings such as a title, a heading of each section and sometimes an abstract. In this definition, most of texts, including technical papers reports and newspapers, are expository. Texts of this kind disclose the main knowledge in a top-down manner and show not only the location of an attribute value in a text but also several key points of the content. This property of expository texts contrasts with that of novels and stories in which an unexpected development of the plot is preferred.

This paper pays attention to such characteristics of expository texts and describes a method of analyzing texts by referring to information contained in the intersentential relations and the headings of texts and then extracting requested knowledge such as a summary from texts in an efficient way.

2. Analysis of intersentential relations

The global sentential analysis is performed by using the information contained in the intersentential relations and the headings of a text by a method combining both the bottom-up and the top-down manner. Various kinds of intersentential relations have been proposed so far by many linguists[1)2)]. By referring to these proposals, intersentential relations are classified tentatively into about 8 items. They are a detail, an additional, a parallel, a rephrase, an example, a temporal succession, a causal and a reasoning relation as described in the following subsections.

(1) Detail relations

If a term t_2 is the topic term in a sentence S_2 and if t_2 is a complementary term of the topic term t_1 in the preceding sentence S_1 as shown in Expr.(1), S_2 is called the detail of S_1.

$$\begin{aligned} S_1&: (\text{PRED: } p_1,\ K_{11}: \underline{\underline{t_1}},\ K_{12}: t_2,\ K_{r1}: t_{r1}) \\ S_2&: (\text{PRED: } p_2,\ K_{21}: \underline{\underline{t_2}},\ K_{r2}: t_{r2}) \\ S_3&: \cdots\cdots \end{aligned} \quad (1)$$

where K:t represents a pair of a case label and a term, and the term with a double underline denotes a topic.

The sentence level of S_1 to that of S_2 depends on the property of the sentence S_3 following to S_2 and the relation among the terms contained in the sentences S_1 S_2 and S_3. If the sentence S_3 is connected to S_1 more closely than S_2, for example, if the sentence S_3 has the topic term t_1 of the sentence S_1 as the topic, it is considered that the principal sentence is S_1 and the sentence level of S_2 is lower than that of S_1.

On the other hand, if S_1 is an introductory sentence of a term t_2 and the articles related to t_2 are described in some sentences following to S_1, or if t_2 is the global topic of the section, the sentence S_2 is considered the principal sentence. The global topic can be easily identified by inspecting the headings of the section the title and the like, whatever it is an attribute name or an attribute value without reading through the whole text.

If the term t_2 in the sentence S_1 belongs to a kind of pronouns such as "in the following ones" or "as follows", the sentence S_2 is set at the same level as that of S_1. At the summarization stage, the system tries to shorten the part consisting of S_1 and S_2 by replacing the pronoun t_2 in S_1 by the main content given in S_2, namely, the main part consisting of t_{r2} and p_2.

[Example 1]

(a) S_1: SGS receives an ordered triple from a user.
S_2: The triple's form is category, input-frames, conditions on the sentence.
S_3: SGS regards the ordered triple as a goal.

S_2 describes the content of a term "ordered triple" in S_1, and S_3 has the topic term "SGS" in S_1. Hence, S_2 is the detail of S_1, and S_1 is the principal sentence.

(b) S_1: In this section, the overview of LFG is described.
S_2: LFG is an extension of context free grammar and has the following two structures.
S_3: One is a c-structure which represents the surface word and phrase configurations, and the other is a f-structure.....

S1 is an introductory sentence of a term "LFG" which is the global topic in a section taken from a text. S_2 has a kind of pronoun "the following two structures" whose contents are described in S_3. Hence, S_2 is the principal sentence and the sentence level of S_2 is the same as that of S_3.

As a special case of detail relations, there are a rephrase relation and an example relation. These intersentential relations between sentences S_1 and S_2 can be identified by referring to their sentential constructions and sentence modifying adverbs such as "in other words" and "for example". The principal sentence of them is, in most cases, the sentence S_1 in an expository text.

(2) Additional relations

If the current sentence has the same sentential topic t_i as that of the preceding sentences and describes another attributes or

functions of the topic, the current sentence is called an additional sentence to the preceding sentences. The sentential form of the relation is

S_1: (PRED:p_1, K_i:$\underline{\underline{t_i}}$, K_{r1}:t_1)
S_2: (PRED:p_2, K_j:$\underline{\underline{t_j}}$, K_{r2}:t_2) (2)

The levels of both the sentences S_1 and S_2 are generally assumed to be the same except for the case that the global topic is put in a predicate part of them. It can be also considered that additional relations hold among various sentential groups of the same level such as chapters sections or paragraphs under a global topic contained in a title.

(3) Other sentential relations

There are other intersentential relations. They are roughly classified into a serial and a concurrent or an extended parallel relation.

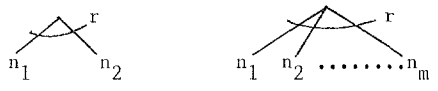
A serial relation such as a temporal succession a causal or a reasoning relation has the same physical location of focus or the same logical object while it has a time shift or a logical inference step shift between adjacent sentential groups.

A concurrent relation has the same time instant of the event occurrences or the same stages of logical inference while it has a distance or a spatial positional shift between the physical or the logical objects described in the adjacent sentential groups.

The level number of a sentence to the adjacent sentential groups in these relations is assigned in a similar way to that of the detail or the additional relation by referring to the intersentential relations and the global topics.

In usual cases, the difference between a principal sentence level and the adjacent sentence level is usually set within one level.

As seen in the above, a sentence or a sentential group has an intersentential relation to some adjacent sentences or sentential groups. The intersentential relation between adjacent sentences is similar to a relation between adjacent words or word groups combined through rewriting rules of a sentence. The intersentential relations are classified into two classes. One of them is a relation such as a detail relation which holds between a principal sentence and the auxiliary or modifying sentential group with a lower level than the principal sentence as shown in Fig.1(a). The other is a juxtaposition relation like an additional relation which holds among several coherent sentences with the same level in usual as shown in Fig.1(b).



(a) (b)

Fig.1 Intersentential relations

In these diagrams a leaf node represents a sentence of a text and an intermediate node denotes a representative sentence of the direct descendents or the principal parts of them. A name r attached to an arc bridging over several branches denotes an intersentential relation.

3. Text analysis

An expository text has a title and consists of several sections. The title shows the main topics of the text. The heading of each section shows local topics of each section and constitutes the attributes of the main topics.

Each of main sections sometimes has an introductory remark followed by the main part. The content of the main part is almost covered with the subframe predetermined by the heading and the title.

The global cohesion of a section is assured by a relation in which each main part of the section shares some items of the same subframe with other main parts.

Based on the above idea of text construction, a text analysis is done after parsing of each sentence. First, each pronoun is replaced by the antecedent noun word with the aids of an anaphora analysis. Then, the intermediate expression of each sentence of the text is transformed into the normal form in which each topic term is inherited together with a double underlined mark. The expressions to be normalized are object-apposition expressions, object-component expressions, predicate-cause expressions, expressions which have a term consisting of a case label, and others[5].

After normalization, the part of topics and the content of each sentence are first identified. Second, intersentential relations between two adjacent sentences are identified indeterministically based on the assumptions of two classes of intersentential relations mentioned in section 2. Third , the main sentence is identified by referring to the intersentential relations and the heading of the section under the main topics of the title. The lower level sentence is indented as a modifier of the main sentence. Sometimes, the knowledge of the specific field is required for better understanding of the relations among main sentential groups and various headings of the text. A case frame of a knowledge base for the specific field is provided in which each slot is filled with the most general term in the specific field. Fourth, a subframe name is prefixed to each main sentential group by referring to the category of the main predicate term of the main sentence and the subframe designated by the heading of the section and the title of the text. The basic subframe names are, for example, FUNCTION, COMPOSITION and PROPERTY in description of actions and physical objects.

As seen in the above, the main work of the text analysis is to identify the main sentential groups and to assign to them a standard attribute name of a subframe in a specified field. These frames and attribute names are used as a key of a specific field for efficiently storing and retrieving the knowledge contained in texts.

The next example of text analysis is taken from a technical paper in language processing.

[Example 2]

Title: A natural language understanding system for data management

Heading of Section: Generating English sentences

Heading of Subsection: The selector

(1)The selector's main job is to construct a graph relevant to the input statement. (2)In constructing this graph the selector first copies the portion of the semantic net which is to be output. (3)It then uses inverse mapping functions to produce a more surface, but still case grammar based representation of the information to be output. (4) Inverse mapping functions map the mumeric representation for date to a more surface one. (5)The selector constructs

modality lists next and chooses a surface ordering rule(SOR) for each verb of the resulting structure. (6)SORs specify the order of the syntactic cases associated to a particular verb to be output.

In the above text the intersentential relations and the levels of sentences are identified, and the label of a subframe is prefixed to each sentence as shown in Fig.2(a) and (b).

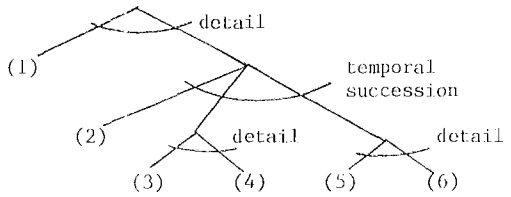


Fig.2(a) The intersentential relations

```
(1)FUNCTION;(PRED:construct,AG:selector,
                          OBJ:graph(.....),
   SUB-PROCESS:
   (2)FUNCTION;(PRED:copy,AG:selector,
               OBJ:portion(....), MANN:first)
   (3)FUNCTION;(PRED:produce,AG:selector,OBJ:more-
               surface...representation(......),
                   MANN:then,INSTR:inverse-
                              mapping-functions
                    (4)(PRED:map,AG:*,OBJ:numeric-
                          representation(...),
                                  ............))
   (5)FUNCTION;(PRED:construct,AG:selector,
                     OBJ:modality-lists,MANN:next)
              (PRED:choose,AG:selector,OBJ:
                 surface-ordering-rule
                (6)(PRED:specify,AG:*,OBJ:
                          order(...)),.....))
```

Fig.2(b) The composition of the text

A symbol "*" denotes a term prefixed to the subframe containing the mark "*" and modified by the subframe.

4. Generation of answering sentences for queries

In this section, sentence generation or text generation for answering a request is described briefly. Text generation is the inverse process of text analysis and is inseparable from text analysis in a sense that the text generation provides an basic idea on text construction for given information to be represented. A given query is parsed and the intermediate expression is constructed. Then the required information is retrieved and transformed into a surface expression in the following steps:

(1) The intermediate expressions related to the main topics of the query are extracted in the order of the level related to the query from the analyzed text or the database storing it under a guide of the frame label and other heading information as well as the index of the terms contained in the text. The level of a description in the text is available for selection of the knowledge source to be extrated.

(2) The intermediate expressions are rearranged in the coherent and readable order, for example, in the occurrence order of the events, and an answer sequence is constructed.

(3) Under a given bounded length the answer sequence is grouped or segmented to several parts and sentential topics are selected to be expanded into surface expressions.

(4) The sentential form of each of the segments is selected to one of phrase, simple, complex and compound surface expressions by referring to the sentential topic.

The summary of the text given in Example 2 is generated from the analyzed results shown in Fig.2(b) by referring to the steps 2 3 and 4. Fig.3 shows two summaries constructed from the descriptions of the text up to level 1 and 3 , where the part enclosed with brackets is the part generated from the descriptions of level 3.

level 1:The selector constructs a graph relevant to the input statement.

level 3:The selector constructs a graph relevant to the input statement. In the construction, the selector performs the following processes. First, the selector copies the portion of the semantic net. Then, it produces a more surface but case grammar based representation with inverse mapping functions [which map a numeric representation to a more surface one]. Finally, it constructs modality lists and chooses a surface ordering rule [which specifies the order of syntactic cases] for each verb.

Fig.3 Generated summaries

5. Conclusion

An experimental system is under construction based on our structured-information extraction system constructed previously. This paper focusses attention on the content suggested by the heading and intersentential structures and assigns a sentence level to each sentence. Ellipsis and restoration problem of known structures on syntax and special field knowledge is not considered here. However, it seems that there are no serious problems in many specific fields at an interactive mode with users.

References

1) Rumelhart,D.E.: Notes on a Schema for Stories, in Bobrow,D.G. and Collins,A. (eds.), Representation and Understanding, pp.211-236, Academic Press, New York (1975).

2) Hobbs,J.R.: Coherence and Interpretation in English Texts, Proc. 5th IJCAI, pp.110-116 (1977).

3) Rigney,J.W. and Munro,A.: On Cognitive Strategies for Processing Text, University of Southern California, Behavioral Technology Laboratories, Tech. Rep. No.80 (1977).

4) Takamatsu,S., Fujita,Y. and Nishida,F.: Normalization of Titles and their Retrieval, Information Processing & Management, Vol.16, pp.155-167 (1980).

5) Nishida,F. and Takamatsu,S.: Structured-Information Extraction from Patent-Claim Sentences, Information Processing & Management, Vol.18, No.1, pp.1-13 (1982).

6) Nishida,F., Takamatsu,S. and Fujita,Y.: Semiautomatic Indexing of Structured Information of Text, J. Chem. Inf. Comput. Sci., Vol.24, No.1, pp.15-20 (1984).