A Method of Automatic Hypertext Construction from an Encyclopedic Dictionary of a Specific Field

Sadao Kurohashi, Makoto Nagao, Satoshi Sato and Masahiko Murakami
Dept. of Electrical Engineering, Kyoto University
Yoshida-honmachi, Sakyo, Kyoto, 606, Japan

1 Introduction

Nowadays, very large volume of texts are created and stored in computer, and as a result the retrieval of texts which fits to a user's demand has become a difficult problem. Hypertext is a typical system to answer this problem, whose primary objective is to establish flexible associative links between relevant text parts and to allow users to select and trace links to see relevant text contents which are connected by links. A difficult problem here is how to construct automatically a network structure in a given set of text data. This paper is concerned with (1) automatic conversion of a plain text set into a hypertext structure, and (2) construction of flexible human interface for the hypertext system. We applied natural language processing methods to locate important conceptual terms in a text corpus and to establish varieties of links between these terms and appropriate text portions.

2 Extraction of thesaurus information

The text corpus we handled as a concrete example was the Encyclopedic Dictionary of Computer Science (hereafter abbreviated as EDCS. Iwanami Publ. 1990. English translation will appear soon from Academic Press). It includes 4500 terms and has the text volume of two million Japanese characters (4 Mega bytes).

The first part of the term description of EDCS is devoted to synonyms, antonyms, abbreviations and broader concept words. This part has typical sentential styles such as,

(i) A is {sometimes, often, also, commonly, ...} {called, written, named, expressed, ...} as B, C, ..., or D.

(ii) A is abbreviated as B.

(iii) A is {an abbreviation, a contraction, ...} of B.

(iv) A stands for B.

(v) We call A B {for short}.

(vi) A is B. A is included in B.

(vii) A is called as B, so C as D.

By finding these sentential patterns the relation between the words A and B is established as follows.

(i) p-link is set up **from a synonym word to a sentence** which defines the synonym relation.

(ii) s-link (by synonym) is set up **from a defined word to defining words** by synonym relation.

Typical sentential styles of intensional definition are:

(i) A is defined as B. A is regarded as B.

(ii) A means B. A connotes B. A is B.

(iii) A is a {kind, form, way, branch, method, ...} of B.

(iv) A is regarded as B, so C as D.

By identifying these patterns in a term description part, the relation between the defined word (A) and the definition sentences is established as:

(i) p-link is set up **from the defined word to the definition sentence** when the defined word is not the headword of the term description. This is the case when the defined word is not so important as a headword of the dictionary, and so a rather simple definition description is embedded in the term description.

(ii) s-link (by synonym) is set up **between the defined words** in the above word extraction process if there are plurals of the defined words.

isa-link is established between a narrower concept word (A) and a broader concept word (B) as A isa B. Here A and B are descriptors which represent other synonym words. isa thesaurus is first established for the words defined by intensional definitions. By utilizing this thesaurus extensional definitions are analyzed to get words of broader/narrower relation defined by these sentences (refer to the next section). Then the thesaurus is reformulated by the addition of new words which have isa relation, and which are obtained from extensional definitions.

3 Extraction of other information

Sentences of extensional definition introduce examples and narrower concept words of the defined words. Typical sentential styles of the extensional definition are:

(i) A is {divided, classified, ...} into B, C, ..., and D.

(ii) A includes B, C, ..., and D.

(iii) There exist B, C, ..., and D {in, for} A.

From these sentences we get the information that A is a defined word, and B, C, ..., and D are narrower (or

Table 1: Functional relations between words.

Relation	Meaning
A isa B	A is a B
A syn B	A is synonym of B
A anti B	A is antonym of B
A hcomp B	A has B as a component
A hprop B	A has a property of B
A hfuni B	A has an intrinsic function of B
A hfuno B	A has an extrinsic function of B
A deal B	A deals with B
A purp B	A has a purpose of B
A used B	A is used in B
A set B	A sets B
A by B	A is done in/by B

lower) concept words. However, there are sometimes the cases where the above sentential styles do not necessarily mean the expected definitions. Therefore to make sure the relations A and B, C, ..., and D are the expected one, we introduced a simple checking procedure as follows.

(i) the semantic category of the narrower concept words must be the same as that of the defined word.

(ii) isa thesaurus is checked to see no contradiction between the defined word and its narrower concept words.

By identifying the above sentential patterns in a term description just after the intensional definition sentences, the relation between the defined word (A) and its narrower term (B, C, ..., D) is established as:

(i) p-link is set up **from a narrower term to its defining sentences** which include the word by the condition that the narrower term is not the headword of the dictionary.

(ii) The extracted words have broader/narrower relation and are included in the isa thesaurus, and are given s-links (by isa).

We tried to extract functional relations between words from the definitional sentences. The extraction is mainly performed by checking the sentential structure and particularly the verb property of the defining sentences, and additional checking is done for the consistency between A and B (Table 1). Links are established from A to B by s-links with meaning attributes shown in Table 1.

EDCS has, besides the term description, the word tree which is similar to a thesaurus, and which shows graphically the structure of the whole area of computer science. Word tree is stored in computer by s-link, and p-link is established from a word in the word tree to the corresponding term description. p-link is also established from an index word to the text portion in which the index word is included. The term description has reference words at the end of the term description. This reference is done by s-link.

4 Retrieval system

Information retrieval can be done from varieties of aspects by tracing p- and s-links. A user can start the dictionary consultation by giving an arbitrary word(W), the meaning of which he or she wants to know.

(i) When W is a headword the term description of W is displayed.

(ii) When W is an index word the term description which includes the explanation of the index word is displayed with the special mark of that explanation part. When there are several terms to W, these candidates are shown with simple explanations, and the user can choose one of them for more detailed description.

(iii) When W is a basic componential word of some compound words these words are shown to the user to select a proper compound word. Then the process (ii) is activated.

(iv) When W is not in the above categories the text search for the whole text data of the dictionary can be started. The full text search will be completed in a few seconds, and all the matched parts can be displayed by KWIC representation.

(v) A user can see on the display the part of the word tree which includes W when it is a headword, and can understand the relative situation of the word W in relation to other words in the wider scope of the computer science field. The user can point any word in the word tree to see its details.

(vi) When the term description of a headword is displayed on the screen, any word on the screen can be marked, and the process of (i) $\sim$ (v) can be activated from the marked word.

(vii) A user can ask what are the words which are related to W by a certain functional relation, and can go to the process (i) $\sim$ (v) with these words.

5 Conclusion

Automatic linking of index words with the corresponding text portions by p-links was compared with the indices in the published book, which were given by human. We can say that a fairly good coincidence is obtained. Some links which are set by automatic process are not proper, but there are many links which were not set up by human by his or her carelessness. The appropriateness of the s-links which are set by automatic process was checked by random sampling, and 92% of the links was recognized as appropriate.

The evaluation result shows that the language processing introduced into the structuring of a large volume of text data of the Encyclopedic Dictionary of Computer Science as a hypertext system was successful. This method is widely applicable for the hypertext construction of varieties of dictionaries.