Proceedings of the IASTED International Conference
Artificial Intelligence and Soft Computing
May 27-30, 1998 - Cancún, Mexico

A METHOD FOR RESOLVING COHESIVE RELATIONS OF UNKNOWN WORDS IN TEXT STRUCTURE

YOSHITAKA SATO
SOFT FRONT, Inc.
N7 W5,KITA-KU,
SAPPORO,060,JAPAN
satoh@softfront.co.jp
(+81-11)700-6011

KENJI ARAKI
Fac. of Engineering,
Hokkai-Gakuen Univ.
S26 W11,CHUO-KU,
SAPPORO,064,JAPAN
araki@eli.hokkai-s-u.ac.jp
(+81-11)841-1161 EXT.860

KOJI TOCHINAI
Fac. of Engineering,
Hokkaido Univ.
N13 W8,KITA-KU,
SAPPORO,060,JAPAN
tochinai@media.eng.hokudai.ac.jp
(+81-11)706-6533

ABSTRACT

This paper describes a method for resolving the problematic relations of unknown words. The purpose of our proposed method is to solve the problem of unknown words, a part of our research about text structure. When a text structure is constructed, a thesaurus is referred to for checking the semantic relations of two words. This means that the relations of unregistered words cannot be processed in our algorithm. To solve this problem, we consider about additional lexical information and the method using this information. These additional information are defined by selecting the word which exists in the thesaurus.

KEY WORDS:

Text Structure, Lexical Cohesion, Thesaurus

1. INTRODUCTION

Recently, the amount of document in computer is increasing explosively. Consequently, efficient text processing systems are required to be developed. In that situation, text structure modeling and analysis has proven to be necessary for natural language understandings[1]. When a text structure is constructed, several knowledge bases are required and one or more strategies is used in its algorithms. On this point of processing, problems exist at the generality of the knowledge base. Every knowledge bases is a subset of whole universal knowledge. This means that knowledge bases do not provide enough information for describing lexical or semantic interpretations. For example, one word which is not registered with any dictionaries cannot be embedded in a text structure as a part of lexical or semantic relationships. This means that the text which has unknown words are not interpreted properly.

This paper describes a method for resolving the problematic relations of unknown words. The purpose of our proposed method is to solve the problem of unknown words, a part of our research about text structure construction. When our text structures are constructed, a thesaurus is referred to for checking the semantic relations of two words. However, one thesaurus does not have all information of the word. This means that the relations of unregistered words cannot be processed in our algorithm. To solve this problem, we consider about additional lexical information and the method using this information.

2. LEXICAL COHESION IN TEXT

In study of text structure, resolving cohesive relations is

one of the major topics. The relations of words could be base elements for text structure construction. Therefore we use lexical cohesion[2] as the first item of the text structure construction. We consider the text structure as aggregation of cohesive bindings.

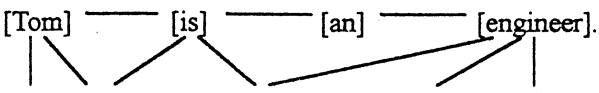
3. OUTLINE OF OUR RESEARCH

3.1 TEXT STRUCTURE

The text structure which we had proposed[3] previously is based upon a lexical chain research[4]. Our text structure is a type of network, which consists of nodes and links.

The nodes are expression form for words in a text. Each node contains information about the surface expression and the grammatical category of a word.

The links are relations between two words. Each link contains a single numeric value, called ``weight''. These weight values are calculated using rules.



[He][works][in][a][factory][which][makes][bicycles].

Nodes: [(word)]
Links: lines between two nodes

Fig. 1: Example of Text Structure

Fig. 1 shows an example of text structure and expression for nodes and links. All two words in a text is basically considered to have a link. However, to make the point of our research clear, we manage subsets of the links. This limitation of the links is based upon the grammatical category of the words. It caused that we can treat the grammar function of words independently.

3.2 TEXT STRUCTURE CONSTRUCTION

(1) CALCULATION OF WEIGHT

Calculation of weight values for links concerns of relations among multiple strategies for expressing lexical cohesion[2,4].

We consider that capability of generalization and robustness of the method for analysis are essential.

(2) CALCULATING FACTOR

The algorithm is calculating the weight value every two words. These weights are calculated by following two elements;

(a) lexical similarities

Lexical similarities is calculated by looking up thesaurus. Thesaurus is database which classify words. Therefore whether two words is in same class of a thesaurus or not indicates the similarity of these two words.

(b) syntax/textual function

Every word in text has its function for the sentence and the text which it is included. For example, subject of a sentence is considered as semantic focus of the sentence, and it can be focus of next sentence. To analyze these function of words, we check only the position of the word in the sentence and the text. This means we use no parser, which is another problematic component of natural language processing system.

(3) COMBINING FACTOR

Each values calculated as described above is the element in a vector. The vector is the expression for a weight of the links.

(a) Lexical similarity ($f_1(w_m,w_n)$)

(b) Distance in text ($f_2(w_m,w_n)$)

$$x(w_m,w_n) = a_1 \cdot f_1 + a_2 \cdot f_2$$

where a_1, a_2 is parameters.

4. PROBLEMS OF UNKNOWN WORDS

There are some problems in our previous research. One of those is unregistered words.

As described above, the semantic relations are obtained from searching one thesaurus. However, no one thesaurus exists that has the whole words. All of thesauri and dictionaries are subsets of perfect set of knowledge base. Therefore, there exist obviously some limitations in data resources. This means that some lexical relationship cannot be treated by using the thesaurus. This limitation is not depends on the number of thesauri which is used.

For example, we show a simple text.

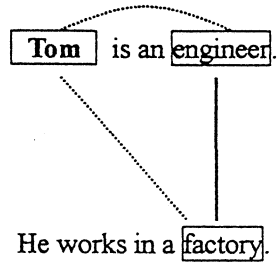


Fig.2: Example of Unknown Words in Text

In this text, "Tom" is a proper noun, and does not exist in Roget's thesaurus, which we use as reference. Therefore we could not manage the relations of "Tom" and "engineer, or "Tom" and "factory".

5. RESOLVING UNKNOWN RELATIONS

5.1 ADDITIONAL INFORMATION FOR UNREGISTERED WORDS

The purpose of our proposed method is solving the problem described above. Firstly, we need to clarify the minimum requirements for data which is needed to do the estimation.

We consider that more lexical information is required. In Fig.2, the word "Tom" can be recognized as "Noun" at our previous method[3]. We add more information to it, such as "Noun-[human]". The additional information is defined as one or more words, and selected from the words registered in the thesaurus.

5.2 CALCULATION OF WEIGHT VALUE FOR LINKS OF UNREGISTERED WORDS

Calculation of weight value is done by using of the additional information. In the case of the reference, the additional information is used instead of the word itself. Therefore, we can process the text structure construction as the same algorithm as our previous study.

For example, in Fig.2, the relation between "[human]" and "engineer" is determined by the system.

6. EXAMPLES

The text for experiment[5] is book for exercises in English reading, which is published for Japanese high school students.

Fig. 3 shows an example text in process.

This text has four proper nouns, "Tom", "London", "Susan" and "Bishopton". None of these words exists in Roget's thesaurus. Therefore these words have to be reworded for obtain the information from the thesaurus. We append generic words to each of the unregistered words.

Fig.4 shows the additional information for unknown words in the example text.

1: **Tom** is an engineer.
[human]

2: He works in a factory which makes bicycles.

3: He has been at this factory for a year: before he came to the factory, he was studying to be an engineer at the University of **London**.
[city]

4: He does his work very well, and some day he is going to be the manager of a big factory — at least, he hope so, and his girl friend, **Susan**, hopes so too.
[human]

5: The factory is a long, low building between the road and the railway about five miles from **Bishopton**.
[town]

6: It was built about ten years ago.

7: There are a lot of people working in the factory, and many of them live in the near towns and villages and travel to the factory every day.

8: Some of them are brought to the factory each morning to buses, and are taken home again in the evening it is always very noisy in the factory, but the workers soon get used to the noise.

Fig. 3: Example Text

Tom → [human]

London → [city]

Susan → [human]

Bishopton → [town]

Fig.4: Additional Information for Unknown Words

7. CONCLUSIONS

In this paper, we described about a method for resolving the relation of unknown words. We consider that auxiliary information is required for processing unregistered words, and the amount of additional information should be minimized. Our research is in progress, and we need to evaluate the proposed method by experiments.

The future direction of our study is application of the algorithm for summarizer or anaphora resolution.

REFERENCES

[1] Zadrozny, W. and Jensen, K., Semantics of Paragraphs, *Computational Linguistics*, 17(2), 1991, 171-209.

[2] Halliday, M. A. K. and Hasan, R., *Cohesion in English* (London and New York: Longman, 1976).

[3] Sato, Y., Araki, K., Miyanaga, Y. and Tochinai, K., A Calculation Method for Cohesive Links in Text, *Proc. of Artificial Intelligence and Soft Computing*, 1997, 236-239.

[4] Morris, J. and Hirst, G., Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text, *Computational Linguistics*, 17(1), 1991, 21-48.

[5] Ito, K., *Eibun Wayaku Enshuu* (Tokyo: Sundai Bunko, 1989).