Proceedings of the IASTED International Conference ARTIFICIAL
INTELLIGENCE AND SOFT COMPUTING, held July 27-31, 1997,
Banff, Canada

A CALCULATION METHOD FOR COHESIVE LINKS IN TEXT

YOSHITAKA SATO
VISION Corporation
Marumasu Bldg. 18,
N 7-Jo, W 1-Chome, Kita-ku
Sapporo, 060 Japan
satoh@visionj.co.jp

KENJI ARAKI
Dept. of Elec. and Info.
Hokkai-Gakuen University
S 26-Jo, W 11-Chome, Chuo-ku
Sapporo, 064 Japan
araki@eli.hokkai-s-u.ac.jp

Y. MIYANAGA KOJI TOCHINAI
Div. of Elec and Info.
Hokkaido University
N 13-Jo, W 8-Chome, Kita-ku
Sapporo, 060 Japan
{miyanaga,tochinai}@hudk.hokudai.ac.jp

ABSTRACT

This paper describes about a text structure construction method. The text structure which we propose is a sort of network. Nodes of the network are words in the text, and links express relation between two words. Each link contain a numeric value, called weight. The weight values is calculated from lexical relationship of the words. Main process of the text structure construction is culculation of these weight values. Our major topic is adapting the weight for the applications. Especially, expressing local and grobal cohesive bindings as single numeric value is the focus of this paper.

KEY WORDS: text structure, lexical cohesion, thesaurus

1 INTRODUCTION

Recently, the amount of document in computer is increasing explosively. It causes that effective text processing is needed. In that situation, text structure modeling and analysis has proven to be necesarry for natural language understandings[1].

When it should be construct with bottom-up approaches, some knowledge bases are required and one or more strategies is used in its algorithms. On this point of processing, problems are generality of knowledge bases and the complexities on the relationship of data and strategies.

Every knowledge bases is a subset of whole universe. This means that knowledge base does not provide enough information for describe lexical or semantic interpretations. Therefore, using more than two different knowledge bases is better for robustness and generality.

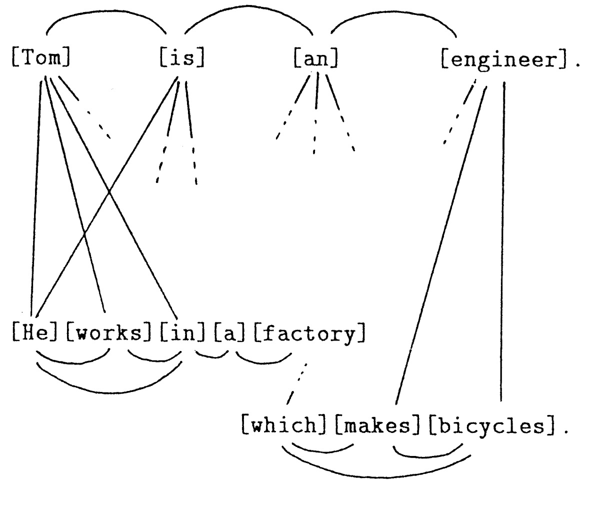
On the other hand, in multi-strategies approach such as in [2], there are no obvious proof for the order of applying each strategy. There are some difficult problem on managing confliction among the rules.

Therefore, our approach described in this paper attempts to apply the rules simultaneously.

2 TEXT STRUCTURE

The text structure we propose in this paper is based upon the network which consists of nodes and links. The nodes are expression form for words in a text, because we consider that word is minimum unit for text processing. Each node contains information about surface expression and grammatical category of a word.

The links are relations between two words. Each link contains single numeric value, called "weight". These weight values are calculated by multiple strategies, which is descrived in next chapter.



Node: [(word)]
Link: line 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 CALCULATION OF WEIGHT

Calculation of weight values for links concerns of relations among multiple strategies for expressing lexical cohesion[3, 4].

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

3.1 CALCULATING FACTOR

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

1. 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.

2. syntax/textual function

   Every word in text has its function for the sentence and the text which it is in. For example, subject of a sentence is considered as semantic focus of the sentense, and it can be focus of next sentence. To analize these function of words, we check only the position of the word in the sentence and the text. It means we use no parser.

3.1.1 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.

1. lexical simirality($f_1(w_m, w_n)$)
2. distanse in text($f_2(w_m, w_n)$)

$$x(w_m, w_n) = \sum_{i=1}^{k} a_i f_i$$

where a_i is parameters, k is number of factor.

4 EXPERIMENT

The purpose of experiment is to confirm the assumption descrived above. Especially, we are interested in adapive weight in cohesive links. Therefore, the experiment needs to done iteratively.

4.1 STANDARDS FOR EXPERIMENT

The experiment is done in following conditions:

1. Grammatical category in focus

   All of the words in text is fundamentaly applied for node. However, not all of the words have lexical or semantic contributions for the text which it is in. Obviously, for example, conjunctions has almost no semantic meanings for whole text.

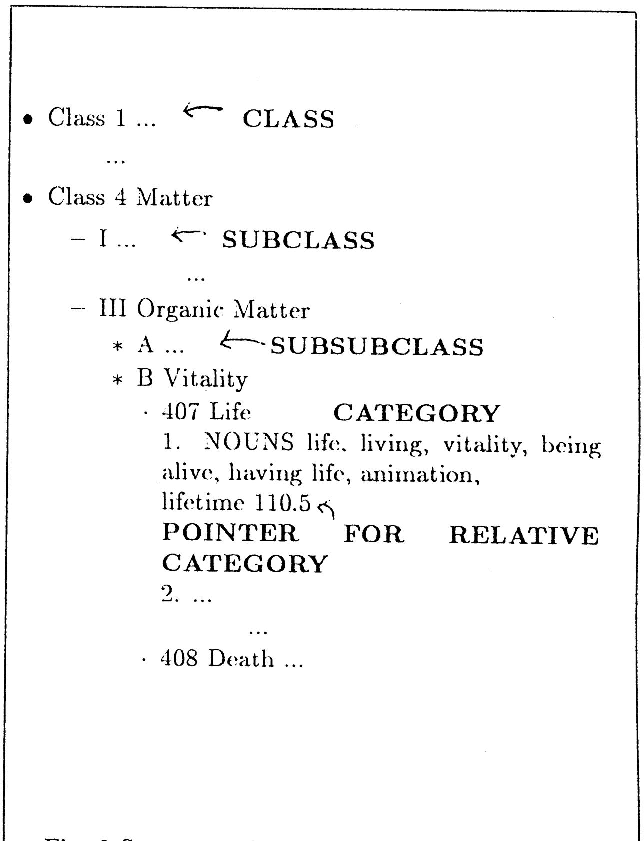
   We select nouns as the grammatical category for the experiment in this paper. Limitation of grammatical category is for making the point simple.

2. Database for calculation of lexical similarity

   For calculate the value of links as the weight of cohesive bindings, it is necessary that the semantic similarity is numeralized.

   The calculation of lexical simirality is based on the method in study of lexilal chain by Morris et allexical. We perform the calculation with using one of their algorithms, which is checking the identification of the word in same category of the thesaurus.

   We also use Roget's thesaurus as knowledge database. Fig. 2 shows the structure of Roget's thesaurus.



- Class 1 ... ← CLASS
  ...
- Class 4 Matter
  - I ... ← SUBCLASS
    ...
  - III Organic Matter
    * A ... ← SUBSUBCLASS
    * B Vitality
      · 407 Life CATEGORY
      1. NOUNS life, living, vitality, being alive, having life, animation, lifetime 110.5 ← POINTER FOR RELATIVE CATEGORY
      2. ...
      ...
      · 408 Death ...

Fig. 2 Structure of Roget's Thesaurus

4.2 EXAMPLES

Fig. 3 shows an example text in process. The text for experiment[5] is book for exercise in English reading for Japanese high school students.

1: Tom is an engineer.

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.

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.

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

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

4.3 DISCUSSION

There are two problems processing this text.

First, it is difficult to separate between general noun and specific noun. For example, the word "factory" in sentence No.3 is specific noun, and same word in sentence No.4 is general noun. This disability of distinguish is relative issue with resolving noun phrases. However, one of our subject is to use simple algorithm, so we consider this problem as a future work.

Next, in the example text structure, some syntax information are not expressed. For example, the words "morning" and "evening" in sentence No.8 are respectively in the clause which binded by coodinate conjunction "and." Our algorithm cannot treat the syntax relation of these two words.

5 CONCLUSION

In this paper, we describe about a text structure and a method for application that represents lexical cohesion in the text. This method is based upon the work about "lexical chain", which expresses lexical similarity of words. We try to add it to another information about lexical similarities and grammatical functions.

However, we consider that we need to confirm the effectiveness of our method.

The limitation of our method in this paper and the future directions of this work is as follows.

- Clarifying the characterictics of the weight of cohesive links

  In this paper, the links have only one numeric value as expression form for the power of lexical bindings. It causes that application of our algorithm is simple. Because it only needed for confirming the degree of lexical bindings that we look up the weight value and compare it the other one. However, we admit that there is difficult problem in calculating the weight value. Our weight calculation method combines more than two values calculated independently with linear combination. In this research, we do not provide any proof for the combination algorithm. This is one of the limitations of our work. To solve this problem, we consider about another experiment which clarify the characterictics of the weight of cohesive links.

- Experiment with another thesaurus and large corpus

  The basis of calculation for weight value of link is thesaurus as database. The numeric value is calculated with looking up the word in the thesaurus. This means that the weight value depends on the structure of the thesaurus in use.

  In this paper, we use Roget's thesaurus for the experiment. Some of the limitations of using it is pointed out by Morris et al[3].

To solve these problem, we consider using another thesaurus, and modify it to weight calculation method.

REFERENCES

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

[2] Carbonell, J. G. and Brown, R. D.,
Anaphora Resolution: A Multi-Strategy Approach,
Proc. of COLING '88, 1988, 96-101.

[3] 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.

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

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