AUTOMATIC TEXT ANALYSIS USING RHETORICAL STRUCTURE THEORY WITH APPLICATION FOR INFORMATION SEARCH AND RETRIEVAL

WONG SHUK MAN, CECILIA

DOCTOR OF PHILOSOPHY
CITY UNIVERSITY OF HONG KONG
APRIL 2004
AUTOMATIC TEXT ANALYSIS USING RHETORICAL STRUCTURE THEORY WITH APPLICATION FOR INFORMATION SEARCH AND RETRIEVAL

WONG SHUK MAN

CECILIA

PhD

2004

CityU
CITY UNIVERSITY OF HONG KONG
香港城市大學

Automatic Text Analysis using Rhetorical Structure Theory with application for information search and retrieval
基於修辭結構理論的
自動文章分析及訊息搜索

Submitted to
Department of Chinese, Translation, and Linguistics
中文翻譯及語言學系
in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy
哲學博士學位

by

Wong Shuk Man, Cecilia
王淑雯

April 2004
二零零四年四月
Abstract

In addition to the typical methods employed in information retrieval systems, e.g. calculating frequency of keywords, pattern matching involving keywords, in this research project, I am proposing an approach to information search and retrieval based not only on the basic element set known as the Dublin Core Metadata Element Set (DCMES), which represents the content or bibliographical information of the data, but also based on the identification of linguistic information about the rhetorical structure of the text. This rhetorical structure information may be inferred from linguistic clues identified in the text. Both types of information are encoded as rules and facts in F-Logic (Frame-Logic). The cues and criteria in identifying rhetorical structure information are based on those developed by Corston-Oliver(1998).

The text base in question consists of abstracts of linguistics journal articles drawn from a collection of over three hundred papers on the topic of Chinese Linguistics. Included in the text base are abstracts from linguistics journals in both Chinese and English. Information retrieval is web-based. Besides offering a search and retrieval capability, the application can also be extended by developing a web interface for authors or publishers to submit their abstracts to the text base.

As the data in this research is linguistic abstracts, part of the focus of the research would be the investigation and analysis on the text structure of the abstracts. Since the usual way of creating an abstract is to extract all the main ideas of the text being described, analyzing abstracts in terms of their structure will be helpful in determining the structure of
the whole article upon which each abstract is based. By identifying the relations among the
different spans in the abstracts, one can be able to realize the general structure of the whole
article. In other words, investigating and analyzing the text structure of discourse in the
smaller-scale, i.e. the abstracts makes it possible to gain insight into those of larger-scale
discourse, i.e. the papers.

The research serves to further the development of ‘smart’ search facilities through the
use of linguistic knowledge about the text. We have based our approach on the existence of
a correlation between the move and rhetorical structures of texts. The result of the research
has demonstrated support for the validity of this assumption.
I would like to express my heartfelt gratitude to my supervisor, Dr. Jonathan James Webster for his unfailing support and encouragement for my study throughout the years. Through his supervision, I am guided to learn a lot in developing my research. His patience and effort are highly appreciated.

I would also like to thank Dr. Pan Haihua and Mr. Lai B. Y. Tom for being members of my qualifying panel and to thank the panel of examiners for sparing their precious time to help. Thank you so much!

Moreover, I would like to take this chance to thank one of the greatest teachers I have ever met in my life, who is Dr. Mary Erbaugh. Although she has left City University of Hong Kong, she is and will be one of the best teachers of mine. Her words of wisdom really help to heal my wounds in times that I encounter difficulties and unhappiness. I hope that she will get well soon and have good health and happy life.
# Table of Content

<table>
<thead>
<tr>
<th>Section</th>
<th>Page Number(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.1. Aims and Objectives</td>
<td>1</td>
</tr>
<tr>
<td>2. Literature Review</td>
<td>4</td>
</tr>
<tr>
<td>2.1. Introduction to abstract genre</td>
<td>4</td>
</tr>
<tr>
<td>2.1.1. What are abstracts?</td>
<td>4</td>
</tr>
<tr>
<td>2.1.2. Abstract as a genre</td>
<td>5</td>
</tr>
<tr>
<td>2.2. Introduction to a move</td>
<td>5</td>
</tr>
<tr>
<td>2.3. Moves in abstract genre</td>
<td>6</td>
</tr>
<tr>
<td>2.4. Adoption of moves and Rhetorical Structure Theory (RST)</td>
<td>7</td>
</tr>
<tr>
<td>in this research</td>
<td></td>
</tr>
<tr>
<td>2.5. Introduction to Rhetorical Structure Theory (RST)</td>
<td>8</td>
</tr>
<tr>
<td>2.5.1. Schemas</td>
<td>9</td>
</tr>
<tr>
<td>2.5.2. Relations defined</td>
<td>11</td>
</tr>
<tr>
<td>2.5.3. Relation Definitions</td>
<td>12</td>
</tr>
<tr>
<td>2.5.4. Nuclearity</td>
<td>14</td>
</tr>
<tr>
<td>2.5.5. Previous research in RST</td>
<td>15</td>
</tr>
<tr>
<td>2.6. Comparison between RST and Systemic Functional Grammar (Halliday)</td>
<td>20</td>
</tr>
<tr>
<td>3. Methodology</td>
<td>23</td>
</tr>
<tr>
<td>3.1. Technology infrastructure of the research</td>
<td>23</td>
</tr>
<tr>
<td>3.1.1. Representing Bibliographical information using F-Logic</td>
<td>29</td>
</tr>
<tr>
<td>3.1.2. Representing Textual/Linguistic information using F-Logic</td>
<td>30</td>
</tr>
</tbody>
</table>
3.2. Approach adopted in this research in generating RST structure

3.2.1. Brief introduction to the Corston-Oliver’s approach in generating RST structure

3.2.2. Employment of the Corston-Oliver’s approach in this research

3.2.3. Procedure of the process for generating RST structure in this research

4. Findings/Result

4.1. Linguistics exploration through inferencing

4.1.1. Bibliographical inferences

4.1.2. Textual inferences

4.2. Found relations between move and rhetorical structures

4.3 Accessing information through Ontobroker

4.4 Query interface for retrieval

4.5 Improvement on precision and recall of information retrieval by textual information/moves and rhetorical information

5. Conclusion

List of References

Appendix I: List of the definitions of the defined RST relations

Appendix II: Corston-Oliver’s heuristic scores

Appendix IIIa: Initial Computer constructed rhetorical structure for sample abstracts 1-3

Appendix IIIb: Computer constructed rhetorical structure for sample abstracts 1-3

Appendix IV: Manually analyzed rhetorical structure for abstracts for sample abstracts 1-3

Appendix V: Correlations between moves and rhetorical structure represented in RST schema
Appendix VI: F-Logic facts of abstract number 10  

Appendix VII: Interface for linguistic features data submission