Human Judgment as a Basis for Evaluation of Discourse-Connective-based Full-text Abstraction in Chinese

Benjamin K T'sou, Hing-Lung Lin, Tom B Y Lai
Language Information Sciences Research Centre
City University of Hong Kong
83 Tat Chee Avenue, Kowloon Tong
Kowloon, Hong Kong
rlbtsou@cpccux0.cityu.edu.hk

Abstract

In Chinese text, discourse connectives constitute a major linguistic device available for a writer to explicitly indicate the structure of a discourse. This set of discourse connectives, consisting of a few hundred entries in modern Chinese, is relatively stable and domain independent. This paper attempts to demonstrate the validity of using discourse connectives in full-text abstraction by means of an evaluation method, which compares human efforts in text abstraction with the performance of an experimental system called ACFAS. Specifically, our concern is about the relationship between the perceived importance of each individual sentence as judged by human beings and the sentences containing discourse connectives within an argumentative discourse.

1. Introduction

Through increasingly convergent interests and cross-fertilization in linguistics and computer science, research into discourse in natural language processing (NLP) has made much progress in the last decade. Discourse as understood by linguists refers to any form of language-based purposeful communication involving multiple sentences or utterances. The most important forms of discourse of interest to NLP are text and dialogue. While textual discourse normally appears as a linear sequence of sentences, it has long been recognized by linguists that these sentences tend to cluster together into units, called discourse segments, that are related in some way to form a hierarchical structure.

In NLP, discourse analysis must go beyond sentence-based syntactic and semantic analysis. Its functions are to divide a text into discourse segments and to recognize and re-construct the discourse structure of the text as intended by its author [Allen 1995]. Results of discourse analysis can be used to resolve many important NLP problems such as anaphoric reference [Hirst 1981], tense and aspect analysis [Hwang 1992], intention recognition [Grosz 1986, Litman 1990] and text generation [McKeown 1985, Lin 1991], etc.

Discourse analysis is also applicable to text abstraction, as demonstrated in Project ACFAS (Automated Chinese Full-text Abstraction System), which is an on-going, computational linguistics research project at the City University of Hong Kong. ACFAS aims to automatically produce abstracts from Chinese newspaper editorials in Hong Kong [T'sou 1992, T'sou 1996] through a new approach based on analyzing the rhetorical structure of argumentative discourse. This process, called Rhetorical Structure Analysis (RSA) [T'sou 1996], is based on the Rhetorical Structure Theory developed by Mann and Thompson for describing the discourse structure of English text [Mann 1986]. A similar approach has been made for Japanese [Ono 1994].

As a brief review of the RSA, please note that in an argumentative discourse, the progression of reasoning commonly involves *explicit* discourse connectives, that are used to express the temporal, causal or rhetorical relationships amongst constituent propositions or clauses. RSA makes use of those discourse connectives appearing in a Chinese text to (1) extract every rhetorically connected discourse segment of the text, and (2) recognize and construct the rhetorical structure of each discourse segment. Using these resultant rhetorical structures, an appropriate abstract may be generated by systematic rhetorical structure reduction to produce abstracts with differential coverage of the details of the underlying argumentation [T'sou 1996].

In modern Chinese text, discourse connectives constitute a major linguistic device available to a writer to explicitly indicate the structure of a discourse. Examples of Chinese discourse connectives include 因此("therefore"), 因爲("because"), 如果("if")...就("then"), 假如("assuming")...那末("then"), 雖然("although")...但是("but"), etc. This set of discourse connectives, consisting of a few hundred entries, is relatively stable in modern Chinese, and is independent of the domain of discourse.

Initial corpus analysis [Ho 1993] has indicated that about 30% of sentences in Chinese editorials contain discourse connectives, which provide a key to the basic understanding of the inherent logical structure within the argumentative discourse. They also provide a potentially useful approach for scaleable and domain-independent full-text abstraction as demonstrated in [T'sou 1996]. Because the flow of argumentation is not exclusively demarcated by discourse connectives, the validity and robustness of this approach require empirical comparison with human efforts in abstraction, which can contribute to the design of a general evaluation method for automatic abstraction in Chinese. Such a comparison would entail human subjects performing abstraction on the same editorials as ACFAS and comparing their results (see also [Watanabe 1996]). Two major questions require answers from carefully designed experiments: (1) Is there relative consistency in human abstraction? (2) Is the existence of discourse connectives a relevant factor in determining the relative importance of the constituent discourse segments?

2. Design of Experiment

A set of 10 Chinese editorials was taken from two well-known newspapers published in Hong Kong and denoted as {E1, E2, …., E10}. These editorials were concerned with controversial events which occurred in Hong Kong. They included the decision to build a nuclear power plant near Hong Kong, the relationship between debt and corruption in the police force,. the unemployment rate of young people, the law and the attitude of the population towards anti-discrimination, etc. These editorials are arche-typical examples of argumentative discourse.

The subjects of experiment included three groups of 25 students each from three prestigious universities in northern China. Two groups were from Chinese departments and one from a computer science department, and all were either final year undergraduates or first year graduate students. They undertook the experiment separately in time and location, and, as far as we can ascertain, these were independent experiments. The subjects were generally brought up in primarily monolingual settings and could understand the issues discussed in the selected editorials, but, without the intimate knowledge, as well as prejudice, of the related background. It was our conscious decision to use Hong Kong newspaper editorials with Mainland Chinese subjects of above-average linguistic competence and intellectual capacity for performance comparison.

Computer print-out instead of the original texts were given to the subjects of this experiment to avoid any confusion and hints preserved in the format of the original texts. The experiments took place under controlled environment in an invigilated classroom setting.

Subjects were given the 10 selected editorials in one batch. They were asked to determine which clauses or sentences in each given editorial contained the most essential information from the author. Subjects were required to work on the editorials sequentially and in prescribed time. Each subject was asked to (1) underline in red about 10% of text which, according to his/her own judgment, contained the most important information (called *key propositions* below) in the editorial, and (2) underline in blue about 15% more of the next most important parts (called *important propositions* below) of the editorial. Subjects were specifically advised to cover as widely as possible (subject to the above constraints, of course) all aspects of the content that the author might have intended to convey.

3. Method of Analysis and Evaluation Metrics

Data analysis of the experimental results as well as performance evaluation of ACFAS were carried out as follows: (1) Target abstracts were generated per editorial per student group according to how the editorial text was marked by the human subjects. (2) Target abstracts for the same editorial were analyzed for similarity and consistency among the three groups. (3) Abstracts generated by ACFAS were compared with the corresponding abstracts generated by the human subjects according to two performance metrics, recall and precision, as defined in Section 3.2.

3.1 Generation of the target abstract

The objective of this step is to select part of a given source text to form a target abstract. The selection criterion is based on how the text is marked by the human subjects of experiment.

(i) Let WK be the weighting factor assigned to a *key proposition* and

WI be the weighting factor assigned to an *important proposition*,

where $0 < \text{WK}, \text{WI} \leq 1$.

We can compute the weighted average of the j^{th} proposition, denoted as PERC-IMP$_j$ (for *Perceived Importance*), according to the following formula:

$$\text{PERC-IMP}_j = \frac{1}{n} \{ (\sum_{i=1}^{n} \text{KEY}_{ij})*\text{WK}+(\sum_{i=1}^{n} \text{IMP}_{ij})*\text{WI} \}$$

where n is the number of subjects,

$$\text{KEY}_{ij} = \begin{cases} 1 & \text{if the } j^{\text{th}} \text{ proposition is marked by the } i^{\text{th}} \text{ subject as a key proposition,} \\ 0 & \text{otherwise} \end{cases}$$

$$\text{IMP}_{ij} = \begin{cases} 1 & \text{if the } j^{\text{th}} \text{ proposition is marked by the } i^{\text{th}} \text{ subject as an important proposition,} \\ 0 & \text{otherwise} \end{cases}$$

(ii) For a given source text, we can sort all propositions of the text according to their perceived importance.

Let α ($0 < \alpha \leq 1$) be the threshold value used to separate those propositions that should be included in the *target abstract* (for PERC-IMP$_j$ $\geq \alpha$) and those that should be excluded (for PERC-IMP$_j$ $< \alpha$). Note that α is introduced to account for the fact that, when we talk about abstraction of a source text, there is a whole spectrum of possible abstracts with different sizes, each corresponds to a different value of α.

For a given α, we can define the *abstract ratio*, β, of the target abstract to be

$$\beta(\alpha) = \frac{\text{size of target abstract}(\alpha)}{\text{size of source text}}$$

3.2 Performance metrics for a text abstraction system

ACFAS is an experimental text abstraction system that is capable of generating multiple abstracts with differential coverage of a source text [9]. We consider only the abstract generated by the top-level output of ACFAS. We define the abstract to source ratio of the top-level output of ACFAS to be

$$\text{ACFAS-RATIO} = \frac{\text{size of top-level abstract of ACFAS}}{\text{size of source text}}$$

The following two performance measures for ACFAS are defined:

$$\text{RECALL}(\beta) = \frac{\text{\# of target propositions generated by ACFAS}}{\text{size of target abstract}}$$

$$\text{PRECISION}(\beta) = \frac{\text{\# of target propositions generated by ACFAS}}{\text{size of abstract generated by ACFAS}}$$

Note that in the above definitions, we explicitly indicate that both RECALL and PRECISION depend on the abstract ratio β of the target abstract that we choose to conduct an evaluation.

4. Similarity Analysis of Human-Generated Abstracts

In this section, results of the experiment described in Section 2 above are analyzed within the framework set out in Section 3 to examine consistency in abstracts generated by different groups of human subjects.

Text abstraction is the process of condensing salient information from a source text. It involves sophisticated and intelligent manipulation of given and assumed world knowledge as well as knowledge of natural language. It is well known that abstracts produced by different human individuals for the same source text can vary depending on the background and education level of the individuals involved. Furthermore, even for the same individual, different abstracts can be generated at different times [Luhn 1958]. While this is true with respect to the behavior of individual human beings, when they are examined as a group, our results below show that abstracts produced by different groups of human subjects with similar educational background are in fact relatively consistent.

Fig. 1 shows the average Perceived Importance scores for the 65 propositions in one of the test editorials in respect of each group of subjects. The two weighting factors are set to be WI=0.8 and WK=1. These two values are chosen to reflect the fact that key propositions and important propositions constitute top 10% and the next 15%, respectively, of the source text according to the instruction given to the subjects of experiment.

Inspection of the three plots of Fig. 1 reveals that while there is a considerable variation in the (three) absolute scores of each of the individual propositions, the overall shapes of the three plots are obviously similar.

The similarity of the plots is statistically assessed by considering each of the propositions as an observation point. For the sake of convenience, the scores given by the 25 subjects in a group are averaged, so that there are 3 scores for each of the observation points. Pearson coefficients of correlation (pair-wise) of the (averaged) scores of the three groups calculated from data for 379 propositions in 5 common test editorials are given below.

	Group 1	Group 2	Group 3
Group 1	1		
Group 2	0.886077	1	
Group 3	0.914838	0.945098	1

As shown above, the correlation coefficients are positive and close to 1. They clearly establish strong consistency amongst the three groups of human subjects with respect to the perception of relative importance of individual propositions in the editorials. Besides confirming that human subjects do indeed generate abstracts in a consistent manner, the above analysis can also be seen as empirical evidence of the validity of the Perceived Importance score suggested in Section 3.

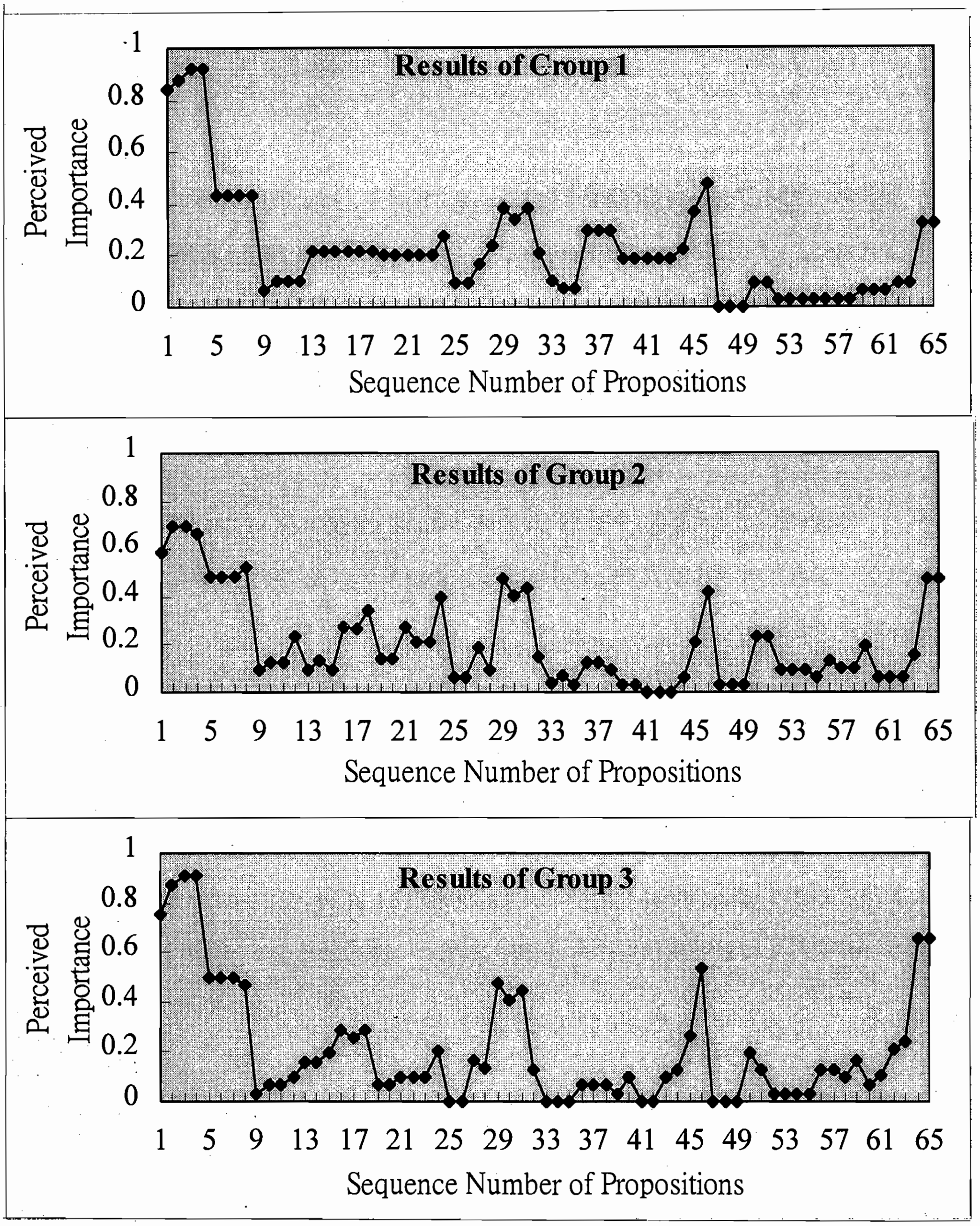

Figure 1 Perceived Importance of an Editorial for Three Groups of Subjects

5. Performance Evaluation of ACFAS: An Empirical Study

In the previous section, we have demonstrated that abstracts generated by different groups of human subjects exhibit a high degree of similarity. Therefore, it seems appropriate to evaluate the performance of a text abstraction system by comparing its output with target abstracts produced by human subjects based on the metric of Perceived Importance. In this section, we report on an empirical study of the performance of ACFAS based on the performance measures of RECALL and PRECISION defined in Section 3. This evaluation was conducted by comparing abstracts generated by ACFAS with those target abstracts produced by the group of 25 computer science students.

5.1 Statistics on the target abstracts of 10 source texts

The average target abstract ratio's of 10 editorials, given as a function of the Perceived Importance threshold, are shown in Figure 2. The two weighting factors are set to be WI=0.8 and WK=1 as discussed above. On the average, only 12.5% of the contents of any source text has received a Perceived Importance of 0.5 or above. This indicates that, within any text, there exists a small, identifiable group of propositions which contains the most important information relevant to the text. This small group of propositions will form the basis of any abstract produced by human subjects.

On the other hand, it may be noted that about 40% of the content of any source text has received a Perceived Importance of less than 0.1. This very likely indicates a high degree of redundancy in human compositions of this genre.

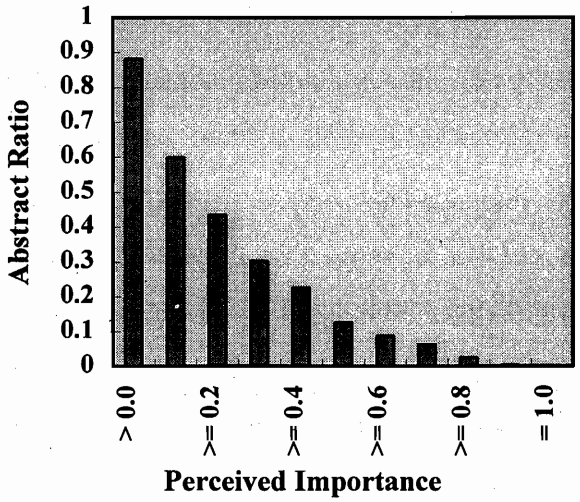

Figure 2 Abstract Ratio as Function of Perceived Importance

5.2 Statistics on the top-level abstract of ACFAS

On the average, the size of a top-level abstract generated by ACFAS is 27.4% of the source text. This is significantly higher than the target abstract ratio of 12.5% (for $\alpha \geq 0.5$) produced by human subjects. This result may be caused by the lack of explicit discourse connectives to determine the relationships between different (yet related) discourse segments. An in depth study of more general types of discourse connectives, including explicit and implicit ones, should improve the present situation.

5.3 Performance evaluation of ACFAS

The average RECALL and PRECISION of the 10 abstracts generated by ACFAS according to how well they correspond with the target abstracts produced by human subjects are shown in Fig. 3 and 4.

As shown in Fig. 3, when the abstract ratio (i.e. the human-generated abstract size as a percentage of the source text) equals to 100%, the average RECALL is 27.4%, which is also the size of the top-level abstract generated by ACFAS. As the value of the abstract ratio reduces, the average RECALL increases modestly until it reaches a maximum value of 36.5% for the abstract ratio of 30%. This improvement of about 10% for the average RECALL is an indication of an inherent relationship between the mechanism of ACFAS and the process of human text abstraction.

Note that when the abstract ratio of 30% further reduces, the average RECALL decreases rapidly. As our abstract ratio is computed by sorting all propositions of the text according to their perceived importance, a small abstract ratio corresponds to the set of propositions that have received high average scores of perceived importance. This result indicates that ACFAS is unable to retrieve some of the most important propositions from the text. After examining the content of the source texts, we find that there is a high probability of finding important propositions in the beginning and the end of these texts (this seems to reflect a typical pattern in argumentative discourse, i.e. problem statement in the beginning and conclusion in the end of a text), but there are relatively few discourse connectives found in this area.. The present strategy of ACFAS is to ignore sentences without explicit discourse connectives between them, therefore, those target propositions located in the beginning and the end of the text will not be included in the ACFAS-generated abstract.

Fig. 4 contrasts the values of RECALL and PRECISION, both as functions of the abstract ratio. We observe that at the maximum RECALL of 36.5%, the average PRECISION is 39.4%. In other words, about 60% of the target propositions are not extracted by ACFAS, and most of them are propositions located in the beginning and end of the source texts.

The conclusion we can draw from this result is that a system like ACFAS, which uses only the existence of explicit discourse connectives to determine the relative importance of the propositions in an argumentative discourse, performs well in the part of text that deals with the argumentative flow and presentation of evidence, but performs poorly where the problem statement is delineated and the conclusion or summarization is presented. Other factors and cues must be used to account for this deficiency.

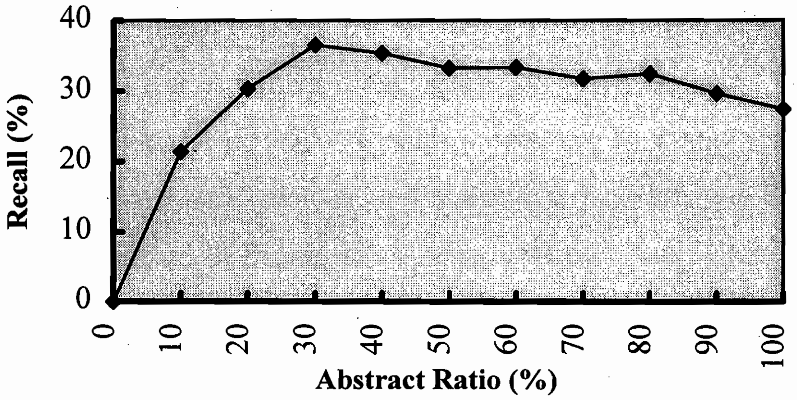

Figure 3 Recall as Function of Abstract Ratio

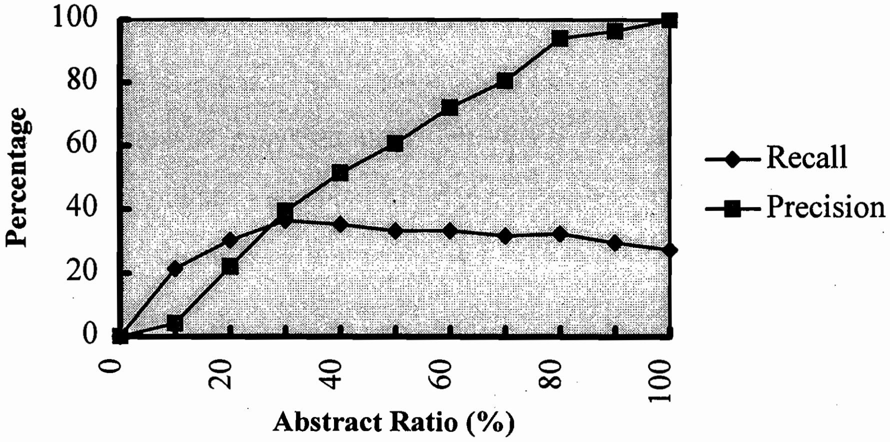

Figure 4 Recall vs. Precision as Functions of Abstract Ratio

6. Conclusions

Text abstraction entails the process of determining which sentences in a text contain the most important information that the author intends to convey to his readers. Our empirical study shows that this set of essential sentences consists of a relatively small fraction of the original text. Based on their comprehension of the text, human subjects, *behaving as a group*, are able to pinpoint this set of sentences relatively easily and consistently.

ACFAS is an automated Chinese full-text abstraction system, which extracts essential sentences from a given text following the analysis of its discourse structure. This process of ACFAS relies mainly on the presence of various discourse connectives in the text. By comparing the sentences identified as important by ACFAS with those identified by human subjects, *who presumably use additional cues*, our study shows that there is a non-random correspondence between these two sets of sentences. Since ACFAS, in its current design, does not include deep semantic processing to understand the meaning of each sentence in a text, we can conclude as follows: Which information in a text perceived by its readers as important depends not only on its semantic content, but also on how it is presented in a text, i.e. its discourse structure.

As a final remark, text abstraction represents a unique human faculty, which involves intelligent manipulation of given and assumed knowledge and natural language. Therefore, it is our belief that no single factor can guarantee its successful execution. Relevant factors or cues that had been used in the design of automated text abstraction systems include keywords, word frequency counts, discourse connectives, rhetorical relations, tense, distance from the beginning and the end of a text, just to name a few. However, there is a general negligence of systematic and quantitative evaluation of the relative contribution of each individual factor to the whole process of text abstraction. The present paper, by concentrating on the factor of explicit discourse connectives within a text, is a step toward improving this situation.

References

Allen, J., *Natural Language Understanding, 2^{nd} Edition*, Reading, Benjamin/Cummings, Redwood City, CA, 1995.

Grosz, B.J. and C. Sidner, “Attention, Intention, and the Structure of Discourse,” *Computational Linguistics* 12:3, 1986, pp.175-204.

Hirst, G., “Discourse Oriented Anaphoral Resolution in Natural Language Understanding: A Review,” *Computational Linguistics* 7:2, 1981, pp. 85-98.

Ho, H.C., B.K. T’sou, Y.W. Chan, B.Y. Lai and S.C. Lun, “Using Syntactic Markers and Semantic Frame Knowledge Representation in Automated Chinese Text Abstraction,” in *Proc. 1st Pacific Asia Conf. On Formal and Computational Linguistics*, Taipei, 1993, pp. 122-131.

Hwang, C.H. and L.K. Schubert, “Tense Trees as the ‘Fine Structure’ of Discourse,” in *Proc. 30th Annual Meeting, Assoc. for Computational Linguistics*, 1992, pp. 232-240.

Lin, H.L., B.K. T’sou, H.C. Ho, T. Lai, C. Lun, C.K. Choi and C.Y. Kit, “Automatic Chinese Text Generation Based on Inference Trees,” in *Proc. ROCLING Computational Linguistic Conf. IV*, Taipei, 1991, pp. 215-236.

Litman, D.J. and J. Allen, “Discourse Processing and Commonsense Plans,” in Cohen et.al.(ed.), *Intentions in Communications*, 1990, pp. 365-388.

Luhn, H.P., “The Automatic Creation of Literature Abstracts,” *IBM Journal of Research and Development*, 2:2, 1958, pp. 159-165.

Mann, W.C. and S.A. Thompson, “Rhetorical Structure Theory: Description and Construction of Text Structures,” in Kempen(ed.) *Natural Language Generation: New Results in Artificial Intelligence, Psychology, and Linguistics*, 1986, pp. 279-300.

McKeown, K.R., “Discourse Strategies for Generating Natural-Language Text,” *Artificial Intelligence* 27:1, 1985, pp. 1-41.

Ono, K., K. Sumita and S. Miike, “Abstract Generation based on Rhetorical Structure Extraction,” *Proc. Coling’94*, 1994, pp. 344-348.

T’sou, B.K., H.L. Lin, H.C. Ho and T. Lai, “From Argumentative Discourse to Inference Trees: Using Syntactic Markers as Cues in Chinese Text Abstraction,” in *Proc. 3rd International Conf. On Chinese Information Processing*, Beijing, China, 1992, pp. 76-93. Also appeared in C.R. Huang, K.J. Chen & B.K. T’sou (ed.) *Readings in Chinese Natural Language Processing*, Monograph Series No. 9, Journal of Chinese Linguistics, 1996, pp. 199-222.

T'sou, B.K., H.L. Lin, H.C. Ho, T. Lai and Terence Chan, "Automated Chinese Full-text Abstraction Based on Rhetorical Structure Analysis," *Computer Processing of Oriental Languages* 10:2, 1996, pp. 225-238.

Watanabe, H., "A Method for Abstracting Newspaper Articles by Using Surface Clues," *Proc. Coling'96*, 1996, pp. 974-979.