A Knowledge-based Machine-aided System for Chinese Text Abstraction[1]

Benjamin K. Tsou
Hing-cheung Ho
Tom Bong-yeung Lai
Caesar Suen Lun
Hing-lung Lin

City Polytechnic of Hong Kong

Hong Kong

Introduction

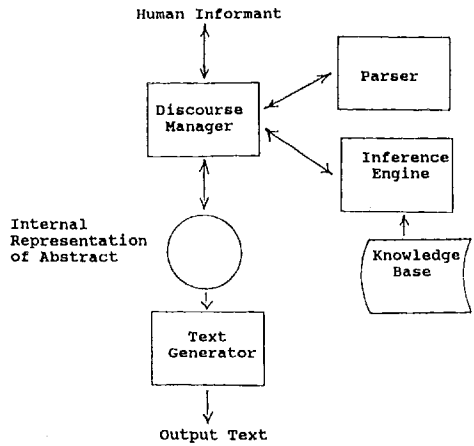
The production of abstracts from input source texts, using computers, is a subject in natural language processing that has attracted much attention and investigative study. It not only poses interesting theoretical challenges but also promises useful practical applications. At the City Polytechnic of Hong Kong, a large-scale research project on automated Chinese text abstraction has entered its third year. The project aims to investigate the issues related to text abstraction through the building of a prototype system.

Recognising the impracticality, at this stage, of attempting to construct a fully automatic system for abstracting random free texts, we have adopted a pragmatic approach by defining three design parameters at the outset : (1) The input texts consist of Chinese newspaper editorials. (2) The universe of discourse is on the safety of the nuclear power plant at Daya Bay (which is situated some 50 Km to the east of Hong Kong). (3) The target system will be fully automatic only at the final text generation stage, but will enlist informant input in the text understanding stage.

The result of our investigations and efforts is a prototype known as Machine-Aided Chinese Text Abstraction System (MACTAS) [6]. To begin the process of text abstraction in MACTAS, an unsophisticated human informant first reads and understands a given Chinese editorial. Based on his understanding of the editorial, he will go through an open-ended question-answering session with the system. At the conclusion of the human-machine dialog, which draws on a previous project [7], MACTAS will generate an abstract of the editorial in Chinese.

System Architecture

The system architecture of MACTAS is illustrated in the diagram below :



[1]Research reported here has been supported by Earmarked Research Grant No. 904007 from the University & Polytechnic Grants Committee of Hong Kong.

The Parser performs linguistic analysis on the response provided by the informant, and delivers relevant syntactic and semantic information to the Discourse Manager. Guided by the Inference Engine, the Discourse Manager supervises the interaction with the informant, producing an internal representation of the abstract. The Inference Engine draws on the knowledge contained in the domain-specific knowledge base to provide the basis for extracting the flow of argumentation of the editorial during the dialog session with the informant. Finally, the Text Generator transforms the internal representation of the abstract into output text.

The remaining sections of this paper will discuss the salient features of each of these system components.

Knowledge Base Management and Inferencing

Knowledge about the universe of discourse is derived from a detailed analysis of newspaper editorials on the subject, and codified into the Knowledge Base, which consists of a set of if-then rules. Rules can be used for deductive reasoning, and thus are best suited for capturing the flow of argumentation in editorials. In MACTAS, these rules are used to trigger queries to be posed to the informant, for the purpose of identifying conclusions in the editorial and the chains of reasoning leading to such conclusions.

The basic building-block of a rule is the predicate, which is an assertion that something is true. Since the Knowledge Base plays a crucial role in the process of agrumentation elicitation, a fundamental requirement is that the queries posed to the informant will be in simple natural language. To this end, nine predicate templates have been designed. The design of such templates is based on linguistic knowledge of Chinese. Each predicate reflects a simple Chinese sentence, based on the subject-predicate sentence structure.

On the macro level, a focus structure is used to provide a framework for partitioning the rules in the Knowledge Base into logically related groups. The purpose of partitioning is twofold. Firstly, it enables the Discourse Manager (discussed below) to identify relevant conclusions in the editorial in an efficient manner. Secondly, it aids the text generation process during the production of the final abstract.

The Inference Engine performs inferencing on the Knowledge Base to support the Discourse Manager in identifying possible conclusion(s) in an editorial, and in assisting the informant to relate the flow of argumentation leading to each conclusion. Key features of the Inference Engine, designed to simulate the dialog between two human beings, include : (1) It does not query the informant about facts either deducible from the Knowledge Base or from the current dialog. (2) Within the context of a given inferencing goal, consecutive queries bear a logical relationship to one another, thus defining the flow of argumentation leading to a particular conclusion. (3) It does not mechanistically pose queries to follow every step that leads to a conclusion. Rather, it is possible to jump ahead more than one step, to minimize the number of queries asked of the informant before a conclusion is reached.

Discourse Management

The Discourse Manager [1] controls the interactive session with the informant, in accordance with a discourse model for Chinese editorials. The design of this discourse model is based on the concept of schemata proposed by McKeown [5]. There are four schemata for an editorial : *(1) Background schema,* recording items that are common to editorials in general, such as name of the newspaper. *(2) Identification schema,* indicating the conclusions put forward in the editorial. *(3) Argumentation schema,* expounding the logical flow of arguments leading to each conclusion. *(4) Recommendation schema,* containing suggestions by the editorial writer for possible actions. The discourse model provides the base framework for organising the internal representation of the abstract which is subsequently used by the Text Generator for producing the output text.

The Discourse Manager performs a four-phase dialog session with the informant, each phase being designed to capture relevant information

for filling a corresponding schema in the discourse model. In particular, it calls upon the services of the Inference Engine for filling the argumentation schema, which constitutes the main body of the abstract.

Parsing

As MACTAS does not aim at full-text understanding for the production of abstracts, the main function of the Parser is to support the Discourse Manager in the task of response evaluation. A data-driven deterministic parser for Chinese sentences has been constructed. The syntactic rules used are adapted from the syntactic framework for Chinese sentences proposed by Zhu [8]. Because Chinese sentences do not always require both subject and predicate in the surface form, a bottom-up, rather than a top-down, parsing algorithm has been implemented. Moreover, to avoid unnecessary backtracking and the build-up of non-occurring constituents, the look-ahead principle of Marcus [4] has been adopted to check contextual information. Thus, tests are performed before a suitable path is selected. For the current prototype, conjoined sentences can be parsed, but not complex sentences with multiple verbs.

Text Generation

The Text Generator performs generation of Chinese abstracts at three conceptual levels : (1) Discourse model level. This determines the structure and content of an editorial, based on the discourse model mentioned above. Its role is to monitor the progress of succeeding utterances so that the output text is well organised. (2) Rhetorical structure level. This generates a paragraph of logically related multiple-sentence text using suitable Chinese conjunctions so that a coherent and rhetorically sound output text can be produced. Its design is based on the rhetorical structure theory proposed by Mann and Thompson [3], adapted here for the Chinese language [2]. (3) Single clause level. This is responsible for generating a grammatically correct clause for a given predicate, using the predicate templates discussed earlier.

Conclusion

Communication using language is a unique ability which distinguishes humans from the other members of the animal kingdom. The ability to simplify information, while retaining essential logical meaning and structure, represents a yet higher order of faculty. The successful attempt through MACTAS to simulate some aspects of this process would not have been possible without recognising the truly interdisciplinary nature of computational linguistics. MACTAS, as it now stands, could offer significant functional value within the devoloping language industry. The comparatively low cost associated with unsophisticated 'informants' in using systems such as MACTAS for the production of abstracts could be more cost-effective than the employment of subject specialists for large-scale and regular monitoring of textual information in specific domains. The abstracts thus produced provide the initial screening of information contained in a growing corpus which may be subsequently selected and examined at a higher level. Further efforts in MACTAS will be aimed at improving: (1) The range of natural language it can process and generate, such as the inclusion of complex sentences with multiply embedded verbs, (2) The refinement of the Discourse Manager on the basis of an enlarged corpus base, and (3) The deductive ability of the system, including consistency checking to cope with infelicities natural to humans.

A Note on Implementation

MACTAS was implemented in a personal computer environment under DOS, using Turbo PROLOG and the Eten Chinese System. To circumvent the limitations imposed by a single PC, the full system runs on two PC's interconnected by an RS-232C interface. A single-PC demonstration version is also available.

References

[1] Ho, H.C., Tsou, B.K., Lin, H.L., Lai, T.B.Y. and Lun, C.S. "A Knowledge-Driven

Discourse Manager For Chinese Editorials." Paper to be presented at the International AMSE Conference on Modelling and Simulation (New Orleans, Oct 1991).

[2] Lin, H.L., Tsou, B.K., Ho, H.C., Lai, T.B.Y., Lun, C.S., Choi, C.Y. and Kit, C.Y. "Automatic Chinese Text Generation Based On Inference Trees." In Proceedings of the ROCLING 1991 Computational linguistics Conference IV (Taiwan, Aug 1991), pp. 215-236.

[3] Mann, W.C. and Thompson, S.A. "Rhetorical Structure Theory : Description and Construction of Text Structures." In Natural Language Generation -- New Results in Artificial Intelligence, Psychology and Linguistics, G. Kempen (Ed.). Martinus Nijhoff Publishers (1987), pp. 85-95.

[4] Marcus, M.P. A Theory of Syntactic Recognition for Natural Language. MIT Press (1980).

[5] McKeown, K. "Discourse Strategies for Generating Natural-Language Text." In Readings in Natural Language Processing, B.J. Grosz et al. (Eds.). Morgan Kaufmann Publishers, California (1986), pp. 479-500.

[6] Tsou, B.K., Ho, H.C., Lin, H.L., Liu, G.K.F., Lun, C.S. and Heung, A.Y.L. "Automated Chinese Text Abstraction : A Human-Machine Co-operative Approach." In Proceedings of the International Conference on Computer Processing of Chinese and Oriental Languages (Changsha, Apr 1990), pp. 57-62.

[7] Tsou, B.K., Lun, C.S., and Heung, A.Y.L. "An Open-ended Chinese Question-Answering System (ECQUAS)." In Proceedings of the 13th International Conference on Computational Linguistics, Helsinki (ed. H. Karlgren) (1990) Vol. 1.

[8] Zhu, Dexi. Yufa Jianyi (Lectures on Chinese Grammar). Commercial Press, Beijing (1982).