AUTOMATIC ABSTRACTING OF TEXTUAL MATERIAL

Stephen L. Taylor
Wichita State University
Wichita, Kansas 67208

Gilbert K. Krulee
Northwestern University
Evanston, Illinois 60201

Lawrence T. Henschen
Northwestern University
Evanston, Illinois 60201

In what follows, we want to describe a system for the automatic abstracting of textual material. In designing the system, major theoretical questions have arisen not unlike those that arise in dealing with any natural language system. In addition, we want to describe some proposed revisions which have important theoretical implications and which should lead to significant improvements in the capabilities of the present system.

Our initial efforts at automatic abstracting began in 1969 as part of a more general question-answering system (Tharp, 1969; Tharp and Krulee, 1969) which made use of short stories about famous discoveries taken from a children's encyclopedia. The original text was mapped into a set of predicates representing the logical content of the story. Although the main emphasis in this system was on question-answering, there were two questions that could be asked that made use of a summarizing and abstracting capability. These were: "Who is the central character in the story?" and "What is the main theme of the story?"

In our current efforts (Taylor, 1975), we have extended Tharp's methods in order to deal more explicitly with the problem of automatic abstracting. The system represents the meaning of a text in terms of the semantic networks of Simmons (1973) which are based on the case grammar relationships of Fillmore (1968). In this form of representation, the nodes are words or concepts while the arcs represent a case grammar relationship that exists between a pair of nodes. Using the graphical techniques of Ramamoorthy (1966), the system identifies a portion of the original network, namely, the maximally connected subgraph (or graphs). Then using the techniques of signal flow graph analysis, the system identifies nodes that are most influential within the maximally connected subgraph. By proceeding iteratively, using this pair of techniques, a subgraph is obtained which serves as an abstract of the original text. As a final step, again using a technique due to Simmons and Slocum (1972), the subgraph is converted back into a set of natural language sentences as the final output.

The results of these initial attempts are reasonably encouraging although certain practical difficulties have been encountered. For example, even a short sample of text (one or two pages) leads to the formation of a complex network that is difficult to store. However, computer time for the processing of these networks is not excessive (less than ten seconds on a CDC 6400) and limitations of space are more serious than of computer time.

During the past year, we have introduced a series of modifications into the original system (Lindner, 1976). One of these has to do with texts that have multiple themes, such as a main theme and some subsidiary themes. In our original system, the first part of the process leads to the identification of several maximally connected subgraphs (MSC's). However, only one of these—the largest--is chosen as the basis for developing an abstract. This subgraph usually does contain the main theme plus some related details. A secondary theme may well be contained in a second MCS. Thus, by choosing only a single MCS, one biases the system towards the development of an abstract that overemphasizes details relating to the main theme while ignoring secondary themes.

Accordingly, we experimented with longer multiple-theme texts and with a procedure that would select the largest MCS plus one or more additional MCS's. The resulting abstract appears to be much improved. For example, using as a text a book review, three important MCS's were identified, the first concerned primarily with a discussion of the author and the second and third with the main character of the book and the theme of the novel, respectively. Thus, an abstract making use of all three leads clearly to a much more balanced presentation than does an abstract of equal length that refers only to the author.

Our original program was unable to handle networks of more than 300 nodes. Basically, the revised technique deals with the text, paragraph by paragraph. As a first step, the network associated with each paragraph is reduced, one at a time, thus obtaining a sequence of reduced networks, one from each paragraph. These networks are recombined into a single network and converted back into an abstract. These abstracts obtained are not unlike those obtained through the use of the original method. Most importantly, with this modification, it is the length of each paragraph that is critical rather than the length of the text as a whole, although, as the number of paragraphs increases, one again runs into difficulties in storing the reduced networks for each of a series of paragraphs.

We now find ourselves faced with the problem of introducing modifications that would lead to a significant improvement in the capabilities of this abstracting system. We propose to introduce some radical changes of which the following are perhaps the most significant.

1. In many respects, networks representative of the meaning of a text make use of logical predicates as the basic semantic units and the higher-level semantic unit that is computed is the product of a "bottom-up" form of analysis with the propositions being related in pairs until the

overall network emerges as a final product. As an alternative, we propose that the higher-level semantic analysis should be "top-down" in the sense of making predictions about the overall thematic structure of the material being processed. Moreover, this analysis should make use of a higher level or semantic grammar much like the thematic grammar that Rumelhart (1975) has proposed for the thematic analysis of children's stories.

2. Secondly, we are assuming that abstracting is normally a dynamic process that should be primarily logical or qualitative in form and making use of what we might refer to as substitution or condensing operators. Formally, these operators might resemble an axiom or theorem in mathematics, stating that certain content can be substituted for certain other content. These operators will "condense" in the sense of taking a network representing a portion of text and replacing it with a simplified network or perhaps a single node. Sometimes, the basis for condensing the text is made explicit in the text, in which case the operator is not unlike a procedure for identifying certain types of phrases in context. For example, one often encounters sentences in the form: "There are two main methods for the production of_____." or "When constructing a_____, one immediately faces three types of problems." Thus, in the abstract, one wants to identify in one case the two main methods and in the other case the three types of problems. Moreover, one probably wants to name the methods or problems while ignoring all of the amplifying details. Unfortunately, under many circumstances, such strong clues may not be given explicitly and must be inferred in much the same way that answers not explicitly contained in a data base can be inferred by problem solving or theorem proving methods. Thus, in our revised system we want to include a capability for "proving" that a set of summarizing statements can be inferred from the original data base.

In short, we are proposing two major modifications to our present abstracting program in order to make the system perform in a moce "human-like" fashion and in order to develop a system with a significantly improved level of competence.

REFERENCES

Fillmore, C.J. The case for case. In E. Bach and R.T. Harms, (Eds.), Universals in linguistic theory. New York: Holt, Rinehart, and Winston, 1968.

Lindner, J.A. Explorations in automatic abstracting by applying graphical techniques to semantic networks. Northwestern University, unpublished master's thesis, 1976.

Ramamoorthy, C. Analysis of graphs by connectivity considerations. Journal of the ACM, 1966, 13, 211-222.

Rumelhart, D.E. Notes on a schema for stories. In D.G. Bobrow ani A. Collins, (Eds.), Re^resienta^tion and Under standing, New York: Academic Press, 1975.

Simmons, R. Semantic networks: Thair computation and use for understanding English sentences. In R.G. Schank and K.M. Colby, (Eds.), Computer models of thought and Language. San Francisco: W.H. FreemanT 1973™

Simmons, R. and Slocum, J. Generating English discourse from semantic networks. Cqyiunlcations of til*-ACM, 1972, 15, 891-905.

Taylor, S. Automatic abstracting by applying B^fel^^l-^0^^^^-,^0 semantic networks. Unpublished doctoral dissertation. Northwestern University, 1975.

Tharp, A. Usin% relational operators to structure long term memory. Unpublished doctoral dissertation, Northwestern University, 1969.

Tharp, A. and Krulee, G.K. Usjj^jreJ^y^onal operators to structure long term memory. Proceedings, IJCAI, May7 1969. Pp. 579-586.