TEXT ANALYSIS: SESSION INTRODUCTION

Donald E. Walker
Artificial Intelligence Center
SRI International
Menlo Park, CA 94025

Text analysis is a promising area for applications of computational linguistics, with regard both to the prospects for technological achievement and to the potential benefits that those achievements can provide. Books, articles, and documents in text form are a major source of information. Being able to use computers more effectively in the analysis of these materials would have significant implications for virtually every intellectual activity. Moreover, developments in this area are essential if mankind is to take advantage of as well as merely to cope with the vast quantity of textual data already stored in computer-readable form, an amount that is increasing rapidly because of the use of computer-directed photocomposition for printing and the practicality of optical character recognition techniques.

I find it useful to distinguish six kinds of operations that can be performed on a given body of text:[1]

(1) Segmentation--break the text up into its constituent elements; at the lowest level this operation identifies the "elementary units" in the text, presuming, of course, that there is a particular set of criteria for their determination; somewhat higher, it segregates grammatical and semantic units within a sentence; more sophisticated procedures delineate discourse structures.

(2) Representation--characterize the information contained in the text; this operation can be as simple as assigning a set of index terms or as complicated as creating a set of predicate/argument structures relating the various elements in the text; summaries and abstracts of the text constitute intermediate representations.

(3) Classification--identify the text as similar to and different from other texts in relation to a set of predetermined categories; this operation establishes the position of the text in some more general framework.

(4) Modification--change the wording of some part of the text; this operation corresponds to the tasks of rewriting parts of the text as well as making corrections; it begs the question of when the modification is sufficiently large to result in considering the text to be new.

(5) Conversion--transform content elements from the text into some other nontextual or at least nonsequential structure; in this operation information is extracted from the text and reorganized according to externally determined criteria.

(6) Differentiation--locate particular constituents within a text; this operation finds those elements that wholly or partially match a given specification.

It should be clear, on reflection, that these operations overlap in complex ways; some presume others; moreover, their effects are strongly context dependent, reflecting the purpose and the particular framework for the analysis. While I make no strong claims for their utility, I believe that it is important for the field to distinguish the different kinds of things that people want to do with texts.

The six papers included in the sessions on text analysis at this conference illustrate the beginnings of a technology that will allow us to address some of the underlying issues. Three of them deal with the problem of conversion; specifically, they show how information can be extracted from a text and formatted for storage in a database.

In "Specialized information extraction: automatic chemical reaction coding from English descriptions," Larry H. Reeker, Elaine M. Zamora, and Paul E. Blower present a system that extracts information on chemical reactions from the experimental sections of papers in specialized chemistry journals, converting it into a format that chemists use to identify that kind of data.

James R. Cowie, in his paper "Automatic analysis of descriptive texts," describes a system for interpreting texts that contain stylized descriptions, like those in catalogues and directories. He shows how examples from a field guide to wild flowers can be processed to identify attributes characteristic of plants, which are then stored in a canonical form.

---

[1] Note that generation, the creation of the text itself, is presumed for this discussion, and that translation of a text into another language is also not included.

"Automatic representation of the semantic relationships corresponding to a French surface expression" describes the procedures Gian Piero Zarri is using to analyze statements about French historical events so they can be translated into the complex semantic data definition language established for the RESEDA system.

All three of these systems entail selecting texts with specialized language characteristics, restricted subject matter, and a limited number of discrete parameters into which the information extracted is to be stored. None of the papers discusses the procedures for accessing the information once stored, although all consider such procedures to be an essential component of a more comprehensive retrieval system. RESEDA is a special case, because it already has a facility established to retrieve data stored in the system and to make inferences based on those data.

In contrast to the preceding papers, "'Expertness' from structured text? RECONSIDER: a diagnostic prompting program," by Mark S. Tuttle, David D. Sherertz, Marsden S. Blois, and Stuart Nelson, begins with a text in which the information is already highly structured. It contains, for a large number of diseases, descriptions of the medical characteristics associated with each, which are stored as an inverted file index. Presented with the findings for a particular patient, the system identifies the disease or diseases whose characteristics correspond most closely to them. The authors are particularly interested in determining how such a text can be modified to enhance the effectiveness of the retrieval process.

The paper by George Vladutz, "Natural language text segmentation techniques applied to the automatic compilation of printed subject indexes and for online database access," focuses on the segmentation of self-contained text fragments into semantic-related constituents. He shows how titles from bibliographic citations can be processed to identify phrases that can be used in a Key Word Phrase subject index. The method, intended for application over a broad range of topics, is based on a predetermined list of nominal syntactic patterns that can be recognized using a small, domain-independent dictionary.

"Using natural language descriptions to improve the usability of databases," by Carole D. Hafner and John D. Joyce, shows a set of procedures for identifying text phrases that correspond to classes of objects in a database. The authors have made extensions to the command language of a relational data management system so that users can name and describe database objects using natural language phrases. Computational linguistic techniques make it possible to recognize partly specified names and to allow partial recovery from ambiguity through an identification of context.

The approaches presented in these papers derive primarily from work on the syntactic and semantic analysis of structures at the lexical or sentence level. None as yet builds on more complex procedures that reflect discourse properties of text coherence and cohesion. Further research on pragmatics and speech acts certainly is essential for some kinds of progress. However, there are other directions to pursue. One that seems important is determining how existing knowledge resources can be applied to the problem. People certainly make use, both explicitly and implicitly, of a variety of aids when they analyze texts. Particularly worth considering are dictionaries, a number of which are now available in machine-readable form, reference materials that contain biographical, geographical, or other specialized kinds of information, and encyclopedias. We need to establish procedures that will allow these resources to be used in text analysis. It is also appropriate to note that we do not have much data on the features of any body of text that might serve as a standard for comparison, much less the detailed studies of the characteristics of texts of all kinds that will be essential to ensure continuing progress in this field. A systematic study of the varieties of texts and of the purposes for which they can analyzed would provide useful guidelines for research at this stage of our understanding.