Book Reviews

Computational Linguistics in Medicine

Werner Schneider and A-L Sagvall Hein, Editors

North-Holland Publishing Co., New York, 1977,
181 pp., $28.25, ISBN 0-444-85040-6.

The convening of an international conference on Computational Linguistics in Medicine and the substantive results reported in many of the papers in this volume of Conference Proceedings testify to the special relevance of research in computational linguistics to the problem of processing medical information. While no discipline, even the most mathematically oriented, escapes a dependence on language to record and transmit its findings, medicine relies for its daily operations on the ability of its practitioners to draw upon information in natural language form, both patient data and the medical knowledge relevant to treatment. Given the diversity and complexity of the problems that are handled, and the vast store of medical knowledge that must be remembered or consulted as part of the treatment process, it is no wonder that medicine is one of the first fields to seek computer aids for accessing and processing natural language information. The purpose of this conference was to explore whether developments in computational linguistics and artificial intelligence have something to contribute to this problem.

Two major directions of research are seen in the papers at this conference and in the field it represents. Broadly speaking, one stream of research draws upon the methods of artificial intelligence and is concerned with general mechanisms of knowledge representation and reasoning, in this case as they apply mainly to clinical decision making. Language processing as such is not much in evidence in these papers, but the need to draw upon knowledge in language form is the background and assumption of much of the work. On the other hand, there is a growing body of research devoted to developing techniques for analyzing and processing the natural language form of medical information. In this work, there is a trend toward representing the data in structures that take account of the semantic relations among terms in medical statements. Though at this time the two areas of research are still quite distinct, a common ground may develop in the future when the AI projects look deeper into their data sources, and the data processors seek more powerful systems for representing information.

The 18 papers in the volume are organized into three sections: Methodological Background (7 papers), Medical Projects and Applications (7 papers), and Aspects of Hardware and Software (4 papers). Cross-cutting these major divisions, the papers can be grouped by topic to indicate the range of present work in this area as represented by the Conference.

Five papers deal with **trends and general methods**. These are: Werner Schneider (Uppsala University Data Center), *The impact of CL and AI techniques on modelling in medicine*; Erik Sandewall (Informatics Laboratory, Linköping University, Linköping, Sweden), *Current trends in artificial intelligence*; Carl W. Welin (Dept. of Linguistics, University of Stockholm), *Semantic networks and natural language understanding*; Uwe Wein (University of Düsseldorf and Uppsala University Data Center), *On the representation of non-procedural knowledge*; R. Pfeifer et al. (University of Zurich and Uppsala University Data Center), *PSY-PAC: A formal system for the modelling of cognitive processes*.

Three papers deal with the **modelling of the process of clinical decision-making and the development of computer aids to the process**. These are: M. N. Epstein and E. B. Kaplan (Section on Medical Information Science, University of California, San Francisco), *Criteria for clinical decision making*; E. H. Shortliffe (Dept. of Medicine, Massachusetts General Hospital, Boston), *A rule-based approach to the generation of advice and explanations in clinical medicine*; Stephen G. Pauker, M.D. and Peter Szolovits (Dept. of Medicine, Tufts University School of Medicine, and Laboratory for Computer Science, M.I.T.), *Analyzing and simulating taking the history of the present illness: context formation*.

Six papers deal with the **representation and coding of the information in medical records**. These are: A. W. Pratt (Director, Division of Computer Research and Technology, N.I.H.), *The use of categorized nomenclatures for representing medical statements*; Ruby S. Okubo and Baldwin G. Lamson (Dept. of Data Processing, UCLA Hospital), *The human interface in natural language retrieval*; Anna-Lena Sagvall Hein (Uppsala University Data Center), *An approach to the construction of a text-comprehension system for X-ray reports*; J. van Egmond et al. (Dept. of Medical Information, University Hospital of Ghent), *Systematization in the registration of medical diagnostic statements*; M. de Heaulme and Ch. Mery (INSERM U 88, and Service de Rhumatologie, Hôpital Cochin, Paris), *REMEDE: An artificial language for clinical documentation*; K. Sauter et al. (4 affiliations, 2 from W. Germany and 2 from Swe-

den), *A data structure model for a medical information system.*

Two papers deal with **medical terminology on the word level**: M. Wolff-Terroine (Dept. of Scientific Information, Institut Gustave-Roussy, Villejuif, France), *Terminology and nomenclatures*; F. Wingert (Institut für Medizinische Informatik und Biomathematik, Univ. Münster, West Germany), *Morphosyntactic analysis of medical compound word forms*. One paper deals with the **design of a large retrieval system**: N. Banerjee (Data and Information Systems Group, Applications Software, Siemens AG, Munich), *CONDOR: Communication in natural language with dialogue oriented retrieval systems*. And one paper deals with a **possible architecture of a LISP machine**: Jack Urmi (Information Laboratory, Linköping Univ.), *Designing LISP oriented hardware*.

Space permits a more detailed treatment of only the main groups of papers. In the **general methods** set, Werner Schneider in the opening paper traces the background to the Conference in the work of a study group of the International Hospital Federation on the application of computer techniques in health care. The 1975 report of the study group, which is summarized in the paper, led to the organization of this Conference in recognition of the importance of research directed toward the development of techniques for the formalization of medical knowledge. Schneider describes and criticizes the early, and still dominant, attitude toward the use of computers in medicine: "The basic idea was that more data -- especially so-called 'hard'-data, i.e. measured by natural science methodology -- would produce more and better information and that more information would lead to better decisions. As a consequence a large number of sometimes huge databanks have been established, mostly based on a rather random choice of which data should be collected and treated by more or less advanced statistical procedures." In contrast, Schneider stresses the importance of developing models and formalizing techniques for the representation of medical knowledge.

In Erik Sandewall's characterization of Artificial Intelligence the dominant methodology of AI is "to implement limited-purpose systems in order to develop general methods and principles, and also in order to develop software tools." Sandewall lists a number of general methods that he sees emerging in AI research (practical, low-key parsing of natural language; assimilation of natural language into a data base; representation of natural-language-based knowledge in a data base; deductive search; search in a problem space; scene analysis) and charts the major applicational areas in terms of their use of these methods. Carl Welin's paper in this group is a straightforward summary of the major constructs of

semantic networks. Uwe Hein's paper proposes criteria for the design of a medium for the representation of knowledge and presents a 3-level system for this purpose, implemented in LISP. The PSYPAC system (Pfeifer et al.) is intended as an instrument for building models of cognitive processes in human behavior. What is proposed is a very general formalism for such descriptions.

To analyze the **clinical decision process** procedurally, to determine which parts of the process can (and perhaps can best) be performed by computer, to design such a system and engineer it for fail-safe operation and for physician acceptance, these are some of the questions faced in computational research on clinical decision making. Epstein and Kaplan provide a modular model of the clinical decision process in which the major modules are Input, Question-Answering, Data Base, and Clinical Decision Making. The latter module, which is based on papers in the medical literature, consists of successive iterations through the functional blocks of data acquisition, data analysis and plan formation with associated feedback between functional blocks. A set of questions that have to be answered for each patient problem is stated, and technical issues, such as the need to limit the search space, are also raised. In addition to the process model, a generalized knowledge representation model (based on prior work of Kulikowski and Weiss) is sketched. The last half of the paper is devoted to the formulation and discussion of criteria for the process and knowledge-representation models, and the issues both medical and technical associated with each criterion. Some of the authors' criteria include medical factors not commonly considered in AI models, e.g., utility-driven search strategy and modification of diagnosis therapy or therapy based on outcome information. There is a brief informative discussion of statistical approaches, production rules, and network representation methods as they apply to the clinical decision problem. The paper is a model of clear exposition and provides a comprehensive framework for thinking about the clinical decision process in precise procedural terms.

E.H. Shortliffe's MYCIN system for computer-based medical consultation is well know in the AI field and beyond. The paper in this volume can serve as a brief introduction to the design criteria of MYCIN and also the design itself for those who have not read other papers or the book on the system. The knowledge base in MYCIN is provided by physician-experts and is represented as production rules. A major design criterion was to provide the system with the ability to explain decisions to the physician user.

The approach to the clinical decision making problem taken by Pauker and Szolovits begins with an analysis of clinical hypothesis formation as it appears in the actual behavior of expert clinicians, with particular emphasis on the formation of the initial hypothesis in response to the history of the present illness. "Before making any decisions about diagnostic tests or treatments, or indeed before asking more than the most cursory, automatic, descriptive questions, the physician must limit his domain of consideration to an appropriately narrow context." The paper analyzes this process of context formation and describes a program, the Present Illness Program, written in CONNIVER and MACLISP, which simulates this process. The domain of the program is a patient who presents with the chief complaint of edema. A supervisory component acquires patient-specific data from the physician, checks it and stores it in a short term memory. The short-term memory also contains "demon" programs, compiled from a knowledge base in long-term memory, whose task is to match stored patterns against the patient-specific findings. The paper describes and evaluates the operation of the program. In addition, the authors offer some general remarks of interest on computational linguistics in medicine.

At some point (and usually at many points) in the health care process and in clinical research, **patient data** must be consulted. Where large numbers of patients are involved, the problems of organizing, coding and retrieving the data are crucial. Since much of the data is initially in natural language form, the approaches and systems which have evolved for actual use are extremely interesting from a computational linguistics point of view. These systems were not developed by computational linguistics, but in a number of respects they represent the most advanced applicational research in CL because of their commitment to provide large scale, real world natural language data processing. One of the earliest systems to adopt a natural language approach to report-storage and retrieval (though it does not perform linguistic analysis of the input string) is the Natural Language Retrieval System of UCLA Hospital , which has been in operation since the mid 1960's (Okubo and Lamson). Four data bases of impressive size are maintained (Surgical Pathology, Autopsy, Nuclear Medicine, Neuroradiology), a total in 1977 of 234,794 documents containing 4,252,385 words. Retrieval requests are formulated in conventional Boolean query logic by a search specialist employing a large thesaurus. The paper discusses factors which would influence the design of an interactive interface to replace the present human interaction of requester and search specialist in formulating the data base query. Suggestions based on extensive experience with the system include providing the user the ability to browse through selected parts of the thesaurus with updated frequency counts for terms and an easy way of referencing synonymous multiple word descriptors and cross-linkages that combine two or more synonym classes to represent another search entity.

Several papers in this group describe patient-data systems that provide a structuring of the natural language input strings in order to facilitate and sharpen retrieval. These systems are particularly interesting for CL because they employ (or imply) a syntactic/semantic model of the input information. Again, because of space limitations, only some of the systems can be discussed here, although all the papers in this group were substantive and worthwhile.

The system for automatic encoding of natural language pathology reports, developed at NIH from the mid 1960's and onward (Pratt), is based on look-up in a semantically structured phrase dictionary, the Systematized Nomenclature of Pathology (SNOP). SNOP terms are divided into four highly structured lists: Topography (names of body sites -- T-terms), Morphology (names of structural changes that occur in tissues as a result of disease -- M-terms), Etiology (causative agents of disease -- E-terms), Function (names of the physiological manifestations associated with disease -- F-terms). The automatic encoder identifies syntactic units, transforms certain morphologically related forms into the word form used in the SNOP vocabulary, looks up the units in the encoding dictionary (SNOP), and combines the resulting codes for the T, M, E, F terms into "TMEF statements" which compose the document surrogate. Translated back into English the complete TMEF statement might be read as a prototype pathology report sentence: This body site T has undergone morphological change M due to the causative agent E resulting in physiological manifestations F. The TMEF statement thus constitutes a fixed formatted record which is efficient for high precision retrieval and at the same time constitutes a syntactic-semantic model of the pathology data in the original natural language pathology report.

It is interesting to compare the TMEF (SNOP) formalization of pathology diagnostic statements with the artificial language REMEDE (de Heaulme and Mery), which is in use for the (manual) encoding of clinical reports in Rheumatology, Endocrinology, Urology, and Vascular Surgery in Hopital Cochin, Paris. As in SNOP, the point of view in REMEDE is that the nomenclature should reflect the semantic structure. The systematized nomenclature can be employed in a formally defined syntax; REMEDE has, for example, such symbols as < ("of"), > ("due to"), * ("treated by") and = ("equal" --

used to introduce the results of a treatment or examination). For vocabulary fields S for Symptoms, T for Topographies, E for Etiology, TR for Treatments and R for Results, the basic clinical sentence in this language is

$$S < T > E * TR = R.$$

Logical operators can be used, and further operators (e.g., for "between" and "qualified by") have been defined, so that relatively complex clinical statements can be expressed. Also a means for including time data is included. This artificial language is sufficiently close to the syntactic form of simple, straightforward natural language sentences expressing the same information that it would seem feasible to automate the coding process directly from natural language input, although this possibility is not discussed in the paper.

Taken as a whole, this volume shows the existence of an area of interest between the processing of medical information and the analysis and processing of language.

Naomi Sager, New York University

Linguistic Structures Processing - Studies in Linguistics, Computational Linguistics, and Artificial Intelligence

Antonio Zampolli, Editor

This book is a collection of good articles. It is not, however, a good collection of articles. The only connection between them is that their authors all lectured at the International Summer School on Computational and Mathematical Linguistics at Pisa in 1974. Each lecturer was asked to contribute a chapter to the book; some of the contributions were specifically written for it, while others are papers that the authors had published elsewhere. Although each article is good by itself, the book as a whole lacks a common theme, a logical progression from one article to another, and a common level of background knowledge expected of the reader.

Three of the articles taken together make a good survey of computational linguistics: *On natural language based computer systems* by Stanley Petrick, *Natural language understanding systems within the AI paradigm* by Yorick Wilks, and *Five lectures on artificial intelligence* by Terry Winograd. Although the articles are three to five years old, the issues they discuss are still among the most active research topics today. One strength is the variety of viewpoints on many of the same systems and issues. One weakness is the skimpy treatment of semantic networks and related graphs: Winograd, for example,

devotes two pages to them out of 123, while using eleven pages to reproduce the same SHRDLU dialog that he has been quoting for the past eight years. One absurdity is the placement of these introductory articles near the end of the book because the chapters are listed alphabetically by their authors' last names.

Three tutorials on techniques are *Synthesis of speech from unrestricted text* by Jonathan Allen, *Morphological and syntactic analysis* by Martin Kay, and *Lunar rocks in natural English* by William Woods. Allen's article is a short survey of the state of the art and current issues in speech synthesis. Woods describes the various phases of the LUNAR system; he doesn't give enough detail to enable a beginner to build his own system, but he gives enough motivation and references to show someone where to go for further information. Kay, however, buries the reader in detail, including 21 pages of traces from his parser. Such detail is acceptable in a technical report, but an article of this sort should put more emphasis on the reasons for these techniques. Some comparisons with the parsing methods of Petrick, Wilks, Winograd, and Woods would be especially useful since they are discussed elsewhere in the same book.

Two articles that relate computational questions to more general issues in linguistics and psychology are *Scenes-and-frames semantics* by Charles Fillmore and *Cognition: The linguistic approach* by David Hays. Fillmore's article meanders for seventeen untitled sections: he presents a wealth of observations that a semantic theory must account for, but he never attempts to systematize his observations or present a tentative theory of his own. Hays, on the other hand, has a short, tightly organized discussion of the psychological implications of cognitive networks. But his article is so vague and devoid of examples that it is hardly more than an extended abstract.

Four other papers, *'The position of embedding transformations in a grammar' revisited* by Emmon Bach, *Focus and negation* by Eva Hajičová, *Some observations concerning the differences between sentence and text* by Ferenc Kiefer, and *John is easy to please* by Barbara Partee, treat theoretical points in linguistics that are also important computationally. Yet none of the authors cite any computational or AI work in their bibliographies or make any attempt to relate their issues to computational methods. These four articles illustrate a frequent failing of interdisciplinary conferences: the speakers talk past one another without ever reconciling their vocabularies or coming to grips with common issues. (In their more recent work, Bach and Partee and their graduate students have been combining Montague