COMPUTERIZED LANGUAGE PROCESSING FOR MULTIPLE USE OF NARRATIVE DISCHARGE SUMMARIES

Naomi Sager*, Lynette Hirschman*,
and Margaret Lyman, M.D.+

*Linguistic String Project, +Department of Pediatrics,
New York University, New York, New York

Summary - At New York University, computer programs have been developed that convert natural language medical records into a structured data base, i.e. into a table containing the same information as the stored documents. In this form specific information can be quickly retrieved, and summaries of the different kinds of information in the documents can be automatically generated. The automatic conversion of the information from its free-text form to a tabular form is called information formatting. This paper describes the application of the information formatting programs to a small set of pediatric discharge summaries for hospitalizations due to sickle cell disease. The programs created a table of approximately 50 columns in which each different type of information in the documents appeared under a separate heading. From this, a retrieval program extracted instances where symptoms of possible infection preceded symptoms of painful crisis, as suggested by the literature on sickle cell disease. In answer to more detailed queries the program checked the time-order of findings within one document. The potential use of such tables in continuing medical education and other applications in the hospital setting are discussed.

Introduction

For most health care settings today there is a rapidly increasing demand for patient-related information to improve understanding and management of all diseases. In addition, such data is needed for review procedures, both internal and external, by local, state and federal agencies, third party payers, accreditation and continuing medical education activities. This demand for patient care data is being met largely by hiring and training more people to review, interpret and re-record what the doctor has written, and by adding to the number of times the doctor must write the same information. Simple tabulations of numbers of patient visits or hospital days is no longer sufficient; now it is: what for? what was done? why? what happened? what could be done differently? that must be addressed for each patient visit or hospital stay.

As long as the information consists of a number (hospital days), or an easily coded laboratory test or drug, it is relatively easily handled by computer processing. Hence billing and accounting procedures, census reports and utilization of services are generally now performed by computer in health care settings of any magnitude. When it becomes necessary to review the events during the hospital days, or the actions taken as a result of the tests, or the effects of the drugs on the patient's disease, computer systems become inadequate. There are simply too many words and phrases being used as essential descriptors of these events for the "numeric data computer programs" to handle.

Designers of computerized medical records systems have also expressed the need for integration of patient information and for comprehensive capture of clinical data. Writing on the basis of experience with the large Kaiser-Permanente Automated Multiphasic Screening Program, and drawing on discussions at the 1972 Conference on Technology and Health Care Systems in the 80's, Davis[1] formulated design objectives for a medical information system, which included:

> "(a) establishment of a data base which would be compatible not only with clinical service (comprehensive individual patient information) but also for systematic retrieval for various research requirements (including longitudinal studies across large numbers of patients);
>
> (b) ability to handle all possible clinically relevant medical data--present and future--and store it in such a way that it could be retrieved efficiently and do this for a projected population of over one million patients." (p. 2)

The objective to handle all possible clinically relevant medical data for both clinical service and research requirements raises fundamental questions about medical information and its formal representation. Pratt[2] emphasizes that, "the data that the medical professional reduces or aggregates by logical inference and deduction to provide for the care of a patient or to communicate a medical concept to a student or colleague or to describe a research opportunity are in the large majority non-numeric in form and are formulated almost exclusively within the constructs of natural language... The data are language data" (p. 101). The objective of the work reported in this paper is to apply computerized linguistic techniques of analysis to medical data recorded in their unrestricted natural

0[illegible]3-6/78/0000-0330$00.75 © 1978 IEEE

language form, so as to represent the information, without loss or distortion, in a structured format suitable for automatic retrieval of qualitative as well as quantitative patient information, for use in patient care, clinical research, and continuing medical education.

This large research objective, automatic structuring of clinical information from the written record, had its origin outside the medical area proper, in investigations into general structural properties of science writing[3] and into the feasibility of utilizing automated language processing in science information retrieval.[4] The Linguistic String Project (LSP) of New York University has developed over the years a powerful system for natural language syntactic analysis[5,6,7] that includes a central program ("parser"), a comprehensive computer grammar of English and an associated computer lexicon of basic science vocabulary (ca. 5,000 words), as well as a number of subsystems important for the analysis of complex sentences. With the success of the parsing program, the question arose as to whether, when the program was used in a given subject area, a meaning-preserving rearrangement of the output parse trees could be performed so as to align informationally similar segments of successive sentences. If the aligned segments were then semantically labelled according to their content, this would yield a regular structure ("information format") suitable for complex information processing and fact retrieval.

We adopted medical records as the testing ground for this hypothesis, on the basis that the medical sublanguage displayed a number of linguistic features that favor automatic text-analysis (specific vocabulary, fact-reporting sentence types, semi-stereotyped formulations). In addition, there appeared to be a need for a program which would bring qualitative and narrative patient information into the structured format of computerized medical records. In 1975, the LSP was fortunate to obtain a grant from the National Library of Medicine to investigate this possibility, working in conjunction with the Bellevue Hospital Information System, which routinely captured free-text pediatric patient records in computer readable form as described under Data Capture, below.

The type of document chosen for investigation was the hospital discharge summary. A study of the laconic "note-style" of medical records had shown that the departures from conventional English syntax in medical records could be absorbed into the computer English grammar[8], and an initial implementation for natural language radiology reports had shown that automatic text-structuring of the kind we envisioned was possible[9,10]. It was felt that the discharge summary, with its wide range of topics covered, its syntactic and semantic complexity, and its importance in medical documentation, would provide a decisive test of the linguistic programs. If the problems in implementing the analysis of discharge summaries could be solved then it would be reasonable to expect that other narrative encounter-documents could be handled by the same techniques. This paper describes the operation of the programs on an initial set of five pediatric discharge summaries received from Bellevue Hospital for hospitalizations associated with sickle cell disease.

Data Capture

The pediatric discharge summary (PDS) was one of many natural language medical documents captured, stored and retrieved for pediatric patient care within the context of an integrated computer-based, health information system developed at Bellevue Hospital between 1966 and 1976, to assist in provision of comprehensive health care to children and youth[11,12,13]. The computer programs employed were those previously developed by Korein and Tick[14,15]. Physicians' reports of outpatient visits, examinations and hospitalizations were either handwritten or dictated, and keyed in on-line by medical typists with computer-prompting of the paragraph structure appropriate to the type of report or document being entered. Daily batch processing rendered some of the documents retrievable in whole or in part at any of the several sites of patient care. Simple computer scanning of selected paragraphs for pre-determined information and re-arrangement of that data (e.g. immunizations and laboratory information) provided immediate availability in tabular form of important features of child health care.

One guiding principle in the use of natural language or narrative medical information (as opposed to the use of coding forms or checklists) was the hope that greatest cooperation from many different health professionals would come from their being allowed to do what was natural for them in that health care setting. The paragraph structure for each document type (125 types in all) was "tailored" to the particular medical or health-related need. The structure provided paragraph headings for otherwise free narrative in order to ensure completeness of history-taking and examination (Figure 1).

Hospital Discharge Summary Information Format

Figure 2 illustrates a small segment of the information format developed for automatic processing of hospital discharge summaries. Figure 2, A. shows the first few sentences of one of the input documents. Fig. 2, B. shows part of the FINDINGS portion of the information format for those sentences and the time information associated with each finding. The complete format has a TREATMENT section and a PATIENT STATUS section, the latter containing a FINDINGS portion (shown in part in Fig. 2, B.) for laboratory data, physical and developmental measurements, and quantitative and qualitative findings from the physician's examination or other sources. Each entry containing a verb or event-noun in the TREATMENT or PATIENT STATUS sections may have a TIME and/or MODIFIER section associated with it. Internally, the struc-

FIG. 1

NATURAL LANGUAGE DATA CAPTURE

A. PARAGRAPH STRUCTURE OF PEDIATRIC DISCHARGE SUMMARY

ID NO. _____ PEDIATRIC DISCHARGE SUMMARY

NAME _____ SEX _____

DATE OF ADMISSION _____ DATE OF DISCHARGE _____

LOCATION _____ BIRTH DATE _____

(REFERRING PHYSICIAN)
(REASON FOR ADMISSION)
(PERTINENT HISTORY)
PRESENT ILLNESS -
SIGNIFICANT PAST HISTORY -
(EXAMINATION ON ADMISSION)
(IMPRESSION ON ADMISSION)
(COURSE IN HOSPITAL)
(STATUS AT DISCHARGE)
(LABORATORY DATA)
(DIAGNOSES)
(PLAN AT DISCHARGE)
(RETURN APPOINTMENT)
(ABSTRACT)
(DOCTOR)

B. FIRST 5 PARAGRAPHS FROM TYPICAL PEDIATRIC DISCHARGE SUMMARY

(REFERRING PHYSICIAN) - NONE GIVEN.

(REASON FOR ADMISSION) - SWOLLEN, PAINFUL HANDS. VOMITING. SYMPTOMS OF 18 HOURS DURATION.

(PERTINENT HISTORY)

PRESENT ILLNESS - THIRD HOSPITAL ADMISSION. PATIENT HAS SICKLE CELL DISEASE. WAS WELL UNTIL 18 HOURS BEFORE ADMISSION. PATIENT BEGAN TO VOMIT. THEN SHE DEVELOPED PAINFUL HANDS. TEMP ELEVATION TO 101 DEGREES WAS NOTED. SHE WAS CONVALESCENT FROM CHICKEN POX AT A VISIT 1 WEEK BEFORE ADMISSION.
SIGNIFICANT PAST HISTORY - TWO ADMISSIONS FOR MENINGITIS. 1 TRANSFUSION REQUIRED AT EACH ADMISSION. HAS BEEN TAKING FOLIC ACID DAILY. PENICILLIN GIVEN WHEN THERE IS EVIDENCE OF INFECTION.

(EXAMINATION ON ADMISSION) - TMP 99.6, PU 120, RR 16, WEIGHT 19.5 LBS. WELL DEVELOPED, WELL NOURISHED. OCCASIONALLY HOLDS AND RUBS HANDS AND FOREARMS. NO MENINGISMUS OR ADNORMAL NEUROLOGIC FINDINGS. DORSAL SURFACES AND PROXIMAL PALMAR SURFACES OF BOTH HANDS ARE SWOLLEN, WARM AND PAINFUL.

(IMPRESSION ON ADMISSION) - HAND-FOOT SYNDROME WITH SICKLE CELL DISEASE.

FIG. 2

INFORMATION-FORMATTED DISCHARGE SUMMARY

A. DISCHARGE SUMMARY (EXCERPT)

(REASON FOR ADMISSION) - SWOLLEN, PAINFUL HANDS. VOMITING. SYMPTOMS OF 18 HOURS DURATION.

(PERTINENT HISTORY)

PRESENT ILLNESS - THIRD HOSPTIAL ADMISSION. PATIENT HAS SICKLE CELL DISEASE. WAS WELL UNTIL 18 HOURS BEFORE ADMISSION. PATIENT BEGAN TO VOMIT. THEN SHE DEVELOPED PAINFUL HANDS. TEMP ELEVATION TO 101 DEGREES WAS NOTED. SHE WAS CONVALESCENT FROM CHICKEN POX AT A VISIT 1 WEEK BEFORE ADMISSION.

B. FINDINGS PORTION OF FORMAT (FOR EXCERPT)

PARAGRAPH	CODE NO.	BODY-PART	BODY-MEAS.	NOR-MALCY	QUANT	SIGN-SYMPT	DIAGNOSIS	TIME
REASON FOR ADMISSION	ADPDS 3.1.1	HAND (PLURAL)				SWOLLEN		
		HAND (PLURAL)				PAINFUL		
	ADPDS 3.1.2					VOMITING		
	ADPDS 3.1.3					SYMPTOM (PLURAL)		OF 18 HOURS DURATION
PERTINENT HISTORY - PRESENT ILLNESS	HIPDS 3.1.2						SICKLE CELL DISEASE	
	HIPDS 3.1.3			WELL				UNTIL 18 HOURS BEFORE ADMISSION
	HIPDS 3.1.4					VOMIT		
	HIPDS 3.1.5	HAND (PLURAL)				PAINFUL		THEN
	HIPDS 3.1.6		TEMP		101 DEGREES			
	HIPDS 3.1.7			CONVA-LESCENT			CHICKEN POX	AT A VISIT 1 WEEK BEFORE ADMISSION

tured sentences are stored in tree form; a program "flattens" the subtrees corresponding to the format rows and generates a table, as shown in Fig. 2, B.

Linguistic Processing

Before computer processing can begin, a preliminary manual linguistic analysis of a representative sample of the documents to be processed must be made. This analysis uses some of the same methods that are later applied automatically to the complete set of documents, as illustrated in the diagram shown in Figure 3, but at this stage these methods are used as the first steps toward discovery of the information format appropriate to the document type.

The function of the first two steps of processing, namely, grammatical analysis ("parsing") and syntactic regularization ("transformation") is to obtain a representation of each sentence which shows in a uniform way the grammatical relations of the words to each other and is free of differences in form that do not affect the meaning. For example, consider the sentence in the EXAMINATION paragraph of Fig. 1, occasionally holds and rubs hands and forearms. We need to establish the association of the verb phrase holds and rubs with its object hands and forearms, via grammatical analysis, and then to expand the construction around the two occurrences of the conjunction and to reveal the four independent statements contained in the construction: holds hands, rubs hands, holds forearms, rubs forearms. For this, and other meaning-preserving rearrangements of English sentence structure, we use well-established paraphrastic linguistic transformations which have been implemented in the LSP system[16,17].

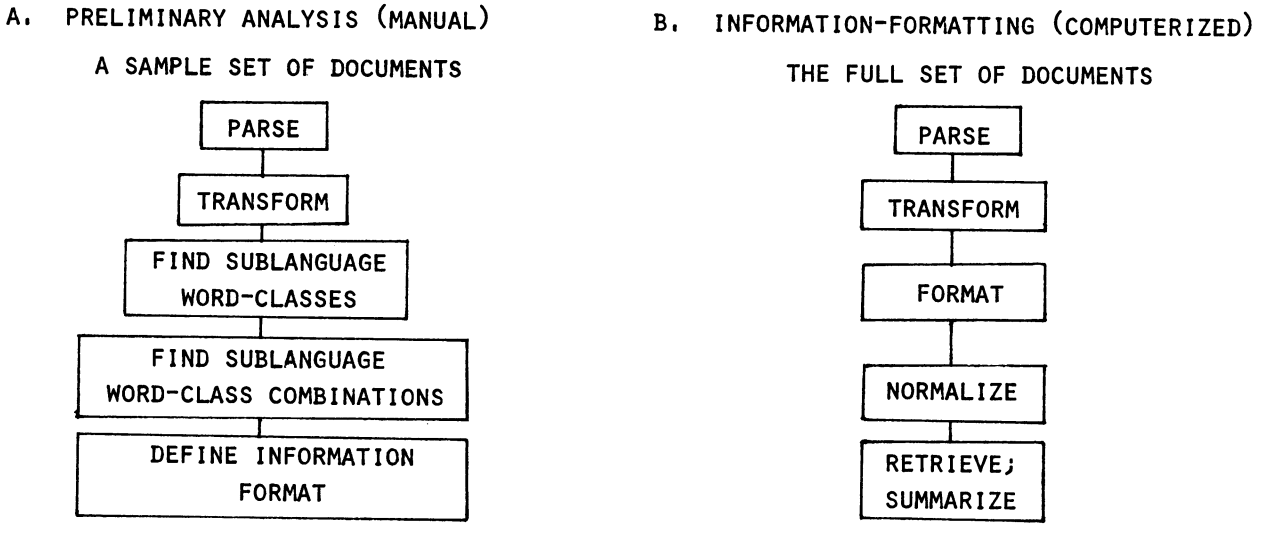
Breaking down the sentences this way into their individual components and arranging the components according to a few standard grammatical relations (mainly, subject-verb-object and head-modifier) makes it possible to discover patterns of word usage that are characteristic of the documents. This specialization of language use in a subject area is called a sublanguage. The grammar of the sublanguage provides the informational categories needed for processing the documents. Words fall into classes on the basis of their occurrence with words of other classes in subject-verb-object or modifier-host relations, obtained by sentence analysis. We thus obtain a class of disease-indicator words (pain, vomiting, etc.) that occur as objects of such subject-verb sequences as Patient complained of ____, Patient developed ____, which is distinct (on the whole) from the class of diagnosis words (sickle cell anemia, meningitis, etc.) which occur as the object of such verbs as diagnose (Patient had sickle cell disease diagnosed at age 4 months). Of course, the semantic affinity of words in a class is an aid to building the word classes when this job is done manually.* But the ability of computer programs to convert free text information into structured form rests on the regularity of linguistic occurrences. Segments which have similar informational standing in the subject matter occur regularly in the same or similar linguistic configurations in their respective sentences.

Once the word classes and their characteristic combinations are described, an overall table (or information format) is defined and a computer

*A clustering program that develops sublanguage classes automatically has also been written.[18]

FIG. 3

LINGUISTIC PROCESSING



dictionary giving both parts of speech and sublanguage classifications for the document words is constructed. This latter draws upon the English computer dictionary of the LSP and will also, we hope, be able to utilize the lexical classifications embodied in established medical terminologies such as the Systematized Nomenclature of Pathology (SNOP), the Systematized Nomenclature of Medicine (SNOMED, extension of SNOP), and the International Classification of Disease (8th and 9th Revisions).

A brief description of the stages of machine processing that convert free-text discharge summaries into a structured data base follows. A full description is in press[19]. As indicated in Fig. 3, the first stage is to parse the document sentences with the LSP parser. This stage begins with a dictionary look-up to associate the stored lexical information with each word occurrence. Then the text is parsed to obtain the basic syntactic relations in the sentence (e.g. subject-verb-object relations and head-modifier relations). Next each sentence undergoes a series of paraphrastic English transformations. These transformations preserve meaning but reduce the number of constructions to a set of basic syntactic relations. Next the parsed, transformed sentences of the text are mapped into the columns of the information format. In general there is a one-to-one correspondence between sublanguage word class and format column. For example, a word of the hospital sublanguage (h-) class h-diag (e.g. sickle cell disease) will be mapped into the DIAG column; or h-bodypart words (e.g. hand) are mapped into the column BODY-PART. However, certain subclasses of words are primarily modifiers; that is, they have meaning in terms of a word or phrase that they modify, for example time phrases, as in HIPDS 3.1.3 (see Fig. 2):

> Was well until 18 hours before admission.

Here, until 18 hours before admission is a time modifier on the adjective well. Modifiers must be formatted in such a way that they retain the information about what they modify. For example, there can be a time associated with the FINDINGS (the TIME column of Fig. 2, B) as well as a TIME column for the TREATMENT section of the format (not shown in Fig. 2, B):

> Patient was admitted on 11/6 for fever of two days.

Here, on 11/6 modifies the medical-action verb (h-vmd) admit (formatted under TREATMENT), and the time expression of two days modifies the sign-symptom word fever.

The parsing, transformation, and mapping into the format are done on a sentence-by-sentence basis. The next step, normalization of the data base, treats the document as a whole and handles connections between sentences (that is, sentences in context), rather than treating each sentence in isolation. This stage does two important things: it fills in missing BODY-PARTs and it establishes time relations between the sentences.

As an example of filling in a missing BODY-PART, take the sentence COPDS 3.1.3:

> Her hands remained somewhat warm and swollen; however pain seemed to subside.

The semi-colon divides the sentence into two independent pieces, each of which becomes a separate "line" in the table of formats, as shown in Fig. 2. The portion of the sentence after the semi-colon does not mention hands again; however, it is clear that this sentence is continuing the topic (her hands) started in the first part of the sentence. Certain symptom words require an associated body-part word, which if not present is sought in the preceding entry. Therefore, the phrase her hands is filled in under the empty BODY-PART column during normalization. Thus, in Table 2, described below, for the entry derived from COPDS 3.1.3, we see that her hands does appear, associated with pain.

The normalization program also computes time relations. As an example let us look at three successive sentences, HIPDS 3.1.3 - HIPDS 3.1.5, and the time relations established by the normalization program.

> Was well until 18 hours before admission. Patient began to vomit. Then she developed painful hands.

First the normalization program expands the sentence Was well until 18 hours before admission into two format entries equivalent to the two sentences:

a. Was well prior to 18 hours before admission.
b. Was not well at 18 hours before admission.

These sentences are assigned the time values:

a. time: < adm - 18 hr
b. time: = adm - 18 hr

The next sentence contains no overt time expression. Therefore since this is a narrative section of the discharge summary (HISTORY), the program assumes that the time of the second sentence is equal to or later than the time of the first sentence (Time b., above). This is because in narrative, time moves in a forward direction unless specifically marked (by another time expression) to the contrary. Therefore, sentence HIPDS 3.1.4 (Patient began to vomit) is assigned the time:

c. time: ≥ time of sentence b.

In the table, we convert all time units into days and round to the nearest day, giving adm - 1 day in this case.

Finally in sentence HIPDS 3.1.5 we have another time expression, then. This tells us that the time of sentence 3.1.5 is greater (later) than the time of sentence 3.1.4:

d. time: > time of sentence c.

However, it too appears as adm - 1 day in the table, since we do not know how much later. After normalization the data base is ready to use.

Application and Results

An example will serve to illustrate the potential usefulness of the discharge summary as data source for preparation of a clinical conference. The medical literature of sickle cell disease, a congenital life-long abnormality of red blood cells leading to chronic disease, describes the "painful crisis" as an associated acute problem, characterized by pain in bones, joints, especially of extremities, abdomen and back. In infants and young children a variety of painful crisis occurs, commonly called hand-foot syndrome, in which bony infarction of hands, fingers, feet and toes produces pain, swelling, tenderness and warmth of one or more of those areas. Painful crisis is often precipitated by infection, often of a mild viral type[20]. Question: In how many instances did patients hospitalized with painful crisis and sickle cell disease also have evidence of possible infection, i.e. signs and symptoms commonly associated with viral infection, prior to onset of pain?

In order to answer this question, we convert it into a retrieval program to operate on the database consisting of the normalized information formats for a set of discharge summaries. This program generates a summary of the relation between painful crisis and possible infection shown in Table 1. This summary tells us that there does appear to be a strong correspondence between signs and symptoms of infection and painful crisis in our data; in fact, in all 4 instances of painful crisis in this small sample, there were preceding signs and symptoms of infection.

In certain cases, the summary table is too succinct. For example, we might wish to see how the program determined that vomiting preceded the development of painful hands in document no. XXX571. For this we generate the more detailed table of document excerpts, showing all occurrences of symptoms of infection as well as signs of painful crisis which were used in arriving at the summary given in Table 1. Table 2 is the excerpt table for document no. XXX571.

Given the question In how many instances did patients hospitalized with painful crisis and sickle cell disease have evidence of possible infection prior to onset of pain, the first step in writing a retrieval program is to formulate the question as a set of Boolean and set theoretic operators on certain defined predicates (written in capitals in the following formula):

TABLE 1

PATIENTS WITH POSSIBLE SYMPTOMS OF INFECTION PRIOR TO DEVELOPMENT OF PAINFUL CRISIS

DOCUMENT #	PATIENT AGE	SYMPTOM OF POSSIBLE INFECTION PRECEDING PAINFUL CRISIS	TIME OF SYMPTOM RELATIVE TO ADMISSION
XXX570	8 MO	105 DEGREE FEVER	ADM - 1 DAY
XXX571	1 YR	VOMITING	ADM - 18 HR
XXX572	1 YR 2 MO	WATERY STOOL PLURAL	ADM - 4 DAY
XXX574	2 YR 5 MO	FEVER, ANOREXIA	ADM - 2.5 WEEK

TOTALS		PERCENTAGES
DISCHARGE SUMMARIES IN SAMPLE	5	
INSTANCES OF PAINFUL CRISIS RELATED TO SICKLE CELL DISEASE	4	80% OF DISCHARGE SUMMARIES
INSTANCES OF PAINFUL CRISIS ASSOCIATED WITH INFECTION	4	100% OF INSTANCES OF PAINFUL CRISIS
INSTANCES OF SYMPTOM OF POSSIBLE INFECTION PRECEDING PAINFUL CRISIS	4	100% OF INSTANCES OF PAINFUL CRISIS

TABLE 2

PAIN AND INFECTION EXCERPT TABLE

DOCUMENT No. XXX571, PATIENT NO. XXX322, PATIENT AGE - 1 YEAR

PAINFUL CRISIS		INFECTION	TIME	TEXT
SYMPTOM/ DIAGNOSIS	ASSOCIATED BODY PART	SYMPTOM/ LAB RESULT	WITH REF. TO CURRENT ADMISSION	
SWOLLEN	HAND PLURAL		ADM	ADPDS 3.1.1 SWOLLEN, PAINFUL HANDS.
PAINFUL	HAND PLURAL		ADM	
		VOMITING	ADM	ADPDS 3.1.2 VOMITING.
		VOMIT	ADM -	HIPDS 3.1.4 PATIENT BEGAN TO VOMIT.
PAINFUL	HAND PLURAL		ADM -	HIPDS 3.1.5 THEN SHE DEVELOPED PAINFUL HANDS.
		TEMP 101 DEGREE	ADM	HIPDS 3.1.6 TEMP ELEVATION TO 101 DEGREES WAS NOTED
RUB	HAND PLURAL		ADM	EXPDS 3.1.3 OCCASIONALLY HOLDS AND RUBS HANDS AND FOREARMS.
RUB	FOREARM PLURAL		ADM	
SWOLLEN	DORSAL SURFACE OF BOTH HAND PLURAL		ADM	EXPDS 3.1.5 DORSAL SURFACE AND PROXIMAL PALMAR SURFACES OF BOTH HANDS ARE SWOLLEN, WARM AND PAINFUL.
SWOLLEN	PROXIMAL PALMAR SURFACE PLURAL OF BOTH HAND PLURAL		ADM	
WARM	DORSAL SURFACE OF BOTH HAND PLURAL		ADM	
WARM	PROXIMAL PALMAR SURFACE PLURAL OF BOTH HAND PLURAL		ADM	
PAINFUL	DORSAL SURFACE OF BOTH HAND PLURAL		ADM	
PAINFUL	PROXIMAL PALMAR SURFACE PLURAL OF BOTH HAND PLURAL		ADM	
HAND-FOOT SYNDROME			ADM	IMPDS 3.1.1 HAND-FOOT SYNDROME WITH SICKLE CELL DISEASE.
WARM	HER HAND PLURAL		ADM + 1	COPDS 3.1.3 HER HANDS REMAINED SOMEWHAT WARM AND SWOLLEN; HOWEVER PAIN SEEMED TO SUBSIDE.
SWOLLEN	HER HAND PLURAL		ADM + 1	
PAIN	HER HAND PLURAL		ADM + 1	
SWELLING	BOTH HAND PLURAL		DISCH	STPDS 3.1.1 SLIGHT RESIDUAL SWELLING OF BOTH HANDS.
HAND-FOOT SYNDROME			ADM +	RXPDS 3.1.1 HAND-FOOT SYNDROME.
SWOLLEN	HAND PLURAL		ADM +	ABPDS 3.1.2 DEVELOPED SWOLLEN, PAINFUL AND WARM HANDS.
PAINFUL	HAND PLURAL		ADM +	
WARM	HAND PLURAL		ADM +	
		VOMITING	ADM -	ABPDS 3.1.3 HAD SEVERAL EPISODES OF VOMITING PRIOR TO ADMISSION.

$$\{ h \in \text{HOSPITALIZATION} \mid (\exists p \in \text{PAINFUL-CRISIS})(\exists i \in \text{INFECTION}) [\text{DURING}(i,h) \wedge \text{DURING}(p,h) \wedge \text{TIME-OF}(i) < \text{TIME-OF}(\text{ONSET}(p))] \}$$

We can read this formula as:

> Form the set of hospitalizations such that:
>
> there exists an instance of painful crisis which occurs during this hospitalization
>
> and there exists an instance of infection which occurs during this hospitalization
>
> and the time of the infection precedes the time of the onset of the painful crisis.

This formula gives the logic of the retrieval program. Next the predicates must be converted into retrieval routines. The set defined by the HOSPITALIZATION predicate is the set of discharge summaries. The predicate DURING with second argument h (a member of HOSPITALIZATION) is programmed as a search through discharge summary h for an instance of PAINFUL-CRISIS or INFECTION (the first argument of DURING). There are two additional time-relational predicates, TIME-OF and ONSET; and finally there are the two medical predicates PAINFUL-CRISIS and INFECTION.

In order to formulate a retrieval routine for the medical predicate INFECTION we need to know how a physician would retrieve this information -- that is, what constitutes evidence for infection? Table 3 shows the physician-defined list that forms the basis for the retrieval routine used to prepare Tables 1 and 2. The list furnished by the physician is supplemented by related words and phrases (e.g. given warm, the word hot is added) and by morphologically related forms (e.g. to fever is added feverish) in order to form word classes. The routine is then defined in terms of a combination of appropriately filled format columns within a given format row in a table like the one shown in Fig. 2.

As an example, Table 4 shows the logic of the routine corresponding to the medical predicate PAINFUL-CRISIS. Part 2 of this routine would, for example, retrieve the information in HIPDS 3.1.5 (Then she developed painful hands), shown in Fig. 2. The format row for this sentence has a word hand in BODY-PART which belongs to the EXTREMITY-WORD class; it has the SIGN-SYMPTOM word painful which belongs to PAIN-WORD; and the column NEG is empty, meaning that it is not negated (so CHECK-REAL will return a value real).

In order to determine whether infection preceded the development of painful crisis, we need another retrieval routine, shown below. This routine uses the COMPARE-TIME routine, which compares the times of any two events in the database. We can translate the question Did any signs or symptoms of infection precede the development of painful crisis? into the following steps:

1) compare the times for the events under the column PAINFUL CRISIS and save the

TABLE 3

MEDICAL INPUT FOR THE DEVELOPMENT OF A RETRIEVAL ROUTINE

LIST OF POSSIBLE SIGNS OF SYMPTOMS OF INFECTION USED IN ROUTINE INFECTION

0) ANY UNNEGATED USE OF THE WORD INFECTION;
1) FEVER (TEMPERATURE GREATER THAN 100.2°);
2) IRRITABILITY (FUSSY/IRRITABLE/CRIES/CRANKY/WON'T SLEEP);
3) COUGH (RALES/RHONCHI/DULLNESS TO PERCUSSION);
4) NASAL CONGESTION (RUNNY NOSE/RHINITIS/COLD);
5) LOSS OF APPETITE (NOT EATING/ANOREXIA);
6) RASH OR SKIN IRRITATION;
7) UPSET STOMACH (VOMITING/DIARRHEA/WATERY STOOLS);
8) HEADACHE, SORE THROAT, ACHINESS, STIFF NECK, OR EARACHE;
9) WBC > 50,000.

TABLE 4

ROUTINE PAINFUL-CRISIS

IF AN ENTRY MEETS CONDITION 1, 2, OR 3 THEN WRITE OUT UNDER PAINFUL-CRISIS THE ITEMS DIAGNOSIS, SIGN-SYMPTOM OR DESCRIPTOR, FOLLOWED BY THE ASSOCIATED BODY-PART (IF ANY).

1. DIAGNOSIS HAS CRISIS-WORD
   AND CHECK-REAL* (DIAGNOSIS) = REAL;

2. BODY-PART HAS EXTREMITY-WORD OR ABDOMEN-WORD
   AND SIGN-SYMPTOM HAS PAIN-WORD OR SWELLING-WORD OR WARM-WORD
   AND CHECK-REAL (SIGN-SYMPTOM) = REAL;

3. BODY-PART HAS EXTREMITY-WORD
   AND DESCRIPTOR HAS MOVE-WORD
   AND CHECK-REAL (DESCRIPTOR) = NEGATED

WORD CLASSES:

CRISIS-WORD = HAND-FOOT SYNDROME, PAINFUL CRISIS, SICKLE CELL CRISIS, CRISIS.
EXTREMITY-WORD = HAND, FOOT, LEG,...
ABDOMEN-WORD = ABDOMEN, ABDOMINAL,
PAIN-WORD = PAIN, PAINFUL, TENDER,...
SWELLING-WORD = SWELLING, SWOLLEN, SWELL,...
WARM-WORD = WARM, HOT,....
MOVE-WORD = BEAR WEIGHT, MOVE, STAND,...

*CHECK-REAL IS A FUNCTION OF ONE ARGUMENT (A FORMAT COLUMN) WHICH CHECKS FOR NEGATION OR UNCERTAINTY ASSOCIATED WITH THAT COLUMN AND RETURNS A VALUE REAL (NO NEGATION OR UNCERTAINTY), NEGATED OR INDEFINITE.

earliest time;

2) compare the earliest time associated with painful crisis to the time associated with each entry in column INFECTION;

3) If any events under INFECTION have an earlier time than the earliest time of painful crisis, answer YES and write a list of these events; otherwise answer NO.

The document for the second hospitalization contains an interesting contradiction with respect to this question. In the HISTORY section, we find the sentences:

> On the day before admission patient developed fever of 105 degrees.... This morning mother noted swelling and tenderness of left tibia and foot.

The question Did infection precede painful crisis? is answered YES on the basis of these two sentences. However, later in the ABSTRACT section, we find the following:

> At onset of fever there was mild extremity swelling and pain.

This of course contradicts the information given in the HISTORY section. This particular type of contradiction is currently not picked up; however, by making the program more sophisticated, it could be recognized and flagged so that the record could be reviewed and corrected.

Implementation

The LSP system described above is coded in

FORTRAN and is currently running on a Control Data 6600, requiring about 75,000 words of memory. The English grammar and transformations (including the formatting) are written in Restriction Language, a special language developed for writing natural language grammars[21,16]. The normalization of the data base and the retrieval program are implemented in LISP 1.5. Each of the four stages for the preparation of the data base (parsing, English transformations, formatting, and normalization) require three to four minutes of computer time per discharge summary. The LSP system has been developed for research purposes and we expect that a system using the same techniques but developed specifically for the processing of medical documents would run significantly faster.

Discussion

Given that all kinds of patient care information -- numbers, words, phrases, sentences -- could be subjected to computer processing, then the single recording by the physician of what was observed and done for the patient could be captured and made to serve many of the demands for these data. Office or outpatient examinations, hospital progress notes and summaries, results of procedures and tests, would become the data source, recorded only once, for immediate or later use.

It is not the purpose of this discussion to review past and present work in computerized medical record systems. Suffice it to say that while some systems contain language data useful to current patient care, these data are not presently utilized beyond occasional scans for presence or absence of selected words or phrases. How much more useful medical record systems might become if narrative information would be analyzed and integrated with quantitative data gathered during patient care.

Many organizations and agencies are striving to improve the quality of medical care, in part through continuing medical education (CME) programs. Postgraduate conferences are one such CME activity. A recent example, selected from many notices received by physicians and other health personnel, is the Postgraudate Conference of the Center for Sickle Cell Disease at Howard University, Washington, D.C. (Nov. 27-28, 1978), that will focus on "Opposing Views and Controversial Aspects of Therapy for the Clinical Care of Patients With Sickle Cell Disease." The participants in this conference are most likely to be those immediately involved in teaching, research and care of patients with sickle cell disease. Suppose the participants could arrive at the conference with an analysis, automatically produced, of discharge summaries of all patients with sickle cell disease hospitalized in their institution. A considerable body of clinical data would thus be available for discussion of "opposing views and controversial aspects."

Similar data would be useful at a clinical teaching conference where there is a review of the clinical features of the disease, including incidence in the population served, discussion of genetic aspects, age at onset, signs and symptoms, changing features with age, management of the basic underlying disease and of its complications, and overall outcomes, to mention only a few. Where the discussant at the clinical conference has a large experience in care of patients with that disease in his institution, a summary of key features of the data could be compared or contrasted with reports in the medical literature, as has been illustrated with a particular example earlier in this paper.

Another area of great concern in medicine (also mentioned, for example, in the announcement of the Sickle Cell Postgraduate Conference) is the need to stimulate controlled clinical trials so that controversial approaches can be placed on a more rational basis. Analysis of narrative data may permit ready identification of patients whose past treatment and course were such that they could serve as controls for any new program -- thus making it unnecessary to withhold a potentially beneficial therapy from some patients while applying it to others.

Some of the applications of "language data" gathered during patient care are listed in Table 5. The likelihood of improving individual patient care by better recall and organization of past information for current usage is obvious. Stimulated in large part by the rising costs of health care, many agencies are now involved in intensive review of both institutional and ambulatory care. With information formatting by computer, the summary of a particular hospitalization could be checked against a set of criteria established as representative of good quality care (medical audit or quality assessment). Only those which failed the check would be subjected to more intensive review by trained (and expensive) persons. A more detailed treatment of this application of computer-formatted discharge summaries has been given recently.[22]

With identification of a new disease or a newly developed laboratory test for an old disease, past patient records are being scrutinized in large numbers, to expand the knowledge about the new disease or locate patients who should have the laboratory test performed. A current example is the appearance of reports of Legionnaire's Disease having been present in hospitals other than in Philadelphia prior to August 1976. Use of a specific test for presence of antibodies to the agent of Legionnaire's Disease has confirmed that a similar outbreak of the disease occurred in a meeting of the Independent Order of Odd Fellows in 1974 at the same hotel where the American Legion met in 1976[23]. Thus the search for an environmental reservoir is greatly focused. The information formats described here would have found cases meeting the definition of this study: "...cough and temperature of 39°C or greater, or any fever and X-ray evidence of pneumonia with onset during the 2 weeks after the last day an Odd Fellows member attended the convention."

TABLE 5

SOME APPLICATIONS OF "LANGUAGE DATA" GATHERED DURING PATIENT CARE

INDIVIDUAL PATIENT CARE

NEW DIAGNOSTIC, MANAGEMENT & FOLLOW-UP PROTOCOLS
RE-EVALUATION REMINDERS ("FLAGS")

HEALTH CARE EVALUATION

UTILIZATION REVIEW
CONCURRENT HOSPITAL REVIEW
MEDICAL AUDIT/QUALITY ASSESSMENT OF PERFORMANCE IN PATIENT CARE
ASSESSMENT OF PATTERNS OF HEALTH CARE
ASSESSMENT OF TRAINING PROGRAMS

POPULATIONS AT RISK

EPIDEMIOLOGIC SURVEILLANCE
SCREENING PROTOCOLS

RESEARCH

IDENTIFICATION OF PATTERNS OF DISEASE
IDENTIFICATION OF GENETIC, FAMILIAL, ENVIRONMENTAL FACTORS IN DISEASE
IMPROVED UNDERSTANDING AND TREATMENT OF DISEASES
USE OF PREVIOUS CASES AS CONTROLS IN CLINICAL TRIALS
IDENTIFICATION OF ADVERSE REACTIONS TO PROCEDURES AND TREATMENTS

EDUCATION

CLINICAL CONFERENCES -- UNDERGRADUATE AND POSTGRADUATE
MORBIDITY AND MORTALITY CONFERENCES

ADMINISTRATIVE

BILLING AND ACCOUNTING PROCEDURES
APPOINTMENT AND SERVICE-PROCEDURE SCHEDULING
CENSUS AND UTILIZATION OF SERVICES
INVENTORY AND PURCHASING FUNCTIONS
DEVELOPMENT OF COST-EFFECTIVENESS AND COST-BENEFIT RATIOS

Conclusion

The potential uses of information-formatted patient records just described assume a capability of the system to process large sets of natural language documents -- perhaps not one million patient records as suggested in Davis' criteria for medical information systems quoted at the beginning of this paper, but nevertheless an increase by several orders of magnitude over the small sample used in this experiment. We do not wish to minimize the differences between a laboratory demonstration of a device in the hands of its designers and a serviceable instrument for routine processing of documents in an operational setting. Much work remains to be done (our next objective)

to make the system robust and efficient for routine use and export.

What we believe we have demonstrated thus far is rather that a natural language processing system founded on general principles of linguistic analysis can carry out the analysis, codification, extraction and collation of medical information delivered to the computer in free-text form. Because the main parts of the system (the parsing program, English grammar, English transformations and formatting mechanism) are not special-purpose to the particular medical documents, diseases, or document-types processed thus far, or even to medical English itself (except in the final stages of the program), there is every reason to believe that the system can be applied to a large variety of data and to large amounts of data without major changes to the linguistic content of the processing programs. This work is based on results about language structure in relation to the information it carries and on regularities of language usage in science sublanguages. For this reason, as we increase the material within any one domain, we should encounter relatively few syntactic and semantic structures not already provided for in the system. A much larger effort would of course be required to extend the magnitude of the application, but this is a different problem from establishing that a technical solution for the structuring of information in natural language exists.

Acknowledgement

This research was supported in part by NIH Grant LM02616 from the National Library of Medicine, and in part by Research Grant DSI 77-24530 from the National Science Foundation, Division of Science Information. The programming of the normalization and retrieval was done by Guy Story; Elaine Marsh assisted in the computer-formatting of the discharge summaries.

References

(1) Davis, L.S., A System Approach to Medical Information. Methods of Information in Medicine 12: 1 (1973).

(2) Pratt, A.W., Medicine, Computers, and Linguistics. Advances in Biomedical Engineering, 97-140, Academic Press, New York, 1973.

(3) Sager, N., Information Structures in the Language of Science. American Association for the Advancement of Science (AAAS) Selected Symposium 3, The Many Faces of Information Science, 53-73, Edward C. Weiss, ed., Westview Press, Boulder, Colorado, 1977.

(4) Sager, N., Evaluation of Automated Natural Language Processing in the Further Development of Science Information Retrieval. String Program Reports (S.P.R.) No. 10, Linguistic String Project, New York University, 1976.

(5) Sager, N., Syntactic Analysis of Natural Language. Advances in Computers 8, 153-188, Academic Press, Inc., New York, 1967.

(6) Raze, C., The FAP Program for String Decomposition of Scientific Texts. S.P.R. No. 2, Linguistic String Project, New York University, 1967.

(7) Grishman, R., Sager, N., Raze, C., and B. Bookchin, The Linguistic String Parser. AFIPS Conference Proceedings 42, 427-434, AFIPS Press, Montvale, N.J., 1973.

(8) Anderson, B., Bross, I.D.J., and N. Sager, Grammatical Compression in Notes and Records: Analysis and Computation. American Journal of Computational Linguistics 2: 4 (1975).

(9) Hirschman, L., Grishman, R., and N. Sager, From Text to Structured Information: Automatic Processing of Medical Reports. AFIPS Conference Proceedings 45, 267-275, AFIPS Press, Montvale, N.J., 1976.

(10) Hirschman, L. and Grishman, R., Fact Retrieval from Natural Language Medical Records. IFIP World Conference Series on Medical Informatics (MEDINFO) 2, 247-251, North-Holland, Amsterdam, 1977.

(11) Lyman, M., Tick, L.J., and J. Korein, Comprehensive Health Care for Children: Bellevue Pediatric Project. New York State Journal of Medicine 68: 17, 2287 (1968).

(12) Lyman, M.S., Health Information System: One Coordinator of Health Care for Communities. N.Y. State J. Med. 72: 6, 698 (1972).

(13) Lyman, M., A Data Support System for Emergency Health Care. Proceedings of Third Illinois Conference on Medical Information Systems, 88-97, University of Illinois at Chicago Circle, Chicago, 1977.

(14) Korein, J., Tick, L.J., Woodbury, M.A., Cady, L.D., Goodgold, A.L., and C.T. Randt, Computer Processing of Medical Data by Variable-Field-Length Format, Parts I, II, and III. In J. of the Amer. Med. Assoc. 186, 132-138 (1963) and 196, 950-963 (1966).

(15) Korein, J., The Computerized Medical Record: The Variable-Field-Length Format System and Its Applications, Information Processing of Medical Records, 259-291, North-Holland, 1970.

(16) Hobbs, J., and R. Grishman, The Automatic Transformational Analysis of English Sentences: An Implementation. International Journal of Computer Mathematics 5: A, 267-283 (1976).

(17) Raze, C., The Parsing and Transformational Expansion of Coordinate Conjunction Strings. S.P.R. No. 11, Linguistic String Project, New York University, 1976.

(18) Hirschman, L., Grishman, R., and N. Sager, Grammatically-based Automatic Word Class Formation. Information Processing and Management, 11, 39-57 (1975).

(19) Sager, N., Natural Language Information Formatting: The Automatic Conversion of Texts to a Structured Data Base. Advances in Computers 17, Academic Press, New York, in press.

(20) Radel, E., Sickle Cell Disease. Pediatric Annals 3: 9, 31-50 (1974).

(21) Sager, N., and R. Grishman, The Restriction Language for Computer Grammars of Natural Language. Communications of the ACM, 18, 390-400 (1975).

(22) Sager, N. and M. Lyman, Computerized Language Processing: Implications for Health Care Evaluation. Medical Record News 49: 3, 20-30 (1978).

(23) Terranova, W., Cohen, M.L., and D.W. Fraser, 1974 Outbreak of Legionnaire's Disease Diagnosed in 1977. Lancet II: 122-4, July 15, 1978.

Dr. Naomi Sager
Linguistic String Project
New York University
251 Mercer Street
New York, N.Y. 10012