LSP 24

COMPUTER PROGRAMS FOR NATURAL LANGUAGE FILES*

Naomi Sager, Lynette Hirschman, Ralph Grishman, and Cynthia Insolio
New York University
New York, New York

ABSTRACT

As a supplement to conventional information services provided by libraries, many organizations build up technical files in particular subject areas. These collections may consist of excerpts from the published literature, supplemented by internal reports, or they may consist entirely of institutional files, as is the case in medical records. What distinguishes these collections is (1) the information is recorded in natural language, either in full prose or in the abbreviated style of notes; (2) the texts, whether in full prose or in the abbreviated style, deal with a circumscribed subject matter. This paper describes computer programs which convert such natural language files into a structured data base. The programs map successive sentences into a table-like structure, called an information format, which is based on the repeating syntactic and lexical patterns found in samples of the language material. In the table, words of similar informational standing are aligned in the same column. The result is an organization of the textual material which permits the information contained in the natural language documents to be searched automatically for specific facts or specific types of information, and makes possible computer preparation of statistical summaries of the different kinds of information in the file. Experiments in computerized formatting of medical records will be described as well as the methods by which the same technology (and many of the same programs) can be applied to other areas.

INTRODUCTION

Until very recently, natural language processing in information retrieval has meant primarily the scanning of large amounts of machine readable natural language material for the occurrence of particular word string combinations. Such methods, while they may be suitable for document retrieval from large data bases, do not provide a characterization of the contents of documents which is adequate for more complex informational operations; for example, checking for the presence of specific facts, or statistically summarizing specific kinds of information. Yet the need for computer programs which can perform such operations is growing. Not only are scientists often interested in extracting and compiling infor-

*For presentation at the 40th Annual Meeting of ASIS, Chicago, September 26-October 1, 1977.

mation from diverse data sources, but many institutions today have large natural language files which contain information that must be accessed and processed for a variety of purposes, often on a routine basis. Some of this material is already in machine readable form and it is likely that more will become so, especially if computer programs can take over some of the burden of processing the textual content.

The Linguistic String Project of New York University has been engaged in developing programs for processing the information in natural language scientific documents, both of the type found in the published literature and of the type found in internal reports, such as medical records. In this paper, we describe programs which transform the sentences in a file of documents in a particular subject area into an informationally equivalent table-like forms, called information formats. When the information is arranged in this way it is possible for a computer to retrieve specific facts, or produce statistical summaries of the information in the file. Later sections of the paper describe an experiment in which information formatting and retrieval programs were applied to a set of English-language radiology reports. First, the formatting programs mapped the sentences of the reports into the information format. Then, a second program obtained answers to specific questions about the content of particular reports by processing the formatted sentences. A statistical summary of certain types of reported facts was also computer-generated from the formatted sentences.

To illustrate the main features of information formats, Table 1 shows a simplified radiology format applied to two short X-ray report sentences. (We call each unit set off by periods a sentence even if, as is often the case in file material, the verb or some other required grammatical element is missing.) As Table 1 illustrates, an information format has rows and columns like a numerical table, but the table entries are words or phrases rather than numbers. Each column contains the words or phrases which carry a particular kind of information in the texts; for example in Table 1 the type of test is given by column 1, the location of the test by column 2, etc. The format columns TEST VERB FINDING correspond to the subject-verb-object relation in the original sentence, so that the major syntactic relations are preserved in formatting. Thus no information is lost, and the original sentences, or paraphrases of them, can be reconstructed from the format entries.

Simple as the examples in Table 1 are (more complicated formatted sentences are shown below), they raise the question: How is it possible for computer programs to convert the information in English sentences into table-like forms? From the outset it should be noted that special programs based on knowledge of the subject matter are not written for each application. Although we check our results with consultants who know the particular subfield of application, we do not ask such experts to propose, a priori, an

Table 1

Simple Radiology Format

	TEST			VERB	FINDING
	TEST	TEST-LOC	DATE	BE-SHOW	MED-FINDING
Br 1.1.1	films	chest	1-31-68	--	post radiation fibrosis
	films	chest	3-26-68	--	post radiation fibrosis
Br 0.1.8	scan	liver	1-29-69	was	normal

Text:

Br 1.1.1 1-31-68 and 3-28-68 chest films--post radiation fibrosis.
Br 0.1.8 Liver scan 1-29-69 was normal.

organization of information for that subfield, which we would then implement in the hope that it would suffice for analyzing subfield texts. First of all, there would be no guarantee that a computer could map sentences into the proposed structure. Secondly, experts might well disagree as to what are the most important relations to be included in the format. Lastly, as it happens, experts on the level needed for such work are simply not available. For all these reasons, but mainly because we attempted to find methods which would be generally applicable and would be assured to work, we base the processing on linguistic regularities which are inherent in the textual material. For each subfield of application we analyze a sample of the documents to find the syntactic and lexical regularities which can be formalized into information structures for the subfield. These textual regularities on the subfield level correspond in a direct way to the information carried by subfield texts. The information formats are a particularly useful way of representing these regularities for information retrieval. In the formats the parts of sentences which carry the same kind of information appear as entries in a single column, and the informational relations among entities in a particular row can be checked by reference to the column headings under which the entries lie.

In order to represent the information contained in the texts of a particular subfield, we must first establish relevant "units" of information and a form to represent these "units". Since we do not necessarily know at the outset the type of queries that the data base will be used to answer, the safest approach is to preserve as much of the structure of the original sentences as possible, while organizing the contents of each sentence into a meaningful pattern (or format). We do this by constructing (manually) a table or information format, whose columns correspond to specialized word classes of the subfield. (The construction of the format is the only manual step in the processing of the texts.) A series of programs then transfers a sentence into the table by putting each word into the table-column corresponding to its subclass. Where a word has no special subclass, it is left attached to the word that it modifies, so that all words are transferred to the format. This creates a table whose entries are the actual sentence words; since the order of columns is arranged to correspond to the occurrence of word classes in a "normalized" sentence, we can reconstruct the original sentence (up to paraphrase) from its representation in the table. As a final step in constructing the data base, it is necessary to "fill in" various pieces of information that are implicit from context, but must be made explicit in the data base; this is done automatically by a set of regularization procedures operating on the formatted sentences.

Each of these steps will now be described for an initial experiment on radiology reports [1, 2]. Our corpus consisted of the reports from the first 13 patients in a follow-up study of cancer patients. This material was furnished by Dr. Irwin D. J. Bross of Roswell Park Memorial Institute. Table 2 shows four sample sentences from this corpus.

CONSTRUCTING THE INFORMATION FORMAT

The relation between a word's pattern of occurrence in a text and its meaning enables us to establish the relevant informational units for a given subfield. Using techniques from structural linguistics, we build up word classes for the subfield by grouping together words which occur in identical or similar environments. (We call the language used in a subfield a sublanguage. We now have a program that will generate sublanguage word classes automatically [3].) For example, in the sentences of Table 2, if we group together the head nouns occurring as subject of show, we find both x-ray and film (ignoring suffixes). If we look at a larger subset of the corpus we find that x-ray and film share a number of other environments (together with the words scan, plate, mammograph); these words form a subclass in this sublanguage, based on similarity of occurrence patterns. They also share an element of meaning, which is related to their similarity of distribution. If we find the subclasses for the frequently occurring words in a given corpus, we will have found the informational "units" characteristic of this particular sublanguage. To the sublanguage classes, we add classes for certain informationally important words in the language as a whole, e.g., negation, time, and modals. From these combined classes we construct the format.

We build up the format by determining the possible co-occur-

Table 2

Sample Formatted Sentences

1. NOT DONE.
2. CHEST FILM 10-22 SHOWED NO CHANGE.
3. X-RAYS TAKEN 1-24 NEGATIVE FOR METASTASIS.
4. 3-2-65 CHEST FILM SHOWS CLOUDING ALONG LEFT THORAX AND PLEURAL THICKENING.

Left adjuncts are placed in ( ) above the head noun; right adjuncts below the head noun.
Material in square brackets [ ] filled in during normalization

	TEST					FINDING						
	NO-TEST	TESTN	TESTLOC	VERB-DONE	TESTDATE	NEG	BE-SHOW	CHANGE	STATUS	MED-FIND	REGION	
											POS	PT-BODY
1)	NOT	[X-RAY]	[CHEST]	DONE	[BETWEEN 4-22-64 and 6-10-64]							
2)		FILM	CHEST		10-22[-64]	NO	SHOWED	CHANGE [SINCE 8-5-64]				
3)		X-RAYS		TAKEN	1-24[-65]				NEGATIVE	(FOR) METASTASIS		
4)		FILM	CHEST		3-2-65		SHOWS			CLOUDING	ALONG	(LEFT) THORAX
		FILM	CHEST		3-2-65		SHOWS			THICKENING		PLEURAL

rences of the word classes. The format must contain a column for each type of sublanguage information; we build up the format from actual patterns of co-occurrence found in a sample subset of the corpus, so that its final structure corresponds to the structure of the sentences in the corpus, "normalized" (rearranged) in ways that do not alter their informational content.

TRANSFERRING THE INFORMATION INTO THE FORMAT

The formatting program operates in three steps. First the sentence is parsed with the Linguistic String Parser English grammar, to provide the syntactic structure of the sentence (e.g., modifier relations, subject-verb-object relations, etc.). Next the sentence is regularized by applying certain information-preserving English transformations. The most important of these are the conjunction expansion (expanding two conjoined elements of an assertion into two complete conjoined assertions -- see example 4 in the table), and the relative clause transformation, which replaces a relative pronoun by its antecedent, e.g., *fracture which may be pathological* → *fracture such that fracture may be pathological*. The parsed, regularized sentences are then placed, piece by piece, into the appropriate format slots by special formatting transformations. This last step involves rules specific to the sublanguage; the first two steps (parsing and English transformations) are general and would be used in the formatting of any sublanguage.

Because of the way that the format is constructed, there is a close correspondence between word class membership and format column. However, for the cases where there is not a one-to-one correspondence between word class and format column, we must have syntactic information to determine which format slot a word should be placed in. For example there is only one slot (TESTN) for words of the class N(oun)TEST (*x-ray*, *film*, etc.). Therefore we need only the word class membership to know where in the format to put the word *x-ray*. However, if we have a negative, there are two format slots for negatives: NO-TEST (negating the existence of a test) and NEG (negating the finding). We must know what the negative modifies to determine which of these slots it should go into:

1) *no metastasis:* *no* goes into NEG because it modifies a N(oun)CONDITION.

2) *no x-rays:* *no* goes into NO-TEST because it modifies an NTEST.

The formatting transformations move the words of a sentence into the appropriate format columns; they make use of information about the syntactic environment of a word where necessary to determine the appropriate format slot. Table 2 shows the formatted sentences produced by this three-step process of parsing, application of English transformations, and application of formatting transformations. (Note that the words in square brackets [ ] are filled in *after* formatting, during the next step of processing.)

FILLING IN IMPLICIT INFORMATION

After the sentences have been automatically converted into tabular form by the formatting procedure, the "implicit" information available from context must be filled in to complete the data base.

One type of filling in has already been done: the English transformations (step two in formatting) have regularized conjunctional constructions, creating conjoined full assertions (see example 4 in the table). In addition, we must find antecedents for pronouns and other types of referential expressions; and we must fill in the information that was omitted because it was clear from context.

The following example illustrates how the antecedent for a referential expression is found. In the phrase 11-5-62 no change from previous chest x-ray, the phrase from previous chest x-ray refers to an earlier x-ray. Previous is the referential time expression; we want to find its antecedent, namely an exact date which identifies the chest x-ray being referred to. To do this, we search backwards in the formatted sentences to find the first "match": a sentence which has TESTN = x-ray or a synonym, TESTLOC = chest, and TESTDATE earlier than 11-5-62. If we find such a line in the format, we have found the antecedent for previous chest x-ray, and can regularize it in the format by adding the specific date (e.g., previous chest x-ray of 10-22-62). If no antecedent is found for a referential expression, then an error message is printed out.

Certain pieces of information are omitted in the original text because they can be found in a preceding sentence, which sets the topic for the following sentences. For example given the "textlet" Chest films 10-15 infiltrate clear. Lung scarring still present., the missing TEST information can be filled in for the second sentence by noting that it continues the topic (chest films 10-15) of the first sentence: chest film 10-15 lung scarring still present. (The "filling in" is actually done on the formatted sentences; the full text is used here merely for illustrative purposes.)

In other cases information is missing and there is no appropriate preceding sentence which sets the topic, e.g., when an entire report consists of the phrase not done. In such cases we fill in certain "default" values that are characteristic of the subfield. For example in radiology reports, if a test name is omitted, and there is no preceding sentence which sets the topic, then the default value for TESTN is x-ray: Not done → X-ray not done.

Once these automatic regularization procedures have operated on the formatted sentences, the data base is ready for use. Table 2 shows the data base for four sample sentences after this final stage of processing.

INFORMATION RETRIEVAL FROM FORMATS

What types of processing would we want to do using the medical records as data? As records are received, validity checks should be performed, flagging internal inconsistencies in the data which indicate errors. When specific information is needed on a patient, a fact retrieval system should be able to provide it. If the records are being accumulated for research purposes, statistical summaries of the data will be needed. The normalized formats present the information contained in the medical records in a form which makes individual components of the information readily accessible. This makes it possible to implement the record processing functions just mentioned with relatively simple, straightforward programs.

The principal feature of the formats which facilitates this processing is that a specific type of information can be directly located by looking under the corresponding column heading. An additional feature is that all entries in a given column have the same directionality. For example, any entry in the CHANGE column is a word which indicates that a change has taken place. If a word occurs in a report which indicates the absence of change (such as unchanged or same), it will be factored into an entry in the NEG column of FINDING and the word change in the CHANGE column. As a result, some questions about a record can be answered by simply testing for the presence or absence of an entry in a particular column. More detailed questions, of course, can still test the value of a particular entry.

To verify that the formats are a felicitous structure for processing the records, we have implemented sample functions of each of the three types cited above--validity checking, fact retrieval, and statistical summary. A good deal of checking is done as a by-product of normalization and other processing. One type of checking verifies test dates. The reports we are processing were prepared four times a year for each patient and covered tests performed since the last report. Each test date in a report is checked to verify that it is not later than the report date (an error) or earlier than the previous report date (a probable error). Any reference to a prior test (e.g., no change since 10-13-63) is checked to verify that a test was performed on that date.

A number of questions, such as whether any X-ray was taken in a given period and when it was done, can be answered directly by examining one or two format columns. We shall describe here the procedure for answering a somewhat more complicated question, namely, whether a report contains any abnormal or suspicious findings. A report contains such findings if any format line of the report contains such findings. Thus the problem reduces to the slightly simpler question, does a format line contain any abnormal or suspicious findings? This is answered by the following procedure:

A. If NO-TEST column is filled (no test was done), answer is no; else

B. If STATUS column (a column for normal findings) is filled and NEG column is empty (e.g. normal), answer is no; else

C. If STATUS column is filled and NEG column is filled (e.g., not normal), answer is yes; else

D. If CHANGE column is filled (and value of column is not a word indicating improvement) and NEG column is empty (unnegated change), answer is yes; else

E. If CHANGE column is filled and NEG column is filled (no change), we locate the format lines reporting the earlier test against which the current test is being compared and apply the procedure to those format lines. The answer obtained for those earlier lines is the appropriate answer for the current format line, since no change is reported. Else

F. If MED-FIND column is filled and NEG column is empty, answer is yes.

G. Otherwise answer is no.

This procedure gives the following results for the four formatted sentences of Table 2.

1. no (test A)
2. answer obtained for formats for test of 8-5-64 (test E)
3. no (test B)
4. yes (test F)

Table 3 shows some statistics which have been obtained automatically by processing the formatted data base. Natural language processing was essential in calculating the recurrence column: the date of recurrence was defined by medical consultants to be the date of the first report of metastases; the logic for obtaining this information from the format is similar to that in the question-answering procedure described above. Such statistics are routinely required in research projects, but they normally involve a review of all the reports by doctors or data clerks.

GENERALITY OF THE METHODS

The programs used to analyze and format radiology reports were not made especially for the radiology material. A comprehensive computer grammar of English had previously been developed [4]. It was recently extended to cover the fragmented sentence forms found in notes and records, a change which increased the size of the gram-

Table 3

PATIENT	SURGERY- DATE	REPORTS	POSITIVE- RECURRENCE	LOCATION	TIME-AFTER - SURGERY
08P03	04-17-67	8	YES	RIBS FEMORAL PELVIS VERTEBRAE SKULL	1 YEAR 9 MONTHS 11 DAYS
08P200	09-09-67	15	NONE	NIL	4 YEARS 7 MONTHS 3 DAYS
09P003	12-04-61	22	NONE	NIL	6 YEARS 11 MONTHS 4 DAYS
10C001	11-09-61	27	NONE	NIL	7 YEARS 2 MONTHS 4 DAYS
10C015	01-31-64	20	NONE	NIL	6 YEARS 0 MONTHS 15 DAYS
10C020	12-02-64	17	NONE	NIL	4 YEARS 7 MONTHS 26 DAYS
10C019	07-28-64	8	YES	PULMONARY	1 YEARS 4 MONTHS 1 DAY
10C021	04-13-65	13	NONE	NIL	4 YEARS 4 MONTHS 5 DAYS
10C024	08-26-65	15	NONE	NIL	3 YEARS 10 MONTHS 4 DAYS
10C026	12-08-65	10	NONE	NIL	2 YEARS 9 MONTHS 16 DAYS
10C027	03-17-66	11	NONE	NIL	3 YEARS 2 MONTHS 2 DAYS
08P008	02-20-62	23	NONE	NIL	7 YEARS 11 MONTHS 13 DAYS
08P009	03-02-62	24	NONE	NIL	9 YEARS 1 MONTHS 4 DAYS

TOTALS

TOTAL NUMBER OF PATIENTS 13
AVERAGE NUMBER OF REPORTS PER PATIENT 16
AVERAGE TIME BETWEEN VISITS 3 MONTHS
NUMBER OF PATIENTS WITH RECURRENCE 2
THEIR AVERAGE TIME - SURGERY TO RECURRENCE 18 MONTHS
NUMBER OF PATIENTS WITHOUT RECURRENCE 11
THEIR AVERAGE TIME TO LAST XRAY 65 MONTHS

mar by only a few percent [5]. The operative parsing and transformation programs of the Linguistic String Project have also been designed for a broad coverage of natural language texts [6,7]. The formatting transformations are general in form, but, of course, must be tailored to the word classes and information structure which are particular to the field of application. The needed word classes are obtained from a sample of grammatically analyzed subfield texts, by grouping together words which occur in the same grammatical relation to particular other words, (e.g. nouns occurring as the subject of particular verbs), as demonstrated in Section 2, above. This procedure was implemented in a clustering program, which was found to generate the desired semantic classes for a subfield of pharmacology [3]. It has also been found that information structures suitable for constructing information formats are a general feature of scientific and technical writing [8]. For example, formats which apply to journal articles in a research field of pharmacology have been obtained [9].

Encouraged by the success of the radiology formatting experiment, we are now applying the analysis and formatting programs to a broader and more complex sector of medical records, a corpus of pediatric patient records which includes both hospital discharge summaries and reports of clinic visits. A simplified verison of a portion of this format is shown in Table 4. Many sentences in medical records, like those in Table 4, contain information about TREATMENT or about the PATIENT STATE, or about both. In the TREATMENT part of the format, INST contains references to the medical institution or instutional representative (M.D., clinic); V-TREAT contains verbs whose subject are INST words. In the PATIENT STATE part, V-PATIENT contains verbs and connectives which link patient and symptom syntactically in the sentence. TIME is a very important informational category for these texts, with many subparts not shown here. Aside from its complexity, the unedited nature of this material also presents problems, in the inconsistent use of punctuation, the heavy use of sentence fragments, abbreviations, etc. However, the same properties of repetition and regularity that characterize the radiology reports, and in a different way the formatted research articles in pharmacology, are also present in this material. Given the generality of the methods and the success of the programs to date, it appears likely that for any collection of texts where the subject is limited and usage is regular, the information in the language material could be automatically converted to information formats.

ACKNOWLEDGEMENT

This investigation was supported in part by Research Grant 1-R01-LM-02616 from the National Library of Medicine, National Institutes of Health, D.H.E.W., in part by Research Grant No. SIS-75-22945 from the National Science Foundation, Division of Science Information, and in part by contract N00014-75-C-0571 with the Office of Naval Research.

Table 4

Partial Medical-Records Format*

PATIENT FIRST HAD SICKLE CELL ANEMIA DIAGNOSED AT AGE 2 YEARS WHEN HE COMPLAINED OF LEG PAIN. HE WAS ASYMPTOMATIC UNTIL AGE 5 WHEN HE WAS ADMITTED TO BELLEVUE HOSPITAL WITH CHEST PAINS. HE WAS HOSPITALIZED FOR A MONTH AND RELEASED

	CONJ	PATIENT	TREATMENT		PATIENT STATE					TIME				
			INST	V-TREAT	V-PATIENT	BODY PART	NORM	SIGN/ SYMPT	DIAGNOSIS	P	#	UNIT	P	REF.PT
1		PATIENT		FIRST HAD DIAGNOSED					SICKLE CALL ANEMIA				AT	AGE 2 YEARS
2	WHEN	HE			COMPLAINED OF	LEG		PAIN						(AGE 2 YEARS)
3		HE			WAS		ASYMPTO MATIC						UNTIL	AGE 5
4	WHEN	HE	BELLEVUE HOSPITAL	WAS ADMITTED	WITH+	CHEST		PAINS						(AGE 5)
5		HE		WAS HOS- PITALIZED						FOR	A	MONTH		
6	AND	(HE)		(WAS) RELEASED										

*The words in parentheses are filled in by the English transformational subprogram.

+The word <u>with</u> is transformationally related to the verb <u>have</u>: <u>patient has symptom</u> → <u>patient with symptom</u>.

REFERENCES

1. Hirschman, L., Grishman, R., Sager, N. From Text to Structured Information: Automatic Processing of Medical Reports. Proceedings of the 1976 National Computer Conference. *AFIPS Conference Proceedings*, vol. 45, 267-275, AFIPS Press, Montvale, N.J., 1976.

2. Hirschman, L., and R. Grishman. Fact Retrieval from Natural Language Medical Records. To appear in the *Proceedings of Medinfo 1977*.

3. Hirschman, L., R. Grishman and N. Sager. Grammatically-based Automatic Word Class Formation. *Information Processing and Management*, vol. 11, 39-57, 1975.

4. Sager, N. A Computer String Grammar of English. *String Program Reports (S.P.R.)* No. 4, Linguistic String Porject, New York University, 1968.

5. Anderson, B., I.D.J. Bross, and N. Sager. Grammatical Compression in Notes and Records: Analysis and Computation. Paper delivered at the 13th Annual Meeting of the Association of Computational Linguistics, Boston, Nov. 1, 1975, *American Journal of Computational Linguistics*, vol. 2, no. 4, 1975. (A Roswell Memorial Institute paper).

6. Grishman, R., N. Sager, C. Raze, and Bookchin, B. The Linguistic String Parser. *Proceedings of the 1973 Computer Conference*, 427-434, AFIPS Press, 1973.

7. Hobbs, J. and Grishman, R. The Automatic Transformational Analysis of English Sentences: An Implementation. *International Journal of Computer Mathematics*, 1976, Section A, vol. 5, pp. 267-283.

8. Sager, N. Information Structures in the Language of Science. *American Association for the Advancement of Science Symposium Volume*, February 1977, in press.

9. Sager, N. Syntactic Formatting of Scientific Information. *Proceedings of the 1972 Fall Joint Computer Conference*, AFIPS Conference Proceedings, vol. 41, 791-800, AFIPS Press, Montvale, N.J., 1972.