MEDINFO 77, Shires/Wolf, editors

FACT RETRIEVAL FROM NATURAL LANGUAGE MEDICAL RECORDS

Lynette HIRSCHMAN and Ralph GRISHMAN

New York University,
New York, New York, USA

This paper describes a procedure for fact retrieval from natural language texts. First the natural language material is processed syntactically and mapped into a table (information format), where each distinct type of information is placed into a separate format column. A set of normalization procedures fills in certain missing pieces of information from context. Finally each query is translated into a procedure which checks the columns of the information format and produces the desired answer. Nine implemented queries are listed for an automatically formatted data base of radiology reports for 13 patients, together with some sample answers for one patient's reports.

1. INTRODUCTION

The Linguistic String Project of New York University is engaged in long-term research on the problem of retrieving information from natural language texts. As a basis for this work, we have developed a number of computational linguistic tools, including a parser, a grammar-writing language, a broad-coverage grammar for the string segmentation of English sentences, and sets of syntactic transformations. Equipped with these tools, we have begun in the past few years to address specific retrieval problems. We have been concentrating on retrieval from medical records, for several reasons: there is a clear need for the processing of medical records in text form; the underlying information structures appear to be simpler than for some other types of texts we have examined; and large quantities of medical records are being captured in machine-readable form.

Our basic approach involves automatically structuring a text into an essentially tabular form, with each column of the table containing one type of information present in the text (such as patient, date, procedure performed, body part involved, etc.). We refer to these structures as information formats. It is our intention to perform all retrieval operations upon such formatted texts.

In a recent paper [1] we reported on a procedure for formatting texts, and described its application to a set of radiology reports. It remains to be shown that the formats we produced constitute an appropriate structuring of the text for data retrieval. To this end, we shall present in this paper a number of requests for information in the radiology reports, and show how they can be answered by straightforward procedures applied to the formatted text.

2. FORMATTING

In this section we shall briefly summarize our method of text formatting; readers are referred to [1] for a more detailed description. Two procedures are involved: the first, a preliminary manual procedure to develop the appropriate format structure (column headings) for the texts in a subfield; the second, an automated procedure to map sentences in the subfield into formatted entries. Parts a and b of table 1 show three sentences from our corpus of radiology reports, and their formatted counterparts.

The procedure for developing the format involves two steps: first, determining the word classes of the subfield; second, actually constructing the format. The word classes are identified by grouping together words which occur in the same syntactic environments. For example, if we delete noun modifiers and ignore affixes, sentences 2 and 3 in table 1 contain instances of the pattern

$$\left\{\begin{array}{l}\text{x-rays}\\ \text{film}\end{array}\right\}\ [\text{show}]\ \text{change}$$

with the verb (show) omitted in sentence 3. Because x-ray and film share several environments, including this one, they are grouped together (along with several other words) into a word class of test nouns.

Once the word classes have been established, the format can be constructed, based on the principle that equivalent pieces of information in different sentences should go into the same format columns. These equivalent pieces of information are identified using the word classes. We begin by taking several sentences, writing them down on successive lines, and lining up words of the same class which occupy the same structural

position in the sentence. (In judging sameness of structural position, we allow for paraphrastic variations in syntax, so that, for example, the word chest will be aligned in chest X-ray and X-ray of chest.) The columns which result constitute an initial version of the format. By taking a larger sample of sentences and continuing this operation, the format is gradually enlarged and refined.

The construction of the word classes and format for this text was for the most part a manual operation. We have, however, written a computer procedure for word class formation [2] and we believe that the entire format construction procedure can be largely automated.

The mapping of text sentences into the format involves three processing steps, and is fully automated. First, a linguistic string analysis (parse) of the sentence is obtained using the Linguistic String Project parser [3] and English grammar. (To handle the medical report narrative, the grammar was enlarged to accept

TABLE 1

Three sample formatted sentences and normalized formatted sentences, taken from 3 separate reports of a single patient (with intervening material omitted).

A. ORIGINAL SENTENCES:
1. Pa chest negative (9-1-64).
2. X-rays of spine show extreme arthritic change but no definite evidence of tumor.
3. 2-25-69 chest film unchanged.

	CONN	FORMAT																
	CONJ	DATA																
		TEST						FINDING									STATUS	MED-FIND
		NO-TEST	VIEW	TESTN	LOC	DATE	NEG	VERB-EL		CHANGE	TIME-PERIOD							
								BE-SHOW	IN-DIC		WHEN	VIEW	TESTN	LOC	DATE			
B. UNNORMALIZED FORMATS																		
1.				pa		chest	9-1-64										negative	
2.					X-rays	(of) spine			show	(extreme) change								arthritic (of) tumor
	but				X-rays	(of) spine		no		(definite) evidence								
3.					film	chest	2-25-69	neg-prefix		changed								
C. NORMALIZED FORMATS																		
1.				pa	X-ray	chest	9-1-64										negative	
2.					X-rays	(of) spine	3-2-65		show	(extreme) change								arthritic (of) tumor
	but				X-rays	(of) spine	3-2-65	no		(definite) evidence								
3.					film	chest	2-25-69	neg-prefix		changed			film	chest	8-6-68			

Left adjuncts are placed in ( ) above the main entry.
NOTE: This is an abbreviated format; a number of empty columns are not shown. The actual format includes columns for region of the finding, and indefinite markers on both FINDING and CONN(ective) for expressions like possible or may in possible lesions or may be related to.

sentence fragments as well as full sentences.) Second, a number of English syntactic transformations are applied to regularize certain constructions, such as conjunctions and relative clauses. This processing is done by the transformational component of the LSP parser [4]. All conjoined structures are expanded to conjoined full assertions; for example, sentence 2 in table 1 is expanded to X-rays of spine show extreme arthritic change but X-rays of spine show no definite evidence of tumor. These first two stages of processing are relatively constant from one type of report to another.

The third step is the mapping of elements from the regularized parse tree into format slots. When a format has been defined, rules are prepared which specify, as a function of a word's subfield word class and position in the parse tree, the format slot into which the word should be placed. These rules are implemented as transformations interpreted by the LSP parser.

3. OBJECTIVE OF OUR PRESENT WORK

Using the procedures summarized above, we successfully formatted 95% of a corpus of 206 radiology reports on 13 patients, containing a total of 248 sentences. We then set out to study the adequacy of this formatted text for information retrieval.

The reports we had formatted were part of a follow-up study of patients who had undergone surgery for breast cancer. We obtained a list of questions which, as part of this study, had been answered manually from these reports. We wanted to show that these questions could be answered by straightforward procedures applied to our automatically formatted text.

4. NORMALIZATION

A great deal of information in the reports, and consequently in the formats, is not stated explicitly, but can be reconstructed from context. The retrieval operations can be expressed quite simply provided that we first reconstruct this information (normalize the formats). Each normalized format contains complete TEST information (test date, test location, test name), and the FINDING, if any, including the TEST information for any text used as a comparison (capitalized names are headings in table 1).

Before any information is filled in, the normalization procedure, in a few cases, will collapse two consecutive formats into a single format; we do this where information belonging in a single format has been split syntactically between two separate assertions (and therefore maps into two separate but incomplete formats).

The following pair of sentential fragments illustrates how missing information is reconstructed:

> Chest films 10-15 rll infiltrate clear.
> Lul scarring still present.

The second sentence contains no TEST information at all; however the first sentence sets the topic: chest films 10-15. We assume continuity of topic: information concerning the topic of a discourse (here a report) remains unchanged in subsequent sentences until explicitly superceded by a new piece of information; on this basis we fill in the missing TEST information in the format of the second sentence by copying it from the format of the first sentence.

In sentence 2 of table 1b, the DATE is missing. In table 1c it has been filled in (normalized) by copying the DATE from the preceding format (not shown in table 1). Before any piece of information is copied, a procedure checks that the information to be copied has not been superceded by new information (i.e. no change of topic has taken place).

In cases where there is no preceding format that gives the TEST information, we fill in certain "default" values for this corpus. Our corpus consists of X-ray follow-up reports on patients with breast cancer. The "default" test for this material is X-ray, and the "default" location is chest. For example, in sentence 1, the normalization procedure fills in X-ray as the TESTN (test name). The "default" values were obtained from an examination of word co-occurrence patterns and frequencies in the corpus. These co-occurrence patterns are checked before any information is filled in (e.g. can the value for LOC (location of test) co-occur with the word in TESTN?). In the case of a missing date, the "default" value is taken as a range, from the date of the preceding report to the date of the current report.

Resolution of time reference requires the most complicated normalization procedures. When a test is said to be unchanged in relation to a previous state, the TESTN, LOC and DATE of the earlier test are reconstructed. The DATE of the earlier test is found by a routine that searches for the preceding comparable test in a patient's reports. Sentence 3 refers to a previous state; in table 1c, the TESTN, LOC and DATE columns under TIME-PERIOD (the earlier test) have been filled in, after finding the earlier test referred to (not shown in the table).

5. QUERIES

Once the formats have been normalized, the conversion of a query to a retrieval routine is straightforward. Each different type of information in the format has its own column; in addition, negatives have been factored out, so that the words in a given column all have the same "directionality". For example, each word appearing in the CHANGE column indicates the existence of a change; the word unchanged is factored into a neg-prefix entry in the NEG column, and changed in the CHANGE column (see sentence 3 in table 1b). In this way we know that when a word occurs in the CHANGE column of the format and the NEG column is empty, a change has taken place; if an entry appears in the NEG column, then the change is negated (i.e. no change has taken place). For many queries it is therefore sufficient to check for the presence or absence of words in several columns. For more detailed information the actual word in the column may be checked, or retrieved as the answer.

The procedure for the query Was a chest X-ray done? has the following steps: we search the formats of the report for a format where all of A), B), and C) are true.

(A) the NO-TEST column is empty (a word in NO-TEST negates the existence of a test)
(B) LOC of TEST = chest
(C) TESTN of TEST = x-ray(s) or film(s).

If such a format is found, then the answer to the question is YES; if none of the formats in the report meets this condition, then the answer is NO. Treating each of the sentences in table 1 as a separate report, this procedure, when applied to the normalized formats in 1c yields the following answers: YES for 1, NO for 2 (criterion B fails for both formats in 2) and YES for 3.

A more complicated example is the procedure for the query: were the findings (of a report) negative? A finding is not negative if there is an unnegated medical finding or change, or if the test is not normal. If there is no change, then we must consult the earlier test given as a comparison. In other cases, the finding is negative. This translates into conditions A)-D).

(A) If MED-FIND column is not empty and NEG column is empty (unnegated medical finding) then the answer is NOT NEGATIVE.
(B) If CHANGE column is not empty and NEG column is empty (unnegated change) then the answer is NOT NEGATIVE.
(C) If NEG column is not empty and STATUS column (a column for normal findings) is not empty (e.g. not normal) then the answer is NOT NEGATIVE.
(D) If the NEG column is not empty, and the CHANGE column is not empty (e.g. no change) then we must find the format that the present format is compared to, and check whether it is negative; this answer gives us the answer for the present format, since there has been no change.

If we get an answer NOT NEGATIVE from one of these tests, we are finished. Otherwise we continue to examine the formats of the report in succession until we get an answer NOT NEGATIVE or until there are no more formats in the report, in which case the answer for the report is NEGATIVE.

Going back to the sentences in table 1, the procedure gives the following answers for the normalized formats in part c (again treating each sentence as a separate format):

Sentence 1 does not meet any of criteria A) to D); answer is NEGATIVE.
Sentence 2: 2a meets criterion B), so the answer is NOT NEGATIVE, and we do not need to look any further.
Sentence 3 meets criterion D), so we must find the answer for the test of 8-6-68 (not shown); the answer for that test is NOT NEGATIVE, therefore the answer for sentence 3 is NOT NEGATIVE.

The procedure outlined above must also take into account the different informational status of formats, which depends on how a format is "connected" to the preceding format. Two formats conjoined by a coordinate conjunction, both have the same informational status; however a format connected to another format by the connective compatible with (e.g. X-ray reveals destruction of right clavicle compatible with progressive osteomyelitis) does not assert the connected material (progressive osteomyelitis) as a finding; it is a possibility. The procedure for this query does not look at any format preceded by such a non-factive connective. (However this information is used by the procedure for the question: Is the finding positive or suspicious?)

6. RESULTS

We implemented a set of nine queries to test the adequacy of our formatting and retrieval procedures. The questions fall into two types: editing questions (concerning the presence and consistency of information in the reports) and "content" questions regarding the first occurrence of metastasis in the patient. We verified our machine generated answers by comparing them to a set of manually prepared answers. The set of questions and the manually tabulated answers were provided to us by Irwin Bross. The questions that we answered are:

Editing:
1E. Was the required chest x-ray done?
2E. Were other procedures carried out?
3E. Did the doctor report visible evidence?
4E. If examinations were performed, were the findings reported?
5E. Is the date recorded?
Content:
1C. Were the findings negative?
2C. Is there some question about the findings?
3C. When was the first metastasis reported?
4C. Where was it?

To illustrate the question-answering, a set of five consecutive patient records from our corpus is given below, together with the answers produced for questions 3C and 4C.

Report Date
10-26-64: None this visit.
03-05-65: X-rays were negative.
06-07-65: No x-rays taken on this examination.
10-11-65: X-rays taken revealed multiple pulmonary nodules with what appeared to be an enlarged inferior hilar node.
01-03-66: X-rays 11-29-65 multiple pulmonary metastasis.

Answers to questions 3C and 4C:
Last negative finding:
BETWEEN 10-27-64 AND 3-5-65
First subsequent non-negative finding:
BETWEEN 6-8-65 AND 10-11-65
First positive finding for metastasis:
11-29-65
Location of metastasis:
PULMONARY

(As noted above, date ranges occur as a result of normalization when no explicit test date is given.)

7. CONCLUSION

We have created a formatted data base of medical reports and shown its adequacy for fact retrieval. In constructing the formats and formatting the text, we have endeavored to develop procedures which are generalizable to other types of texts. This approach to retrieval from natural language text should be useful in several areas. For example, in medical research studies, these techniques should facilitate the gathering of statistics from large collections of medical records. We are currently preparing formats for hospital discharge summaries; using more complex processing routines, we plan to automate in part the screening of hospitalizations for peer review.

In order to use this approach for on-line retrieval of specific patient information, a user interface will be required. We are developing a system for translating natural language queries, such as the questions presented above, into data base retrieval procedures.

ACKNOWLEDGEMENT

We would like to acknowledge the contribution of Dr. Barbara Anderson and Dr. Irwin D. J. Bross in evolving an approach to coding natural language medical notes, and for the basic data used in the tests reported here. This investigation was supported in part by Research Grant 1-R01-LM-02616-01 from the National Library of Medicine, National Institutes of Health, DHEW; in part by Public Health Service Research Grant No. CA-11531 from the National Cancer Institute; in part by Research Grant No. SIS75-22945 from the National Science Foundation, Division of Science Information; and in part by contract N00014-75-C-0571 with the Office of Naval Research. The programs for normalizing the formats and answering the queries were prepared by Jim Litsas.

REFERENCES

[1] L. Hirschman, R. Grishman, and N. Sager, From text to structured information -- Automatic processing of medical reports, Proc. 1976 National Comp. Conf., AFIPS Press, 1976, 267-275.

[2] L. Hirschman, R. Grishman, and N. Sager, Grammatically-based automatic word class formation, Information Processing and Management, vol. 11, 1975, 39-57.

[3] R. Grishman, N. Sager, C. Raze, and B. Bookchin, The linguistic string parser, Proc. 1973 National Comp. Conf., AFIPS Press, 1973, 427-434.

[4] J. Hobbs and R. Grishman, The automatic transformational analysis of English sentences: an implementation, Intl. J. Computer Math., in press.