Natural Language: Sager et al

TRANSFORMING MEDICAL RECORDS INTO A STRUCTURED DATA BASE

Naomi Sager, Lynette Hirschman, Ralph Grishman and Cynthia Insolio

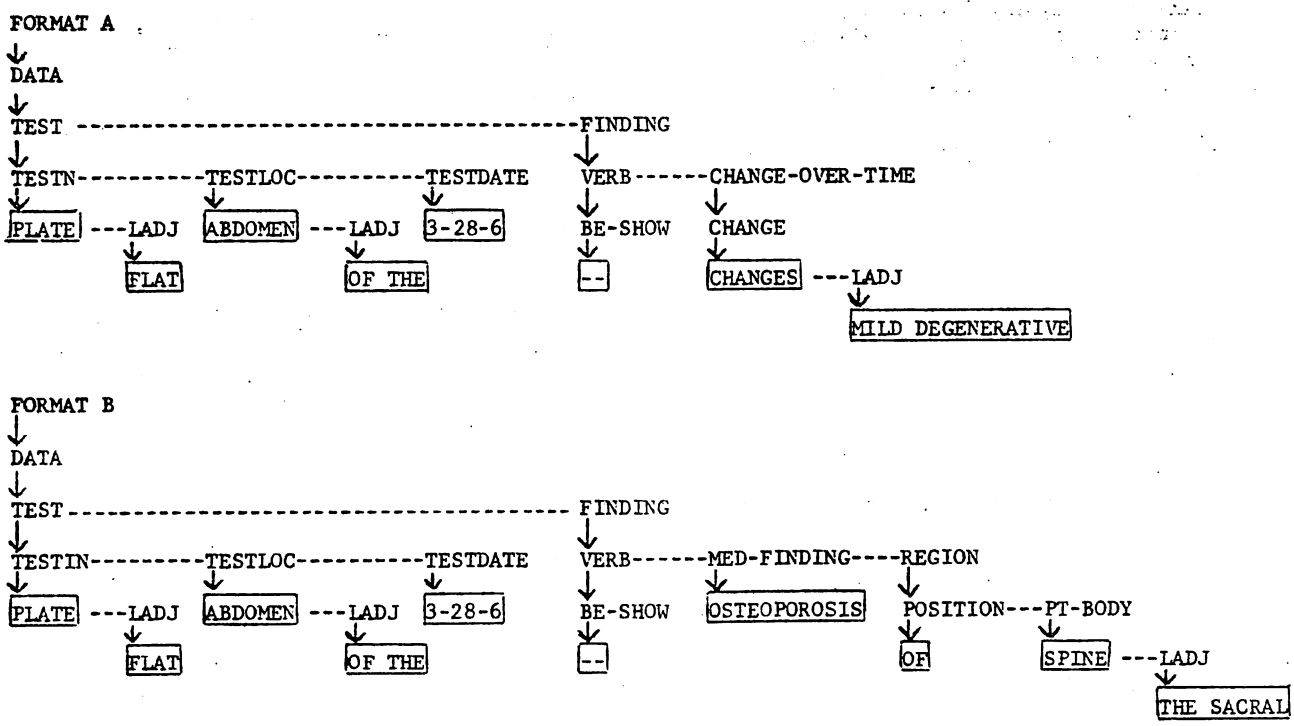
New York University
Linguistic String Project
251 Mercer Street
New York, N.Y. 10012

The N.Y.U. Linguistic String Project (LSP) is presently engaged in applying its programs for natural language processing to medical records. The programs transform the free narrative input into a structured data base suitable for automatic information processing, such as question answering, editing of records, or statistical summaries of the data. In order to determine the appropriate structures for a given type of material we first perform a manual linguistic analysis on a sample of the texts prior to processing. From this we obtain a set of word classes and a tabular form (called an information format) for this type of material. We then apply the series of processing programs to the sentences of the texts. Each sentence is parsed with the Linguistic String Parser English grammar in order to obtain its grammatical structure; then certain standard English transformations are applied to regularize the grammatical form of the sentence. Finally a set of "formatting transformations" map the words of the sentence into the slots of the information format, or table, for this material in such a way that the sentence is reconstructible up to paraphrase from its representation in the table.

The first experiment* in automatic information formatting involved the processing of a set of follow-up X-ray reports on patients who had had surgery for breast cancer. The corpus consisted of 159 consecutive (i.e. not specially selected) reports on 11 patients. It contained a total of 188 sentential units ranging in complexity from short fragments (e.g. x-rays negative) to long sentences (e.g. Reexamination shows some scarring and thickening over the right apex which is perhaps slightly more evident than it was before but nothing is seen that is typical of tumor involvement). To each sentence of the corpus we applied the formatting program, which parsed the sentence, performed conjunction expansion and certain other English transformations on it, and then mapped this structure into the format. This program successfully formatted 176 of the original 188 sentences (94 percent).

An output for one of the X-ray entries is shown in the following figure.

3-28-68 FLAT PLATE OF ABDOMEN -- MILD DEGENERATIVE CHANGES AND OSTEOPOROSIS OF THE SACRAL SPINE



* Lynette Hirschman, Ralph Grishman and Naomi Sager, "From text to structured information--automatic processing of medical reports," Proc. of the National Computer Conference, 1976.

As can be seen in the output, a formatted X-ray entry contains a TEST section, which includes slots for the test name, body location, and date, and a FINDING section which includes slots for a verb (or dash) followed by slots for information about change and specific medical findings. [Headings with no word values in a particular occurrence are not printed.] It should be emphasized that while the format slots have a semantic interpretation indicated by the choice of headings (e.g. TEST, FINDING), the format structure is arrived at by syntactic programs of general applicability. The semantic content specific to the field of application comes into play mainly in the third stage of processing (mapping the words of a sentence into format slots). Words of a sentence are mapped into format slots based on their membership in sublanguage word classes.*

Once the medical narrative has been mapped into information formats, this data base can be used in a variety of ways. We are currently at work on a program to extract various medical statistics from the x-ray data base (e.g. number of patients with recurrence of metastasis; time from operation to time of first suspected recurrence of metastasis; location of new metastasis, etc.). An AI project of the Computer Science Department at the Courant Institute of Mathematical Sciences is preparing to use this data base with a natural language front end to process questions and answer them with information from the data base.

The LSP is now working on a second experiment in formatting medical records. The natural language material for this experiment consists of hospital discharge summaries, which are texts of roughly 1-3 pages in length giving the background, reason for admission, physical examination, laboratory data, narrative of the course in the hospital, diagnosis and discharge status of a patient's hospital stay. While this material contains a much greater variety of information and is more varied in style and content than the X-ray records, it still has the restricted and repetitive features of a sublanguage; this enables us to formulate the word subclasses and syntactic regularities that give us the structure of the information formats for these documents. If this more ambitious undertaking is successful, it should then be possible to do a wide variety of medical information processing tasks without having to input the patient record in a special format. For example, one of the applications we are working on is the automation of routine screening of hospitalizations for health care evaluation. This process would have as input the discharge summary (in free narrative form) and the criteria which are to be applied in order to determine whether the case warrants peer review. In this application the system must be able to answer such questions as whether particular symptoms were noted, whether certain procedures were carried out, whether certain variables were monitored, etc.

* N. Sager, Syntactic Formatting of Science Information, Proc. FJCC 1972, AFIPS Press, 791-800.