Analyzing Language in Restricted Domains:

Sublanguage Description and Processing

edited by

Ralph Grishman
New York University

Richard Kittredge
University of Montreal

6 Automatic Structuring of Sublanguage Information

Carol Friedman
Courant Institute of Mathematical Sciences
New York University

ABSTRACT

Computer processing of free-text input can be used to obtain a database of patient information that contains in a structured form the relations among the medical events and observations recorded in the narrative. The information in the narrative can be structured by a computerized procedure because the information is expressed in a small number of sublanguage sentence types. For example, there is a treatment sentence type, a test and result sentence type, and a patient state sentence type. Each type is a syntactic relation among medical and English word classes. Once defined from an analysis of sample patient documents, these types can be used as target information structures into which the narrative portions of medical records can be mapped. This paper describes an implementation of text processing that utilizes English syntactic analysis and sublanguage word class combinations to determine to which of the known sublanguage sentence types an input occurrence conforms and to transform the occurrence into a standard form for the information carried by sentences of that type.

One of the goals of our work at the Linguistic String Project is the creation of an automated procedure whereby medical data from the narrative portions of patient documents are mapped into a structured form, making the information in the data available for use as a database. Such a procedure is possible for several reasons:

1. The narrative portions of patient documents form a sublanguage having a grammar that consists of: (a) a syntactic specification that is a modification of the English Grammar; (b) a vocabulary that can be classified into
both general English syntactic categories and semantic categories of the medical domain; (c) a statement of sublanguage sentence types as particular sequences of the semantic word classes in English syntactic relations.

2. The information in the documents falls naturally into a small number of informational sentence types; and the sentence types, as defined by the sublanguage grammar, correspond to the different types of information given in the documents.

3. Complex text sentences are composed of elementary sentences of the stated types, joined by connectives, or relations, between the sentence types.

For example, there is a test + result sentence type for sentences such as Culture was positive for streptococcus. And there is a patient-state sentence type for sentences such as Patient had meningitis. On the other hand, the sentence, Impression of meningitis was confirmed by finding cloudy cerebrospinal fluid, consists of two informational sentence types related to each other. It contains: (a) a test + result sentence type: finding cloudy cerebrospinal fluid; (b) a patient-state sentence type under an evidential operator: impression of meningitis was confirmed. The relation between (a) and (b) is by (by means of).

Another example of a single sublanguage sentence type is the medication type, as in Patient was given tylenol 2 tabs qid. However, mild increased pain for 3 weeks after stopping azathioprine consists of: (a) a medication sentence type: azathioprine was stopped; (b) a patient-state sentence type: mild increased pain for two weeks. The relation between (a) and (b) is the conjunction after.

Once the elementary units are defined from an analysis of a representative sample of the sublanguage text, they can be used as target structures (which we call information formats) into which the patient-record narrative can be mapped. These formats represent the semantic regularities of the sentences in the documents and therefore allow the system to have certain expectations depending on the sentence types. The systems can determine that a significant field is missing from a format and then use its knowledge of the sentence type to find and fill in the missing information. For example, in the patient-state format for increased pain, the format subject which should be either the whole patient, a body part, a body measure, or a body function, is missing. With the knowledge of what type of word is missing the system can search for it in a previous occurrence.

INFORMATION FORMATS FOR CLINICAL NARRATIVE

A format can be thought of as a prototype sentence or a template for a given type of information. The format has certain fixed fields or columns, and some of the fields have subfields.

Figure 6.1 shows a simplified version of the format that corresponds to the treatment-with-medication sentence type (FORMAT3 in the system). The prototype has the following information: Certain hospital personnel treated the patient with a certain medication, in a certain amount, with a certain frequency, and in a certain manner. The fields with an *" can have modifiers, which will be explained shortly. Figure 6.1 shows a structured form of the sentence Patient was treated with ampicillin 200 mg/kg/d Q6 hr. Several other sentences of this type are:

Patient was given IV ampicillin. 
Increase aspirin to 20 tabs daily. 
Continue penicillamine 500 mg daily.

FORMAT1, shown in simplified form in Fig. 6.2, is the template for general management information. It consists of four fields: an INST (for institutional personnel) field, a PT (patient) field, a VERB-MD (management verb) field, and a VERB (nonsublanguage verb) field. FORMAT1 can be thought of as having the following information: Certain hospital personnel followed a certain management procedure with regard to the patient. Under each column in Fig. 6.2 are examples of words that can be mapped into that column. The sample words under a column all have the same (or similar) medical subclasses. For example, admit, appointment, check-up, and schedule are all medical management verbs (which we classify as H-VMD). Several sentences that would map into FORMAT1 are:

FIG. 6.1 Information format for “Patient was treated with ampicillin 200 mg/kg/d q6 hr.”
Two fields in FORMAT1 have numbers next to them, indicating that these fields may have certain modifiers associated with them. These modifiers contain additional information such as evidential status, time, negation, or degree of certainty. The modifiers are also represented in standard templates. Figure 6.3 shows the templates for the time modifiers EVENT-TIME and TIME-ASP. EVENT-TIME contains time information such as for 2 days prior to admission. TIME-ASP contains fields for various aspects of time—begin, end, change, repetition, and time period. A modifier is placed immediately to the right of the column it modifies. The remaining modifiers are shown in Fig. 6.4. MODS contains fields NEG for negation, MODAL for nonfactuality, and EVID for evidential information. There is also a BP-MOD for auxiliary bodypart information, a QUANTITY for quantity information, and a TENSE for tense information. Notice that in Figs. 6.3 and 6.4 some of the modifiers may also have modifiers. Retaining the appropriate
hierarchy for the modifiers is important in order to preserve the information content of the narrative. For example, no fever is different from no increase in fever. In the format of no fever, no is the negation modifier of the format field SIGN-SYMPTOM (containing fever). In the format of no increase in fever, no is the negation modifier of the format field CHANGE (in TIME-ASP containing increase); increase is in TIME-ASP, modifying the format field SIGN-SYMPTOM (containing fever).

Another format (FORMAT2) is shown in simplified form in Fig. 6.5. FORMAT2 contains treatment-without-medication information. There is an INST (institution) field, a PT (patient) field, a VERB-TR (treatment) field, and a nonmedical VERB field. This format contains the following information: Certain institutional personnel performed a certain treatment on the patient. Some examples of FORMAT2 sentences are:

An appendectomy was performed on 5/3/81.
Pt was given chemotherapy.
LMD advised against physiotherapy.

FORMAT4 (see Fig. 6.6) is the test + result sentence type. It is a large format because it covers the spectrum of test and result combinations. Sentences of this type are:

CSF culture showed no growth.
Blood-gases normal.
Chest xray normal.
Culture was positive for streptococcus.
WBC 1000.

6. AUTOMATIC STRUCTURING OF SUBLANGUAGE INFORMATION

FORMAT5 (see Fig. 6.7) contains patient-state information. It is actually a family of formats for such sentences as the following:

Patient has sickle cell disease.
Right shoulder with calcium tendinitis.
Breathing with difficulty.
P.E. essentially normal.
BP was 120/72.

There are four additional formats that are not shown in detail because they do not contain specifically medical information. However, they are described here briefly. FORMAT00 contains English sentential operators, such as: Doctor felt that . . . ; Her mother noted . . . . This format usually is connected to another format in the sublanguage. For example, Her mother noted patient had increasing pain consists of two formats: an English sentential operator type: her mother noted; and a patient-state type: patient had increasing pain. FORMAT00 is important because it may affect the factuality of the format it is related to. A second format, FORMAT0, contains patient descriptor information. It has a field PT-DESCR, which has the following subfields: AGE, RACE, SEX, and FAMILY. All the other formats also contain a PT-DESCR field in case that information should occur within a sentence of the given format type (PT-DESCR was omitted from the format figures for simplicity). Another format is FORMAT6, which contains behavioral information such as Patient worked as a typist. There is also a template (CONNECTIVE) for housing conjuncts of various types and verbal connectives (verbs connecting two sentence types). Its structure is very simple: CONNECTIVE has one subfield naming the type of connective

Simplified

FORMAT2 (TREATMENT NOT MEDICATION)

FIG. 6.5 Medical word classes in FORMAT2.
FIG. 6.6 Medical word classes in FORMAT4.

(CONJOINED, SUBORDINATE-CONJ, etc.). This field contains the value of the connective (and, while, etc.). CONNECTIVE may also have EVENT-TIME, TIME-ASP, and MODS modifiers.

PROCESSING THE NARRATIVE

The processing has been divided into consecutive modules so that each module performs a distinct logical function and refines the output of the preceding module. Modularizing the process clarifies the steps of the analytic process and permits different and relatively independent implementations for each function. Figure 6.8 shows an overview of all the stages of processing. The modules are described in sequence in the following paragraphs, with greater detail supplied on the sublanguage-dependent stages.
The parser also uses a lexicon in which words are assigned both English syntactic categories and medical subclasses based on their usage. Figure 6.9 contains some sample dictionary entries. Words are assigned medical subclasses on the basis of where they occur in assertions relative to other words of the assertion and on the basis of the medical information they contain. For example, pain, rash, and fever all occur in the environment patient developed. They are all informationally disease indicators. Therefore, we assign them the subclass H-INDIC. On the other hand, joint, arm, and leg all occur in pain in ___. They are all body part words. Therefore, we assign them the subclass H-BODYPART.

The Semantic Selection Module

The semantic selection module utilizes patterns of medical subclass co-occurrence in particular syntactic relations to give a semantic characterization to the syntactic structures. The patterns are also used to help resolve sublanguage ambiguity and to improve the parse tree. Selection was first implemented in the LSP system as part of the parsing grammar (Grishman, Hirschman, & Friedman, 1982). The selection constraints were list driven for

DEVELOPED TV: (H-GROW, H-BEGIN),
VEN: (H-GROW, H-BEGIN).
PAIN N: (H-INDIC, SINGULAR).
IN P.
LEFT ADJ: (H-LOC),
TV,
VEN.
LEG N: (H-BODYPART, SINGULAR).

FIG. 6.9 Some simplified dictionary entries.
ease of maintenance and portability from one sublanguage to another. The present implementation does selection in two points in the processing: in the selection module and (for conjunctions) in the parsing stage. The list-driven mechanism to determine the sublanguage patterns has been retained from the first implementation; the use made of the information and the lists themselves have been altered.

In this module, subclass co-occurrence patterns for particular syntactic relations are matched, and the success or failure of the match is recorded along with a semantic characterization of the structures when possible. Figure 6.10 shows some partial entries for the list P-NSTGO-HOST, which consists of subclass co-occurrence patterns for a noun N and its prepositional right adjunct PN. Well-formed N P N subclass sequences are placed in the list P-NSTGO-HOST in the following order (examples, in parentheses, are taken from the first entry to the list P-NSTGO-HOST shown in Figure 6.10):

- the preposition P (in)
- the subclass of the head noun in the PN phrase (H-BODYPART)
- the semantic label of the PN phrase (BODYLOC-PN)
- the list of the allowable subclasses for the host of the PN phrase (H-INDIC, H-DIAG, . . . )

Consider the phrase swelling in arm, where in arm is a prepositional right adjunct of swelling. The subclass pattern is H-INDIC IN H-BODYPART, and it is on list P-NSTGO-HOST. The success of the match is recorded as follows: swelling is assigned a property (i.e. node attribute) called SELECT-ATT, which has the value H-INDIC; arm is assigned a SELECT-ATT with the value H-BODYPART. In addition, the PN structure is assigned an attribute called TYPE-ATT with value BODYLOC-PN.

The recording of a match is important for several reasons. For one, a successful match is preferable to an unsuccessful one. In the phrase swelling in arm for 2 days, for 2 days is a prepositional right adjunct of arm in the syntactic analysis. The pattern is H-BODYPART FOR NTIME1. The pattern FOR NTIME1 is a recognizable PN pattern; it is a time phrase and in our system is assigned the label TIME-PN. A TIME-PN can not have an H-

LIST P-NSTGO-HOST =

\[
\text{IN'}: (\text{H-BODYPART} : (\text{BODYLOC-PN} : (\text{H-INDIC}, \text{H-DIAG}, . . . )))
\]
\[
\text{H-INST} : (\text{NO-TYPE} : (\text{H-VTR}, \text{H-VMD}, . . . ))).
\]

\[
\text{FOR'}: (\text{NTIME1} : (\text{TIME-PN} : (\text{H-INDIC}, \text{H-DIAG}, \text{H-VTR}, \text{H-VMD}, . . . ))).
\]

FIG. 6.10 Partial entries from a selection list.

6. AUTOMATIC STRUCURING OF SUBLANGUAGE INFORMATION

BODYPART host. However, swelling is another possible host for in 2 days. This pattern is H-INDIC FOR NTIME1 and is acceptable for this relation. Therefore, this analysis is the preferable one. The parse tree is rearranged to reflect the improved analysis.

The most important reason for recording a successful pattern match of a structure is to provide a semantic characterization of that structure that will be used to obtain its final formatted representation. In our example, for 2 days is a time phrase relating to the event swelling. In the final format, for two days will be mapped into an EVENT-TIME modifier, modifying the format field SIGN-SYMPTOM containing swelling.

The recording of the result of a match is important also because it helps to resolve sublanguage ambiguities. A sublanguage homograph is a word that has more than one sublanguage class. For example, growth has the subclass H-INDIC, as in growth is a tumor, and the subclass H-GROW as in growth was normal. Thus, the phrase normal growth (where normal is a left adjunct of growth) has two possible patterns: H-NORMAL H-INDIC and H-NORMAL H-GROW. The first pattern match fails, but the second one is successful. Therefore, growth in this occurrence is assigned a SELECT-ATT with value H-GROW only. All processing that follows will treat growth informationally as an H-GROW word only.

Another example of a sublanguage homograph is felt. Felt has three subclasses:

- H-EXAMTEST as in No masses felt
- H-VPT as in Patient felt pain
- VSENT3 (an English sentential operator) as in It was felt that surgery was required.

In It was felt that surgery was required, the object is a sentential object. For the purpose of pattern matching, we give it the value SENTOBJ. There are three verb + object patterns for this sentence corresponding to the three different subclasses of felt. The patterns H-EXAMTEST SENTOBJ and H-VPT SENTOBJ fail, but the pattern VSENT3 SENTOBJ succeeds. Therefore, felt is assigned a SELECT-ATT with value VSENT3 only, and it will be treated as a sentential operator in all the processing that follows.

A word that has only one medical subclass is not a sublanguage homograph but may still present a problem of sublanguage ambiguity: its usage may or may not be a medical one. That is, a word with a medical subclass may be used in the narrative as an ordinary English word and not as a medical word. This phenomenon occurs most frequently with verbs. For example, wear has the subclass H-BEH for behavioral information. In Patient wore comfortable shoes, wear is a behavioral verb, but in the day wore on, wear is not. When a SUBJECT VERB OBJECT subclass pattern fails, the verb is as-
signed a SELECT-ATT with value FAIL-SEL, signifying that a sublanguage pattern failed with that verb. This is a signal that the medical subclass of the verb should be ignored and the verb is to be treated as an English syntactic verb only.

A similar procedure is followed when a LEFT-ADJUNCT NOUN subclass pattern fails. A failure in this situation means that the medical subclass of the left adjunct is to be ignored. For example, white has the subclass H-RACE. The pattern H-RACE H-PT as in white female is acceptable. However, the pattern H-RACE H-RX as in white pills is not. In that situation, white is not being used to denote the race of the patient or the pills. White is assigned a SELECT-ATT with value FAIL-SEL in the latter phrase. Therefore, informationally it will remain as a left adjunct of pills in the final format, and the H-RACE subclass will be ignored.

Selection has yet another important function. There are certain words that may be higher order operators or connectives in the sublanguage, joining one basic medical event to another. In the lexicon, these words are given a subclass that we call H-CONN for connective. The following are some examples of connectives:

due to as in headache due to fever
from as in rash from penicillin
for as in tylenol for headache
relieve as in tylenol relieved pain
cause as in fever caused headache

However, these words may also be used as ordinary prepositions or verbs. For example, in He visited a clinic for rheumatic patients, for is not connecting one event to another. Another example is tylenol for 2 days, where for 2 days is a time phrase associated with tylenol and is not a separate event. The selection component determines whether or not words with the H-CONN subclass are connectives. If a word is determined to be a connective, it is assigned a SELECT-ATT with value H-CONN. This characterizes a structure containing a word with a SELECT-ATT equal to H-CONN as being a structure consisting of two separate events related by that connective.

Another case where selection is extremely useful can be demonstrated by the compound noun relation. This is a highly degenerate structure where semantic co-occurrence patterns are used to help determine the basic underlying structure. Consider the pattern H-RX H-INDIC, as in penicillin rash. The underlying structure is a Noun + Prepositional Phrase (N P N) with the pattern H-INDIC FROM H-RX, or rash from penicillin. This phrase consists of two related patient events and will be treated as such. Now consider the compound noun pattern H-RX H-INDIC, as in headache remedy. The underlying structure for this pattern is the N P N phrase H-RX FOR H-INDIC, or remedy for headache. In this situation, the H-INDIC (headache) is not directly asserted about the patient but is implied, because the patient was given a remedy that was for a headache. Thus, for this pattern, whereas the two related events are H-RX (remedy) and H-INDIC (headache), the H-INDIC event (headache) is marked as implied.

Another pattern, H-BODYPART H-PROC (chest xray), can be transformed into H-PROC OF H-BODYPART. However, here the formats are the same for both forms. In both phrases, there is only one medical event xray modified by the body part chest, with the relation that the chest was xrayed.

One very important type of selection operates in the parsing stage rather than afterwards (in the selection module). This is sublanguage selection for conjunctural structures. Our grammar handles a rich range of conjunctural structures; therefore, to reduce ambiguity and obtain the best parse possible, it is crucial to restrain possible conjoinings in the parsing stage. In general, nouns or adjectives are allowed to conjoin only if they are in the same subclass or in a class of conjoinable subclasses. These constraints are discussed in greater detail in Hirschman (1982).

THE REGULARIZATION MODULES

The next two modules consist of regularization transformations. They map the structures into canonical forms without changing their informational content. One component consists of general English transformations and the other sublanguage-specific ones. An example of an English transformation is the one that expands conjunctural structures. Thus patient had fever and chills is transformed into two assertions: patient had fever and patient had chills. These two assertions are joined by the connective and. An example of a sublanguage-specific transformation is the one that breaks up an assertion into two fragments when there is a connective in the assertion (that is, the connective was assigned a SELECT-ATT with value H-CONN). Thus headache was due to fever is transformed into two fragments fever and headache connected by the relation be due to.

As a result of these two stages of processing, the original narrative sentences are structured into elementary assertions or fragmentary assertions joined by connectives, and, most importantly, these assertions conform to the basic sentence types of the sublanguage. Thus, the various linguistic forms that were used in the narrative to express the medical information have been reduced to standard forms: basic medical events and their relations. At this point, the particular sentence type that is occurring can be determined by
the presence of its characteristic subclasses in the elementary assertion and fragment.

The Formatting Module

It is the function of this component to map the structured trees into the appropriate information formats. When the narrative reaches this module, it has been structured into elementary or fragmentary assertions joined by connectives. Each assertion corresponds to a basic sentence type of the sublanguage, and each connective structure contains the type of connective (subordinate conjunction, coordinate conjunction), its value (i.e. while, and), and possibly modifier information. For each assertion and connective, the formatting component builds the appropriate format and maps the words of the assertion (or the connective) into the appropriate fields of the format. The mapping itself is a fairly straightforward procedure. In general, the syntax directs the order of the mapping, and the semantic subclasses (not the original subclasses but the subclasses assigned by the selection module) determine which fields the words are to be placed in. The processing in the preceding modules has broken each sentence into its component medical events and labeled them accordingly, so that the correct format can be built for it. Once the format is built, the semantic labels on the words (SELECT-ATT values), and phrases (TYPE-ATT values) are sufficient to drive the mapping, except for a few cases that are treated differently.

The LSP had a first implementation of medical formatting that produced a single composite format (Hirschman & Sager, 1982). In the present implementation, the mapping transformations produce individual formats for the different medical sentence types and are list driven, making this component easy to modify and also adaptable for other sublanguages. One list establishes correspondences between the semantic subclasses and the nonmodifier fields of the format. For example, a word that has the SELECT-ATT H-INDIC is mapped into the SIGN-SYMPTOM field of FORMAT5. Another list establishes correspondences between the semantic subclasses and the modifier fields of the format. In this case, however, the proper modifier must be built and inserted next to the appropriate format field. In addition to the list driven transformations, the formatting component also has some sublanguage-specific transformations. These involve mostly quantities and units such as 10 years old, 100 mg, 150 lbs, and 3 days ago, which map into one of the fields AGE, DOSE, QUANT, EVENT-TIME, or RESULT.

The important result of formatting the narrative is that the medically significant informational fields for all the various ways of reporting the information are the same. Figure 6.11 illustrates this by showing that the significant fields are filled identically for various forms of reporting pain in leg.

CONCLUSION

This paper has described an implementation of medical text processing based on English parsing and a medical sublanguage grammar. After parsing, sublanguage co-occurrence patterns are checked for well-formed subclass combinations in the syntactic relations given by the parse, and the parse tree is adjusted to reflect the preferred patterns. Structures are labeled according to their semantic content, and, where a word is in more than one medical subclass, the class conforming to its usage in the given context is selected. Sentences or phrases that contain more than one medical event (fever due to headache) are separated into two elementary structures joined by a connective. Finally, the semantically labelled elementary structures are mapped into information formats, which are generalized sublanguage sentence forms for housing the different types of patient information in the documents. The patient information that was originally given in narrative form is now organized and codified according to medical categories, so that it can be queried in terms of the medical categories and relations in the material.
ACKNOWLEDGMENTS

This research was supported in part by National Library of Medicine grant number 1-R01-LM03933, awarded by the National Institutes of Health, Department of Health and Human Services.

REFERENCES


7 General Semantic Patterns in Different Sublanguages

Elaine Marsh
*Naval Center for Applied Research in Artificial Intelligence*
*Naval Research Laboratory*
*Washington, D.C.*

ABSTRACT

A comparison of the semantic word class co-occurrence patterns in two different domains, a class of Navy messages about shipboard equipment failure and a set of medical discharge summaries, discloses that, despite superficial differences, the sublanguages share semantic relationships among classes of objects. A set of general semantic patterns is proposed. The semantic patterns, which are contextually determined and which represent stereotyped situations, can be considered instances of frames. By sequencing the semantic patterns, scriptal knowledge about discourse structures in a sublanguage can be represented. The domains of discourse of the different sublanguages can be viewed as sister nodes in a generalization/specialization hierarchy of discourse domains. The parent node represents the general domain of system failures. The shared/different semantic patterns of the sublanguage are shown to correspond to the relative positions of the discourse domains in the hierarchy. As supporting evidence, the domain of discourse of an artificial language, IEEE ATLAS programming language (for describing electronic signal failures), has also been placed in the hierarchy; its semantic patterns are a specialization of the patterns of its parent domain (equipment failures).

INTRODUCTION

A sublanguage is characterized by limited subject matter, lexical, syntactic, semantic and discourse properties, and text structure properties that are distinct from those of the standard language. Sublanguages under study