COLING 82, J. Horecký (ed.)
North-Holland Publishing Company

ANALYSIS AND PROCESSING OF COMPACT TEXT

Elaine Marsh and Naomi Sager

Linguistic String Project
New York University
New York, New York
U.S.A.

This paper describes the characteristics of compact text as revealed in computer analysis of a set of physician notes. Computer processing of the documents was performed using the LSP system for natural language analysis. A numerical breakdown of syntactic and semantic patterns found in the texts is presented. It is found that four major properties of compact text make it possible to process the content of the documents with syntactic procedures that operate on full free text.

INTRODUCTION

Linguistic mechanisms of compacting are common in situations where specialists record facts to be used by others in the same field. Compact text is found in notes and records within institutions (1), and in network messages among collaborators. It is found in a different form in everyday dialogue (2), and in headings and headlines (3).

(1) Positive for heart disease and diabetes.

(2) Q: How are you?
R: Fine.

(3) REAGAN NAMES WOMAN JUSTICE

Similar shortened forms also occur in published research protocols and non-numeric tables (as in building codes, geological surveys).

The NYU Linguistic String Project has developed a computer program to analyze and process the compact text of specialized technical areas, using a general parsing program and English grammar augmented by procedures specific to the given subject area [1,2]. In particular, the system has been tailored for the computer analysis of free-text medical records [3]. The note-taking style of medical records [4] uses a remarkable amount of reduced English sentence forms. For instance, in the documents reported on here, 49% of occurrences were incomplete sentences. Similar types of reductions have been described in other technical sublanguages [5]. In this paper, we report the results of using the LSP processor to obtain a precise description of the syntactic and semantic properties of a body of compact text.

DATA AND METHODS OF ANALYSIS

The data of this study consist of eight medical discharge summaries, each 1-2 pages in length. These physician reports were dictated and then transcribed by a medical typist into computer readable form. As a preliminary to machine processing, there was minimal pre-editing of the documents for such necessary formatting changes as

inserting a blank between a word and its following punctuation mark and occasional spelling corrections. Abbreviations were maintained as used in the text and were treated as dictionary entries linked to their full form in the dictionary.

A computer dictionary of the document words was obtained by look-up in the medical parsing dictionary developed by the LSP, which was augmented for new words appearing in these texts. The dictionary classifies words according to their major parts of speech (e.g. noun, verb, adjective), as well as certain English subclasses (e.g. plural, past) and special medical subfield classes. The medical classes are based on cooccurrence properties of the words as seen in a larger survey of the material and are checked for semantic consistency by a physician-consultant. In total, the medical classes currently number sixty-six. These are used in selectional constraints applied during parsing to resolve syntactic ambiguities. A smaller set of eighteen medical classes determine the major semantic sentence types discussed below. Descriptions of the 18 medical classes are given in Table 1, drawn from [6], where all the medical classes are defined and details of the text processing procedure are described.

The input sentences from the discharge summaries are structured by a three stage system of (1) parsing, (2) syntactic regularization, and (3) mapping into an information format.

The first step is to parse the document sentences with the Linguistic String Project parser [7]. This step begins with a dictionary lookup to associate the stored lexical information with each word occurrence in the sentence. The parsing utilizes a medically tuned grammar, which includes productions for the compact sentence syntactic types discussed below, and also productions for special sublanguage constructions (e.g. dose expressions, penicillamine 250 MG PO QD). The sentence parse identifies grammatical relations which hold among parts of the sentence, principally subject-verb-object relations and host-modifier relations. Also built into the grammar is a selectional mechanism which disambiguates multiply classified words based on the type of subject-verb-object or host-modifier relationships permitted in the sublanguage [8].

The second step is syntactic regularization. Each sentence undergoes a series of paraphrastic English transformations regularizing the syntax within the sentences in order to reduce the variety of syntactic structures to a set of basic syntactic relations [9]. The syntactic regularization does not alter the information content of the sentences. In addition, reduced word forms, such as abbreviations, are replaced by their full forms.

The final processing step is information formatting, which maps the words of the parsed, transformed sentence into a tabular representation of the information contained in the sentences. A word is mapped into the format column which corresponds to the information content of the word. In general, there is a 1-to-1 correspondence between the sublanguage word class and a particular format column. For example, a word of the medical class DIAG(nosis), e.g. meningitis, would be mapped into the DIAG column of the format. Formatting is based on cooccurrence patterns found in the text and on the lexical information obtained from the computer dictionary of medical vocabulary. The sets of filled format columns represent the semantic patterns found within the data. These semantic patterns will be discussed below.

The data was run on the LSP natural language processing system as implemented in FORTRAN and run on a Control Data 6600, requiring about 75,000 words of memory. The LSP system also runs on the CDC CYBER, the VAX, and UNIVAC 1100 machines. The English grammar, regularization component, and information formatting component are written in Restriction Language, a special high level language developed for writing natural language grammars.

B-FUNC	description of body-functions, hearing, movement.
B-MEAS	standard body-measures, weight, temperature, blood pressure.
B-PART	location of test or symptom, hand, spine, joint
EXAM	test or technique performed during physical exam percussion, hear (rales).
DEVEL	patient growth word, growth, puberty, birth.
DESCR	neutral descriptor term, yellow, flat, pale.
DIAG	diagnosis word, meningitis, sickle cell disease.
INST	institution, clinic, or doctor, emergency room, hematology, local doctor.
LAB-RES	result of laboratory test or culture, generally contains agent words, pneumococcus type 18, pathogen.
MED	medication or specific treatment, penicillin, ampicillin, transfusion.
NORM	word indicating normalcy or change towards normalcy, normal, negative, convalesce, improvement.
PT	patient
QUANT	numerical quantifier, possibly with unit.
S-S	non-normal sign or symptom, crisis, cold, headache.
TEST	laboratory test, including x-rays, chemistry, bacteriology, and hematology, x-ray, urinalysis, hematocrit.
V-MANAGE	general patient management, admission, follow, report, decision.
V-RESPOND	patient or symptom response word, respond to, controlled by.
V-TREAT	general treatment verb or noun, treatment, therapy.

Table 1

Medical Word Subclasses

RESULTS

In a test set of narrative hospital discharge summaries that were computer-parsed, 49% of the sentences were syntactically incomplete sentence forms ("fragments"). The fragments were of six types that can be related to full sentence forms on the basis of the elements which were regularly deleted: (i) deleted verb and object (or subject and verb), leaving a noun phrase (4); (ii) deleted tense and verb be (5); (iii) deleted subject, tense, and verb be (6); (iv) deleted subject (7); (v) deleted subject, tense, and verb be (passive predicate) (8); and (vi) deleted subject, tense, and verb be (infinitival complement) (9).

(4) Stiff neck and fever. (63%)
(5) Brain scan negative. (22%)
(6) Positive for heart disease and diabetes. (8%)
(7) Was seen by local doctor. (5%)
(8) Treated for meningitis. (2%)
(9) To be followed in hematology clinic.
(1 instance = .08%)

Viewing these fragments as deletion forms, it is possible to fill them out to full sentences that would be accepted as paraphrastic by "native speakers" of medical language. For example, occurrences (4)-(9) can be related to sentences (10)-(15) respectively.

(10) Patient had stiff neck and fever.
(11) Brain scan was negative.
(12) Patient/test/exam [depending on context] is positive for heart disease.
(13) Patient was seen by local doctor.
(14) Patient was treated for meningitis.
(15) Patient is to be followed in hematology clinic.

In a set of 8 analyzed hospital discharge summaries, we found 41 cooccurrence patterns of the subject-verb-object (SVO) type, stated in terms of 18 participating word classes. (There were further patterns of the host-modifier type and some larger patterns involving connective between SVO types.) The SVO patterns could be grouped into six more general types by defining a "superclass" (RESULT) consisting of the classes SIGN-SYMPTOM, LAB-RES, QUANT, NORMALCY, DIAGNOSIS, and DESCRIPTION that occurred with one of the subject classes PATIENT, BODY-PART, TEST. For example, sentence (5) is of the type TEST RESULT, where scan is the name of a test (occurring with BODY-PART word brain as modifier), and negative (in the class NORMALCY in the superclass RESULT) is the finding of the test. Examples of the sentence patterns are provided in Table 2.

CONCLUSIONS

The computer results presented above, considered along with manual analysis of other document sets, lead to several major conclusions about the characteristics of compact text:

I. Repetitive ungrammaticality is grammatical for the text set.

Within a given set of data, there are recurrent ungrammatical constructions. These forms can be characterized and made a part of the parsing grammar. The departures from grammaticality are limited and can be related in a regular way to full sentente types in English.

II. Word choice is quasi-grammatical.

In repetitive single-topic text, word subclasses that are specific to the subject

TABLE 2: Medical Semantic Patterns

PATTERN			
I (22%)	B-PART + TEST V RESULT = {LAB-RES/NORM/QUANT/S-S/DIAG/DESCR}		
		LAB-RES	H. Influenzae type B from CSF.
		NORM	Spinal fluid was negative.
		QUANT	Hematocrit 22.6 percent.
		S-S	Chest x-rays suggest pleural effusion.
		DIAG	Chest x-rays revealed bilateral pneumonia.
II (53%)	EXAM/FUNCTION V RESULT (excluding LAB-RES)		
	B-PART	S-S	Developed swollen hands; Abdomen showed no organomegaly.
		DIAG	Lingular pneumonia.
		NORM	ENT negative; Lungs clear.
		DESCR	Conjunctivae were pale.
	B-FUNC	NORM	Appetite is good; She slept well; Eating well.
		S-S	Pain on dorsiflexion of left foot.
	B-MEAS	QUANT	Temp 99.6, Pulse 120, Respiration Rate 16, Weight 19.5 lbs.
		NORM	Temp normal.
		DESCR	Low grade temp finally cleared.
	B-PART-EXAM	NORM	DTR's are normal.
		S-S	Heart murmur heard; Slight tenderness to touch.
		QUANT	Liver palpable 6 cm.
	Ø	DIAG	Meningitis; Has sickle cell disease.
		S-S	Patient began to vomit; Patient developed mild cold.
		NORM	She remained well; Pt had a complete recovery.
		DESCR	Patient was active; Occasionally rubs hands.
III (2%)	DEVEL	RESULT	
		NORM	Well developed; Growth and development of patient was normal.
		S-S	Pt is product of complicated pregnancy.
		DESCR	Delivery was spontaneous.
		QUANT	Product of gravida 10 para 9 pregnancy.
IV (9%)	INST V-MANAGE {DIAG/S-S}		
		DIAG	1st admission to BH for meningitis.
		S-S	She was seen in Emergency Room because of a temp of 105.
		Ø	To be followed in hematology; Seen in Pediatric Emergency Service.
V (13%)	(INST) V-TREAT-with MED		Treatment with ampicillin; Partial exchange transfusion was given; Was given phenobarbitol; Resume prophylactic penicillin orally;
VI (1%)	{PT/S-S} V-RESPOND-to MED		
	PT		She responded well to penicillin.
	S-S		Seizures were controlled by valium.

Notes:
1) Examples for each pattern are illustrative, not exhaustive.
2) In many cases B-PART word, TEST word are deleted when reconstructable from context, e.g. CSF grew out pneumococcus = CSF culture grew out pneumococcus.
3) Mention of PATIENT is omitted in Patterns I-V.
4) Key: ( ) = optional element
{ } = choice of one among elements in braces

matter are found in particular combinations. These patterns are so marked that deviations can be considered ungrammatical for the discourse. For example, in medical records, (16) would be possible, while (17) would not.

(16) Patient admitted to hospital on 11/5/81.
(17) *Meningitis admitted to hospital on 11/5/81.

III. Deletions are reconstructable.

Deletions are reconstructable on the basis of both syntax and regularity of subclass patterning. It was seen above that deleted (reconstructable) elements are either function words known to be deleted in other English forms (e.g. be), or distinguished words of the sublanguage (e.g. patient).

IV. Texts are convergent.

While it would be improper to say "when you've seen one, you've seen them all," compact texts within a given area are remarkably similar. In the set of eight documents referred to above, six generalized semantic patterns occurred in the first document processed. No new types were recognized in the remaining seven documents.

On the lexical level, while new vocabulary is found in each new document (so-called "seepage"), this tends to taper off. In a prior study of journal articles, it was found that seepage after the 7th article remained at about 20%. In the processing of medical records, we found that, when changing from one medical subfield to another, the new vocabulary in a set of documents containing 2200 distinct lexical items was 27%.

The above four properties of compact text: grammaticality despite syntactic deviation, regular patterning of subject-specific vocabulary, recoverable deletions, and convergence as new texts are analyzed, make it possible to process the content of documents with syntactic procedures that operate on the full free text.

ACKNOWLEDGMENTS

This research was supported in part by National Science Foundation grant number IST79-20788 from the Division of Information Science and Technology, and in part by National Library of Medicine grant number 1-ROl-LM03933 awarded by the National Institutes of Health, Department of Health and Human Services.

REFERENCES

1. Sager, N. (1981). Natural Language Information Processing: A Computer Grammar of English and Its Applications. Addison-Wesley, Reading, Massachusetts.
2. Sager, N. (1978). Natural Language Information Formatting. Adv. Comput. 17, (M.C. Yovits, ed.), 89-162. Academic Press, New York.
3. Hirschman, L. and N. Sager (1982). Automatic Information Formatting of a Medical Sublanguage. Sublanguage: Studies of Language in Restricted Semantic Domains (R. Kittredge and J. Lehrberger, eds.). Walter de Gruyter, Berlin.
4. Anderson, B., Bross, I.D.J. and N. Sager (1975). Grammatical Compression in Notes and Records. Am. J. Comput. Linguist. 2:4.
5. Lehrberger, J. (1982). Automatic Translation and the Concept of Sublanguage. Sublanguage (as in ref. 3, above).
6. Sager, N. and L. Hirschman (1978). Information Structures in the Language of Science. String Program Repts. 12. Linguistic String Project, New York Univ.
7. Grishman, R., Sager, N., Raze, C. and B. Bookchin (1973). The Linguistic String Parser. Proc. 1973 Nat. Comput. Conf., 427-434, AFIPS Press, Montvale, N.J.
8. Grishman, R., Hirschman, L., and Friedman, C. (1982). Natural Language Interfaces Using Limited Semantic Information. Proc. 9th Int. Conf. Comput. Ling.
9. Hobbs, J. and R. Grishman (1976). The Automatic Transformational Analysis of English Sentences. Int. J. Comput. Math., Sect. A, Vol. 5, pp. 267-283.