UTILIZING DOMAIN-SPECIFIC INFORMATION FOR PROCESSING COMPACT TEXT

Elaine Marsh

Linguistic String Project
New York University
New York, New York

ABSTRACT

This paper identifies the types of sentence fragments found in the text of two domains: medical records and Navy equipment status messages. The fragment types are related to full sentence forms on the basis of the elements which were regularly deleted. A breakdown of the fragment types and their distributions in the two domains is presented. An approach to reconstructing the semantic class of deleted elements in the medical records is proposed which is based on the semantic patterns recognized in the domain.

I. INTRODUCTION

A large amount of natural language input, whether to text processing or question-answering systems, consists of shortened sentence forms, sentence "fragments". Sentence fragments are found in informal technical communications, messages, headlines, and in telegraphic communications. Occurrences are characterized by their brevity and informational nature. In all of these, if people are not restricted to using complete, grammatical sentences, as they are in formal writing situations, they tend to leave out the parts of the sentence which they believe the reader will be able to reconstruct. This is especially true if the writer deals with a specialized subject matter where the facts are to be used by others in the same field.

Several approaches to such "ill-formed" natural language input have been followed. The LIFER system [Hendrix, 1977; Hendrix, et al., 1978] and the PLANES system [Waltz, 1978] both account for fragments in procedural terms; they do not require the user to enumerate the types of fragments which will be accepted. The Linguistic String Project has characterized the regularly occurring ungrammatical constructions and made them part of the parsing grammar [Anderson, et al., 1975; Hirschman and Sager, 1982]. Kwasny and Sondheimer (1981) have used error-handling procedures to relate the ill-formed input of sentence fragments to well-formed structures. While these approaches differ in the way they determine the structure of the fragments and the deleted material, for the most part they rely heavily, at some point, on the recognition of semantic word-classes. The purpose of this paper is to describe the syntactic characteristics of sentence fragments and to illustrate how the domain-specific information embodied in the cooccurrence patterns of the semantic word-classes of a domain can be utilized as a powerful tool for processing a body of compact text, i.e. text that contains a large percentage of sentence fragments.

II. IDENTIFICATION OF FRAGMENT TYPES

The New York University Linguistic String Project has developed a computer program to analyze compact text in specialized subject areas using a general parsing program and an English grammar augmented by procedures specific to the subject areas. In recent years the system has been tailored for computer analysis of free-text medical records, which are characterized by numerous sentence fragments. In the computer-analysis and processing of the medical records, relatively few types of sentence fragments sufficed to describe the shortened forms, although such fragments comprised fully 49% of the natural language input [Marsh and Sager, 1982]. Fragment types can be related to full forms on the basis of the elements which are regularly deleted. Elements deleted from the fragments are from one or more of the syntactic positions: subject, tense, verb, object. The six fragment types identified in the set of medical records are shown in Table 1 as types I-VI.

A feature of fragment types that is not immediately obvious is the fact that they are already known in the full grammar as parts of fuller constructions. The fragment types reflect deletions found in syntactically distinguished positions within full sentences, as illustrated in Table 2. For example, in normal English, a sentence that contains tense and the verb *be* can occur as the object of verbs like *find* (e.g. *She found that the sentence was ambiguous.*). In the same environment, as object of *find*, a reduced sentence can occur in which the tense and verb *be* have been omitted, as in fragment type I (e.g. *She found the sentence ambiguous.*). In the same manner, other reduced forms reflected in fragment types also represent constructions generally found as parts of regular English sentences.

The fact that the fragment types can be related to full English forms makes it possible to view them as instances of reduced SUBJECT-VERB-OBJECT patterns from which particular components have been deleted. Fragments of type I can be represented as having a deleted tense and verb *be*, of type II as having a deleted subject, tense, and verb *be*, etc. This makes it relatively straightforward to add them to the parsing grammar,

TABLE 1. FRAGMENT TYPES

	MATERIAL DELETED	FRAGMENT STRUCTURE	MEDICAL EXAMPLE	NAVY EXAMPLE
I	TENSE, BE	A. N + PASSIVE PRED	PAIN NOTED IN HANDS AND KNEES. NO MURMURS HEARD. NO MEDICINES INDICATED.	CORE MEMORY UNIT NOT YET RECEIVED. OVERHAUL NOT COMPLETED. IMPROPER REPAIR WORK PERFORMED.
		B. N + PROG PRED	AN IRRITABLE CHILD CRYING.	PROCUREMENT ACTION BEING INITIATED. 'A' ENGINE OPERATING NORMALLY.
		C. N + ADJ PRED	EXTREMITIES NORMAL. MOTHER AND FATHER ALIVE AND WELL. ACETONE NEGATIVE.	SYSTEM INOP. ONE OF FOUR AC UNITS NOT AVAILABLE. PMS ADEQUATE.
		D. N + PN	WD FEMALE IN MARKED DISTRESS.	REPAIRS IN PROGRESS.
		E. N + Q	PROTEIN 25 SISTERS 19 AND 16.	------
		F. N + N	SKIN - NO ERUPTIONS.	BOILER 1A CONTAMINATION 27EPM CHLORIDE.
II	SUBJ, TENSE, BE [VERBAL PRED]	A. PASSIVE PRED	TOLD TO RETURN TO CLINIC. EXHIBITED GOOD FLEXION.	SENT BY FLASHING LIGHT.
		B. PROG PRED	------	AWAITING RECEIPT OF REPLACEMENT. TROUBLESHOOTING.
III	SUBJ, TENSE, BE	A. ADJ PRED	SLIGHTLY EDEMATOUS. ALLERGIC TO PENICILLIN.	UNKNOWN AT THIS TIME. UNABLE TO RADIATE CW1.
		B. PN PRED	ON PROPHYLAXIS WITH BICILLIN.	FOR COMDESRON ELEVEN.
IV	SUBJ, TENSE, VERB [OR TENSE, VERB, OBJ]	NP	BACTERIAL MENINGITIS. NO DISCHARGE OR INJECTION.	NORMAL DETERIORATION. MINOR DAMAGE TO SHIPS AT ANCHOR.
V	SUBJ, TENSE, BE	INFINITIVAL PRED	TO BE FOLLOWED IN MEDICAL CLINIC.	------
VI	SUBJECT	TENSE, VERB, OBJ	WAS SEEN BY LOCAL MD. HAS BEEN TAKING PENICILLIN.	CAN CONTINUE ASSIGNMENT. HAVE CHECKED STABILITY OF 150 PSI SYSTEM.
VII	2 SUBJECTS	VERB,VERB,OBJ	------	REQUEST ADVISE ARRANGEMENTS FOR PICKUP. INTEND DELAY REPAIR ATTEMPTS ON REDUCERS.

TABLE 2. DELETION FORMS IN NORMAL ENGLISH

I. DELETED TENSE, VERB BE

A. N + PASSIVE PRED	THE KING HAD HIM BEHEADED.
B. N + PROGRESSIVE PRED	WE OBSERVED BILL TALKING TO HIMSELF.
C. N + ADJECTIVE PRED	SHE FOUND THE SENTENCE AMBIGUOUS.
D. N + PN	THEY FOUND HIS IDEA OF INTEREST.
E. N + Q	JOHN THOUGHT HIM 25 OR YOUNGER.
F. N + N	THEY CONSIDERED HER THEIR SAVIOUR.

II. DELETED SUBJECT, TENSE, VERB BE [VERBAL PREDICATE]

A. PASSIVE PREDICATE	THE MAN, FINISHED WITH HIS WORK, WENT HOME.
B. PROGRESSIVE PREDICATE	MARY LEFT WHISTLING A HAPPY TUNE.

III. DELETED SUBJECT, TENSE, VERB BE

A. ADJECTIVE PREDICATE	GRACIOUS AS EVER, SHE WELCOMED HER GUESTS.
B. PN PREDICATE	THE GUARD, IN GREAT ALARM, CALLED THE POLICE.

IV. DELETED SUBJECT, TENSE, VERB BE

NOUN PHRASE	THE CHILD, A CLUMSY DANCER, TWISTED HER ANKLE.

V. DELETED SUBJECT, TENSE, VERB BE

INFINITIVAL PREDICATE	THEY TOOK THE TRAIN TO AVOID THE TRAFFIC.

and, at the same time, provides a framework for identifying their semantic content by relating them to the corresponding full forms.

The number of fragment types that occur in compact text of different technical domains appears to be relatively limited. When the fragment types found in medical records were compared with those seen in a small sample of Navy equipment status messages, five of the six types found in the medical records were also found in the Navy messages. Only one additional fragment type was required to cover the Navy messages. This type appears in Table 1 as type VII, in which two subjects have been deleted (Request advise arrangements for pick up).

While the number of fragment types is relatively constant, the distribution of fragment types varies according to the domain of the text. Table 3 shows distributions for each of the fragment types identified in Table 1. For example, in Table 3, while fragment type IV, from which subject, tense, and verb have been deleted, is most frequent in medical records, it is a much less frequent type in the Navy messages. On the other hand, type VI, from which a subject has been deleted, is relatively infrequent in medical records, but much more frequent in Navy messages.

In addition, the different sections of the input differ with respect to the ratio of fragments to whole sentences and in the types of fragments they contain. For example, the different sections of the medical records that were analyzed (e.g. HISTORY, EXAM, LAB-DATA, IMPRESSION, COURSE IN HOSPITAL) were distinguished by differences in the distribution of the fragment types. The EXAM paragraph of the medical texts, in which the physician describes the results of the patient's physical examination, contained a relatively large number of fragments of type III, especially adjective phrases. The COURSE IN HOSPITAL paragraph contained a larger number of complete sentences than the other paragraphs.

TABLE 3. DISTRIBUTION OF FRAGMENT TYPES

TYPE	MEDICAL	NAVY
I.	22%	36%
II.	1%	6%
III.	12%	11%
IV.	61%	15%
V.	1%	0%
VI.	2%	28%
VII.	0%	4%

III. RECONSTRUCTION OF DELETIONS

The deletions which relate fragment types to their full sentence forms fall into two main classes: (I) those found virtually in all texts and (II) those specific to the domain of the text.

Just as the fragment types can be viewed as incomplete realizations of syntactic S-V-O structures, the semantic patterns in sentence fragments can be considered incomplete realizations of the semantic S-V-O patterns. In general terms, the structure of information in technical domains can be specified by a set of semantic classes, the words and phrases which belong to these classes, and by a specification of the patterns these classes enter into, i.e. the syntactic relationships among the members of the classes [Grishman, et al., 1982; Sager, 1978]. In the case of the medical sublanguage processed by the Linguistic String Project, the medical subclasses were derived through techniques of distributional analysis [Hirschman and Sager, 1982]. Semantic S-V-O patterns were then derived from the combinatory properties of the medical classes in the text [Marsh and Sager, 1982]; the semantic patterns identified in a text are specific to the domain of the text. While they serve to formulate sublanguage constraints which rule out incorrect syntactic analyses caused by structural or lexical ambiguity, these relationships among classes can also provide a means by which deleted elements in compact text can be reconstructed. When a fragment is recognized as an instance of a given semantic pattern, it is then possible to specify a set of the semantic classes from which the medical sublanguage class of the deleted element can be selected.

On a superficial level, the deletions of be in fragment types Ic-f and IIIa-b, for example, can be reconstructed on purely syntactic grounds by filling in the lexical item be. However, it is also possible to provide further information and specify the semantic class of the lexical item be by reference to the semantic S-V-O pattern manifested by the occurring subject and object. For example, in type If fragment skin no eruptions, skin has the medical subclass BODYPART, and eruptions has the medical subclass SIGN/SYMPTOM. The semantic S-V-O pattern in which these classes play a part is:

BODYPART-SHOWVERB-SIGN/SYMPTOM

(as in Skin showed no eruptions). Be can then be assigned the semantic class SHOWVERB. Protein 25, type Ie, enters into the semantic pattern:

TEST-TESTVERB-TESTRESULT

and be can be assigned the class TESTVERB, which relates a TEST subject with a TESTRESULT object. Assigning a semantic class to the reconstructed be maximizes its informational content.

In addition to reconstructing a distinguished lexical item, like the verb be, along with its semantic classes, it is also possible to specify the set of semantic classes for a deleted element, even though a lexical item is not immediately reconstructable. For example, the fragment To receive folic acid, of Type VI, contains a verb of the PTVERB class and a MEDICATION object, but the subject has been deleted. The only semantic pattern which permits a verb and object with these medical subclasses is the S-V-O pattern:

PATIENT-PTVERB-MEDICATION

Through recognition of the semantic pattern in which the occurring elements of the fragment play a role, the semantic class PATIENT can be specified for the deleted subject. Patient is one of the distinguished words in the domain of narrative medical records which are often not explicitly mentioned in the text, although they play a role in the semantic patterns.

The S-V-O relations, of which the fragment types are incomplete realizations, form the basis of a procedure which specifies the semantic classes of deleted elements in fragments. Under the best conditions, the set of semantic classes for the deleted form contains only one element. It is also possible, however, for the set to contain more than one semantic class. For example, the type Ia fragment Pain also noted in hands and knees, when regularized to normal active S-V-O word order as noted pain in hands and knees, has a deleted subject. The set of possible medical classes for the deleted subject consists of {PATIENT, FAMILY, DOCTOR}, since a fragment with a verb of the OBSERVE class, such as note, and an object of the SIGN/SYMPTOM class, such as pain, can enter into

SUBJECT	VERB	OBJECT	
FAMILY	OBSERVE	SIGN/SYMPTOM	(MOTHER OBSERVED FEVER.)
PATIENT	OBSERVE	SIGN/SYMPTOM	(PATIENT OBSERVED FEVER.)
DOCTOR	OBSERVE	SIGN/SYMPTOM	(DOCTOR OBSERVED FEVER.)

FIGURE 1. EXAMPLES OF SUBJECT-VERB-OBJECT PATTERNS

any of the S-V-O patterns in Figure 1. The choice of one subclass for the deleted element from among elements of the set of possible subclasses is dependent on several factors. First, properties of paragraph structure of the text place restrictions on the selection of semantic class for a deleted element. The fragment noted pain in hands and knees would select a DOCTOR subject if written in the IMPRESSION or EXAM paragraph of the text, but, in the HISTORY paragraph, a PATIENT or FAMILY subject could not be excluded. A second factor is the presence of an antecedent having one of the semantic classes specified for the deleted element. If a possible antecedent having the same semantic class can be found, subject to restrictions on change of topic and discourse structure, then the deleted element can be filled in by its antecedent, restricting the semantic class of the deleted element to that of the antecedent. However, an antecedent search may not always be successful, since the antecedent may not have been explicitly mentioned in the text. The antecedent may be one of a class of distinguished words in the sublanguage, such as patient and doctor, which may not be previously mentioned in the body of the text.

Thus, semantic patterns derived from distributional analysis permit the specification of a set of semantic classes for deleted elements in texts characterized by a large proportion of sentence fragments. This specification can facilitate the reconstruction of deleted elements by limiting choice among possible antecedents.

IV. CONCLUSION

In this paper, seven deletion patterns found in technical compact text have been identified. The number of fragment types is relatively limited. Five of the seven occur in the full grammar of English as subparts of fuller structures. These syntactic fragment types can be viewed as incomplete realizations of syntactic SUBJECT-VERB-OBJECT structures; the semantic patterns in sentence fragments are found to be incomplete realizations of the semantic SUBJECT-VERB-OBJECT patterns found in full sentences. Semantic classes can be specified for deleted elements in sentence fragments based on these semantic patterns.

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-R01-LM03933 awarded by the National Institute of Health, Department of Health and Human Services.

REFERENCES

Anderson, B., Bross, I.D.J. and N. Sager (1975). Grammatical Compression in Notes and Records. American Journal of Computational Linguistics 2:4.

Grishman, R., Hirschman, L., and C. Friedman (1982). Natural Language Interfaces Using Limited Semantic Information. Proceedings of 9th International Conference on Computational Linguistics (COLING 82), Prague, Czechoslovakia.

Hendrix, G. (1977). Human Engineering for Applied Natural Language Processing. Proceedings of 5th IJCAI, Cambridge, Mass.

Hendrix, G., Sacerdoti, E., Sagalowicz, D., and J. Slocum (1978). Developing a Natural Language Interface to Complex Data, ACM TODS 3:2.

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,

Kwasny, S.C. and N.K. Sondheimer (1981). Relaxation Techniques for Parsing Ill-formed Input. American Journal of Computational Linguistics 7:2.

Marsh, E. and N. Sager (1982). Analysis and Processing of Compact Text. Proceedings of the 9th International Conference on Computational Linguistics (COLING 82), Prague, Czechoslovakia.

Sager, N. (1978). Natural Language Information Formatting: The Automatic Conversion of Texts to a Structured Data Base. In Advances in Computers 17 (M.C. Yovits, ed.), Academic Press, New York.

Waltz, D. (1978). An English Language Question Answering System for a Large Relational Data Base, CACM 21:7.