AUTOMATED PROCESSING OF MEDICAL ENGLISH

Introduction

The present interest of the scientific community in automated language processing has been awakened by the enormous capabilities of the high speed digital computer. It was recognized that the computer which has the capacity to handle symbols effectively can also treat words as symbols and language as a string of symbols.

Automated language processing as exemplified by current research, had its origin in machine translation. The first attempt to use the computer for automatic language processing took place in 1954. It is known as the "IBM-Georgetown Experiment" in machine translation from Russian into English. (1,2)

The experiment revealed the following facts:

a. the digital computer can be used for automated language processing, but

b. much deeper knowledge about the structure and semantics of language will be required for the determination and semantic interpretation of sentence structure.

The field of automated language processing is quite broad; it includes machine translation, automatic information retrieval (if based on language data), production of computer generated abstracts, indexes and catalogs, development of artificial languages, question answering systems, automatic speech analysis and synthesis, and others.

Approaches to automatic information retrieval, quantitative studies of generic relations between languages and style analysis, have been based to a great extent on statistical considerations, such as frequency counts of linguistic units (phomemes, morphemes, words, fixed phrases). In each of these approaches linguistic analysis was considered to be a useful but insufficient method for automated information processing because of the many unresolved problems in language analysis.

Implementation of statistical techniques for automated indexing, classification and abstracting has proved useful despite certain limitations caused by our lack of knowledge of language.

Some of the major problems in language processing are:

a. There is no method for storing in a computer the speaker's knowledge of the universe.

b. Syntactic and semantic ambiguities pose severe difficulties for implementation.

c. Difficulties associated with the design of a general purpose formal semantic language (intermediate language) into which an input natural language could be mapped by an algorithm.

d. Recognition and interpretation of logical inferences which are contained implicitly in natural language.

e. Lack of computer-oriented dictionaries and microglossaries.

In recent years mathematically oriented studies of the nature of natural languages have been directed to the development of formal models of grammars, such as context free, context sensitive and

transformational grammars. Formal characteristics of these models of grammars, their generative power, and their adequacies and inadequacies may be found in the literature (3, 4, 5, 6, 7).

Several noted scientists, such as Bar-Hillel (8), have expressed a pessimistic view in regard to practical implementation of machine translation.

Nevertheless, there is merit in continuing efforts for more fundamental research in the area of formal and applied linguistics and computer applications. Even if we are not able to resolve all the problems in language processing at once, limited goals can be attained and tested for validity by design of a model for language processing within a restricted language domain, such as medicine.

Some Characteristics of Medical English

Aware of the many problems associated with automated processing of natural language, we have limited our efforts, for the present, to the language domain used in pathology diagnoses, a subset of Medical English.

Medical diagnosis may be described as the process used by the physician to determine the nature of disease, or as the art of distinguishing one disease from another. The name which is assigned to a disease implies the unique configuration of signs and symptoms believed to be characteristic of the condition which has been diagnosed. The diagnosis can be regarded as a summary of the more complete medical document in a conventicnalized medical style.

Medical diagnoses are characteristically free of verb phrases. The copulative verb "to be" is frequently implied by the use of comma. Often, the pseudosentence structures appear to be grammatically illogical. Nevertheless, these structures carry semantic meaning and are generally understood by others in medicine. Modifiers frequently occur in discontinuous sequence with the nouns they modify. Anaphoric expressions are commonplace.

The terminology consists of a mixture of Latin, Greek and English derivatives. Not uncommonly, diagnostic statements exhibit features of all three languages. Evidence of heterogeneous linguistic origin is also found in single word forms. The language is rich in the use of compound word forms which are segmentable into single constituents.

The distinctive semantic features of diagnostic statements may be categorized as follows:

- anatomic site affected, or body system involved in the disease process;
- disease condition, including structural changes ranging from gross observations to intracellular ultrastructural changes;
- causative agent of the abnormality;
- disease manifestations, including physiological and chemical changes, observable manifestations, and symptoms reported by the patient;
- therapeutic agents or processes used;

- causal relationships among disease entities;
- method or source of diagnosis.

Two or more of these distinctive semantic features may be combined in a single conceptual unit, e.g., "measles" implies both the specific infectious disease manifested, and the etiology, the rubeola virus; while "pneumonia" describes the inflammatory disease process or condition, as well as the anatomic site affected, lung.

On the other hand, the precise designation of the location at which a disease entity has manifested itself may require a complex statement for adequate description of the semantics relative to anatomic site affected, e.g., a lesion may be found in the "apicoposterior segment in the upper division of the upper lobe of the lung."

After mentioning some of the pecularities of Medical English we will turn to the description of the system for automated processing of Medical English which is now under development at the Division of Computer Research and Technology, National Institutes of Health. (9, 10, 11, 12, 13)

Systematized Nomenclature of Pathology

In any type of automated language processing at least two basic components are required:

a. Lexicon

b. Grammar

Experience in machine translation revealed that commercial dictionaries consisting mostly of word lists are not suitable for automated language processing. Their main disadvantages are that they are out-of-date, incomplete and inaccurate for the purpose of morphological, syntactic and semantic analysis.

In our work we have been using as a lexicon base the Systematized Nomenclature of Pathology (SNOP) (14), the structure of which is described below:

SNOP is a special purpose lexicon created by pathologists to assist them in the organization and retrieval of information. The SNOP language consists of a relatively rich word vocabulary and a primitive grammar.

A term or conceptual unit is listed in only one of the four semantic categories of the vocabulary and is assigned a unique numerical code within the given information class.

The four semantic categories of the SNOP are:

Topography (T) - the body site affected

Morphology (M) - the structural changes resulting from disease

Etiology (E) - the causative agents (micro-organisms, drugs and chemicals)

Function (F) - the physiological manifestations associated with disease, including symptoms and a limited number of specific infectious diseases.

The conceptual units in each category are information content words which are used by the pathologist to convey in a condensed form the concept being described. The basic syntactic structures of the dictionary entries are: single nouns, adjective phrases, attributive noun phrases and participial phrases.

The code of dictionary entries is divided into four separate and independent fields T, M, E, F. Within a given field, terms are assigned a four digit number. The first digit refers to the section of the field, while the other three digits indicate progressively finer subdivisions. These groupings reflect natural relations among the terms insofar as possible. The code structure permits selection of more specific terms by moving down the list and of more generic terms by moving up the list (see Appendix 1).

Semantic linkage pointers are also incorporated in the dictionary which enable the cross-referencing of information within the same or another category.

We intend to extend the semantic codes and also to allow the insertion of additional semantic information. The individual codes will be linked together to represent the content of the message, and messages will be linked to represent the content of a diagnosis or other relevant medical document.

Grammar

Automated processing of Medical English (APME) consists of a

series of computer programs which given as input a body of medical text, will produce as output, a linguistic description and semantic interpretation of the given utterance.

In the initial stage of our research, our attention was focused on pathology diagnoses since this domain of discourse intersects with all other medical specialty areas. Consequently, research work in this area should be applicable to other medical specialties.

The APME parsing algorithms consist of series of interrelated linguistic and programming operations which are described briefly as follows:

Morphological Analysis embodies the identification and transformation of terminal morphemes and recognition of a limited set of prefixes. The input to the morphological analysis and subsequent morphosyntactic transformations is the unedited Medical English of a pathologist. The transformation procedure consists of a set of rules by which the adjective forming suffixes are substituted by a set of nominalizing suffixes (adjective to noun transforms) or plural nominal suffixes are substituted by their singular allomorphs (plural to singular noun transforms) and, finally, transforms of nouns to their synonymous or near-synonymous forms (i.e., DISENTERY → DISENTERIA). (12)

The main reason for the transformation of terminal morphemes is to provide a means for the successful retrieval of word forms which occur in the text in a derivative form but are listed in the SNOP dictionary in an alternative form.

We have been preparing rules for morphosemantic segmentation of composite word forms which are mainly derived from Greek or Latin. The decomposition implies the recognition of constitutents, the order of constituents, how they are intertwined and the assignment of semantic value to productive components.

For example, the composite 'BRONCHITIS' is segmented into two components C1-T and C2-M, namely:

a. Terminal morpheme-ITIS (C2-M) which is the semantic marker of inflammation process (semantic category 'M'), and

b. The stem morpheme BRONCH(US) (C1-T), the site of the body where the inflammation process occurs (semantic category 'T'). The semantic structure of the class of word forms such as 'BRONCHITIS' is expressed as

$$W(T, M) \rightarrow (C1\text{-}T) + (C2\text{-}M)$$

where the order of semantic components C1-T and C2-M is fixed and cannot be reversed.

It is expected that the analysis of the semantic structure of components into their semantic constituents and the formalization of the structural and semantic relationships among them will be useful for the preparation of computer-oriented medical micro-glossaries.

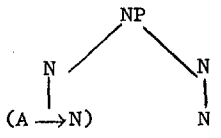
Syntactic and Semantic Analysis

Syntactic and semantic analysis of medical text involves the determination of sentence structure and its semantic interpretation.

This problem can be resolved by the implementation of transformational rules by which different types of noun phrases will be transformed into a set of semantically equivalent phrases. For example:

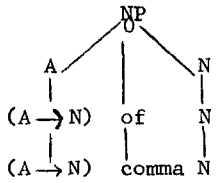
NP → N_1 + N_2 (muscle atrophy) ⇒ N_2 + OF + N_1 (atrophy of muscle) ⇒ N_2 + COMMA + N_1 (atrophy, muscle) ⇒ (N_2 → A) + N_1 (atrophic muscle) ⇒ (N_1 → A) + N_2 (muscular atrophy).

Two tree diagrams are required for the representation of cumulative derivations in an abbreviated form.



(muscle atrophy)

(muscular atrophy)



(atrophic muscle)

(atrophy of muscle)

(atrophy, muscle)

2. The problem of paraphrasing is closely related with automatic recognition of synonymous or nearly-synonymous expressions which are not found in the lexicon. Even if we would be able to assign an approximative value to so-called "unknown terms" by the implementation of deductive rules, the physician or the expert in the related field will have to make the decision about the synonymity of the term in question.

In both cases it is assumed that the semantic content of the message is understood. Otherwise, if the transformational rules were based only on the formal structure of noun phrases they could generate phrases which are not compatible with the semantic content of the message. Consequently, the transformational or paraphrasing rules should not be based only on syntactic features but also on semantic features of components of the given semantic unit, i.e., on the deep structure of the message.

3. The boundaries of the semantic unit do not correspond, necessarily, to the boundaries of the noun phrase. For this reason, we prefer the term "kernel phrase" instead of noun phrase. However, it seems reasonable to identify first the semantic correlations within the frame of the syntactic construction (noun phrase, adjectival phrase, verb phrase, etc.).

4. The noun phrase structure grammar comprises - among others - the analysis and synthesis of adjectival phrases and attributive noun phrases.

The adjectives were classified in terms of their semantic compatibility with semantic classes of head nouns. The following properties of adjectives and nouns were considered:

Adjectives:

a. Feature of transformability ($A \rightarrow N$; $A + ed \rightarrow N$; $A + ing \rightarrow N$)

b. Semantic class membership (Topography, Morphology, Etiology, Function, General)

c. Admissible semantic correlations with semantic classes of head nouns NPx $\rightarrow$ Ax + Nx where (x) denotes the type correlation.

Nouns:

a. Feature of transformability (N $\rightarrow$ N; N $\rightarrow$ A);

b. Semantic class membership (T,M,E,F,G as above)

For example:

ACUTE STAPHYLOCOCCAL PNEUMONIA

A-TR/G/M + A-TR/E/M + N-$\overline{\text{TR}}$/M $\rightarrow$ (NP) M,E where

A-TR/G/M (acute) = Adjective (A), transformable (TR), 'General' (G), semantic correlation with N(M);

A-TR/E/M = Adjective, transformable, semantic class membership-'Etiology' (E), modifying N(M);

N-$\overline{\text{TR}}$/M = noun, non-transformable ($\overline{\text{TR}}$), belonging to semantic category 'Morphology' (M).

(NP)M,E = adjectival kernel phrase belonging to the semantic category 'M' (morphology) and category 'E' (etiology).

The semantics assigned to adjectivals reflect, to a greater or lesser extent, the semantic and syntactic relations between the adjectivals and the head nouns that they modify. Rules were written for strings of adjectives which are preposed or

> postposed to the head noun as immediate constituents or in discontinuous sequence. Analogous criteria were used to prepare rules for recognition of attributive noun phrases. (13)

Z. Harris suggested that it might be possible to establish, what he calls, "classified relationships" between adjective and noun in advance, i.e., to have dictionary pairing of words and their classifiers for a scientific language because the metaphorical usage is highly restricted in scientific discourse. It seems reasonable that adjectival phrases or appositional noun phrases which are perceived as single conceptual units should be listed and treated in the same way as compounds.

Any treatment of English adjectives shows that the attribution and predication is far from simple. For instance, an adjective may modify a noun which is not present explicitly. An example is the phrase "parathyroid (gland) adenoma" meaning "parathyroid gland adenoma" where the adjective "parathyroid" is semantically related to the absent noun "gland."

The complexity of the analysis of adjectival phrases becomes apparent in attempting to semantically classify adjectives. Many adjectives may be members of two or more classes, i.e., the given A_{xy} can modify either N_x or N_y, e.g., general adjective 'abnormal' can be correlated semantically with N_T or N_M or N_E or N_F, thus, yielding multiple readings. Even further refinement of the classification of adjectives will not completely solve the difficulty.

Semantics

It was already mentioned before that our goal is the successful mechanical recognition, interpretation and subsequent storage of Medical English for meaningful and timely retrieval of medical data in accordance with the needs of the user.

When the boundaries of kernel phrases are identified by the implementation of syntactic and semantic algorithms the next step is the establishment of semantic correlations among the major semantic units of the utterance. The semantics of kernel phrases are conveyed by a set of relational predicates.

Relational predicate having a propositional function f(x) describes the type of semantic relationship among the major kernel phrases as mentioned before.

For example, the statement: "Pneumonia, due to staphylococcus" can be formalized as

$$\left[(R_1) \quad (M,E)\right]$$

where R_1 is the causative relational predicate "due to," "pneumonia" is the kernel phrase belonging to the semantic category 'M' (morphology) and "staphylococcus" being a member of the category 'E' (causative agent). The relational expression "due to" can be substituted by other equivalent expressions such as "caused by" or "resulting from," since they designate a similar relationship between the same semantic categories. The notion of argument pair (M,E) is not limited to single constituents as in the example above but it refers to kernel phrases in which the number of constituents is variable. For example, the statement mentioned above can be expanded as follows: "Acute and

severe pneumonia caused by staphylococcus albus" where the left-bound phrase "acute and severe pneumonia" (M) and the right-bound phrase "staphylococcus albus" (E) are related by the relational predicate "caused by" (R_1). Consequently, the statement $\left[(R_1) \quad (M,E)\right]$ holds in both cases.

The classification of relational predicates R_1, R_2 - - R_n will reveal different types of relations among the semantically marked kernel phrases belonging to the semantic categories T,M,E,F or perhaps to other categories which may result from further semantic analysis. The precision of relational predicate rules may be increased by refinement of admissible co-occurrences of subclasses of the basic semantic constituents of kernel noun phrases.

For example, statement: $\left[(R_x) \quad (M\ 12^{**},\ T\ 11^{**})\right]$ is to be interpreted that there is a certain relationship R_x (not defined here) between the subclass M12** (fractures) and subclass T 11** (bones) where ** implies any member of the respective subclass.

The syntactic structure of relational predicates consists either of a single functional element, such as the preposition 'in,' or compound expressions, such as 'due to,' or it can be a punctuation mark such as comma.

The semantic rules can be regarded as the axioms of the system and therefore theorems derived from these axioms will describe various properties of Medical English which can be tested for truth or falsity for completeness, and for ambiguity.

The body of theorems will constitute the grammar of a formal intermediate language having the semantic classes T, M, E, F and others to be defined, as its vocabulary.

The axioms of the system are based upon the distribution of semantically significant elements. The key components consist of operators and their operands. The operands, linguistically related to noun phrases, are members of particular semantic categories, and the operators are theorems describing the relationship among the operands.

Relationships among syntactically connected parts of sentences will be tabulated and associated with the meaning of semantic categories in the SNOP dictionary. The relationship between syntactic and semantic units is expressed by a set of linguistic operators which have the function of relational predicates in the intermediate language.

Conceptual analysis and preparation of algorithms for translating restricted Medical English into well-formed expressions will require definition of the functional form of relational predicates in the intermediate language and their syntactic features as linguistic operators in natural language.

We assume that heuristic-type rules for automatic problem-solving will be expanded beyond the present logical framework and combined with linguistic techniques for automatic processing of natural language. This will lead to the development of a more powerful theory of problem-solving, in general (15, 16).

Conclusions

The development of a methodology for machine encoding of diagnostic statements into a file, and the capability to retrieve information meaningfully from data file with a high degree of accuracy and completeness is the first phase towards the objective of processing general medical text.

The results of morphological, syntactic and semantic analysis are translated via computer programs into the existing intermediate language which comprises four information fields: Topography, Morphology, Etiology, and Function, as they are listed in the SNOP dictionary. Once the units are identified and mapped into the intermediate language, they could be replaced interlingually by equivalent units of another syntactic and semantic system. The replacement could be implemented mechanically through the selection of the equivalent syntactic structure, in our case, SNOP noun phrases in their intermediate language form.

The most amazing aspect of language is the fact that despite its enormous complexity human beings are able to use it with success as a communication tool. If we are ever able to discover and describe the process of human thought, we will be closer to the resolution of many problems associated with the formalization and subsequent automatization of natural language. It is not our intention to tackle all the problems inherent in natural language. We believe that we will be able to refine our algorithms and further develop a system which will

process medical text by applying the formalized linguistic analytic procedures for the storage of data in such a way that the users' requirements can be met.

SECTION 2: RESPIRATORY TRACT

20 - RESPIRATORY TRACT

2000 Respiratory tract, NOS

201 - Upper Respiratory Tract

2010 Upper respiratory tract, NOS
Nose, accessory sinus and nasopharynx combined sites

202 - Lower Respiratory Tract

2020 Lower respiratory tract, NOS
Larynx, trachea, bronchi and lungs combined sites

21 - NOSE

2100 Nose, NOS
2101 Mucous membrane of nose
2102 Respiratory region of nose
2103 Olfactory region of nose
2104 Nasal gland
2105 Olfactory gland
2106 Cavernous plexus of nose
2107 Lamina propria of nose
2108 Nasal meati

211 - External Nose

2110 External nose
2111 Roof of nose
2112 Dorsum of nose
2113 Apex of nose
2114 Ala nasi
2115 Nasal septum, mobile portion

212 - Nasal Cartilage

2120 Nasal cartilage
2121 Greater alar cartilage
2122 Lateral nasal cartilage
2123 Nasal septal cartilage
2124 Lesser alar cartilage
2125 Vomeronasal cartilage

213, 214 - Internal Nose

2130 Internal nose
2131 Nares
2132 Nasal vestibule
2133 Nasal fossae
2134 Nasal septum, NOS
2135 Choanae
2136 Nasal turbinate, NOS
2137 Inferior nasal turbinate
2138 Middle nasal turbinate
2139 Superior nasal turbinate

22 - ACCESSORY SINUS

2200 Accessory sinus, NOS
Paranasal sinus
Accessory nasal sinus
2201 Mucous membrane of accessory sinus
2202 Accessory sinus gland
2203 Lamina propria of accessory sinus

221 - Maxillary Sinus

2210 Maxillary sinus, NOS
Maxillary antrum
2211 Right maxillary sinus
2212 Left maxillary sinus
2213 Mucous membrane of maxillary sinus
2214 Maxillary sinus gland
2215 Lamina propria of maxillary sinus

222 - Frontal Sinus

2220 Frontal sinus, NOS
2221 Right frontal sinus
2222 Left frontal sinus
2223 Mucous membrane of frontal sinus
2224 Frontal sinus gland
2225 Lamina propria of frontal sinus

223 - Ethmoid Sinus

2230 Ethmoid sinus, NOS
Ethmoid antrum
2231 Right ethmoid sinus
2232 Left ethmoid sinus

References

1. DOSTERT, L.E.: Machine Translation and Automatic Language Processing: Vistas in Information Handling; Spartan Books; Washington 1963

2. GARVIN, P.L.: The Georgetown - IBM Experiment of 1954: An Evaluation in Retrospect: Papers in Linguistics in Honor of Leon Dostert; Ed. by W. M. Austin; Mouton Co.; The Hague, 1967

3. SEDELOW, S.Y., SEDELOW, W.A. JR.: Stylistic Analysis; Automated Language Processing; John Wiley and Sons, Inc.; New York 1967

4. CHOMSKY, N.: Aspects of The Theory of Syntax; MIT Press; Cambridge, Mass., 1965

5. Automatic Language Processing Advisory Committee, LANGUAGE and MACHINES - COMPUTERS IN TRANSLATION AND LINGUISTICS: National Academy of Sciences, Washington 1966

6. BOBROW, D.G.: Syntactic Theory in Computer Implementations; Automated Language Processing; John Wiley and Sons, Inc.; New York 1967

7. Seminar on Computational Linguistics; Edited by A. W. Pratt, A. H. Roberts and K. Lewis; U.S. Department of Health, Education and Welfare; National Institutes of Health; Public Health Service; Publication No. 1716; U.S. Printing Office 1967

8. BAR-HILLEL, Y.: Language and Information; Addison-Wesley, Reading, Massachusetts 1964

9. PRATT, A.W., THOMAS, L.B.: An Information Processing System for Pathology Data; Pathology Annual-66; Appleton-Century-Crofts Publishers; 1967

10. DUNHAM, G.: Pathology Diagnoses Language Encoder; Division of Computer Research and Technology; National Institutes of Health; Internal Report.

11. EPSTEIN, M.: A System for Pathology Data Processing; Division of Computer Research and Technology; National Institutes of Health; Bethesda, Maryland 1969

12. PRATT, A.W., PACAK, M.: Identification and Transformation of Terminal Morphemes in Medical English; Methods of Information in Medicine; Vol. 8; No. 2, April 1969; F. K. Schatauer Verlag, Stuttgart, New York

13. PACAK, M., DEFRANCESCO, H: Adjectival Phrases in Medical English; Division of Computer Research and Technology; National Institutes of Health; Bethesda, Maryland; in preparation.

14. Systematized Nomenclature of Pathology; College of American Pathologists, Chicago, Illinois 1965

15. GARVIN, P.L.: The Place of Heuristics in the Fulcrum Approach to Machine Translation; Lingua, Vol. 21; North-Holland Publishing Company; Holland, 1968

16. SLAGLE, J.R.: Experiments with a Deductive Question-Answering Program; Comm. ACM 8, 12, 1965