Extending a Natural Language Parser with UMLS Knowledge

Alexa T. McCray
National Library of Medicine
Bethesda, Maryland 20894

Abstract

Over the past several years our research efforts have been directed toward the identification of natural language processing methods and techniques for improving access to biomedical information stored in computerized form. To provide a testing ground for some of these ideas we have undertaken the development of SPECIALIST, a prototype system for parsing and accessing biomedical text. The system includes linguistic and biomedical knowledge. Linguistic knowledge involves rules and facts about the grammar of the language. Biomedical knowledge involves rules and facts about the domain of biomedicine. The UMLS™ knowledge sources, Meta-1™ and the Semantic Network, as well as the UMLS test collection, have recently contributed to the development of the SPECIALIST system.

Introduction

The SPECIALIST system [1-3] is being developed to include both linguistic and biomedical knowledge. Linguistic knowledge includes lexical information, and rules of morphology, syntax, and semantics. The lexicon, which forms a central part of the system, has general English lexical items as well as items specific to the domain of biomedicine. The lexicon currently contains approximately 40,000 lexical entries, which when expanded to the full set of inflectional variants, is actually over 75,000 lexical forms. Each lexical entry encodes morphologic, syntactic, and semantic information. This information is used by the grammar rules as they attempt to produce structured representations of phrases and sentences. Morphologic information encodes the inflectional and derivational variants of lexical items. Inflectional variants include singular and plural forms of nouns; positive, comparative and superlative forms of adjectives; and person, number, and tense of verbs. Derivational variants are part-of-speech alternations such as "treat" and "treatment", "aberrant" and "aberration", and "able" and "ability". Syntactic information includes information about syntactic category, allowable complements, positional information, and codes for allowable transformations. These latter codes are for rules that relate pairs of sentences to each other, e.g., active to passive sentences with the same basic meaning. The semantic information includes information for rules of logical interpretation; for example, rules that interpret ellipted or otherwise missing elements.

The biomedical knowledge needed by SPECIALIST includes knowledge of the important concepts in the domain of biomedicine, the relations among these concepts, and rules to process these concepts and relations. The two UMLS knowledge sources, Meta-1 and the Semantic Network, provide the sort of biomedical knowledge that we require. (See [4-5] for recent descriptions of the UMLS project). Meta-1 contains information about a large number of biomedical concepts that appear in several controlled vocabularies. It includes single and multi-word concepts, definitions, lexical category information, hierarchical contexts, and interrelationships among many of the concepts. Further, each concept in Meta-1 is assigned to at least one of the basic semantic types, or categories, included in the Semantic Network. The network consists of 131 semantic types and 35 relationships between them.

In order to test the extent to which natural language processing techniques may improve access to information, we have developed a database module. The module processes files such as MEDLINE® citation records, creates an index for the items in all relevant fields, and provides for Boolean retrieval of these items. One of our major sources of textual material for the system is the UMLS test collection of queries and citation records [6].* The data for the test collection were selected primarily from 2,000 search request forms submitted to the NIH and NLM libraries. We chose 155 questions in three broad topic areas: clinical medicine research, basic science research, and health services research. The questions were searched on a subset of MEDLINE. The subset totals 167,000 citations and is that portion of the citations that were added to MEDLINE with a 1986 publication date and which included an abstract. The searches were con-

*The collection was developed for use as an evaluation tool in the UMLS project. It is available to interested researchers. The collection has also been included in Virginia Disc One, one of 3 CD-ROM's under development to illustrate state of the art methods in information retrieval.

0195-4210/91/$5.00

ducted by an expert NLM searcher whose search strategy emphasized recall over precision. The retrieval results were then evaluated for relevancy by a subject matter expert.

The SPECIALIST System

Linguistic Knowledge

As noted above, the lexicon is a central part of the SPECIALIST system (see [3] for some discussion of this point). Lexical records are expressed as frames with slots and values. The required slots for any entry are 'base', 'cat', and 'variants'. The base form is what is often called the 'citation' form of the lexical item. The value of the slot 'cat' gives the syntactic category of the lexical item. The value of the variants slot is either a list of codes, such as 'inv', (invariant) 'reg' (regular), or 'glreg' (greco-latin regular), or in the case of irregular forms, as in the word *bad*, whose record is shown below, the actual variant forms are listed.

```
{base=bad
entry=1
        cat=adj
        variants=irreg(worse,worst)
        position=attrib(1)
        position=pred
        position=attribc
        compl=infcomp:nsr
        nominalization=badness}
```

Since this adjective may both premodify a noun and be used as a predicate adjective (e.g., "This is a bad situation", and "The situation has gotten quite bad") it is given the codes 'attrib' and 'pred' respectively. This is in distinction to an adjective like "main" which may only appear in attributive position (e.g., "the main problem"). The '1' in the code attrib(1) means that this is an adjective of quality. The adjective 'bad' may also take a discontinuous infinitival complement, as in "That was a particularly bad problem to solve." Since all of this information is coded in the lexical entry, the parser checks the entry at parse time and then invokes the appropriate syntactic and semantic rules based on the coding. Coding for verbs is extensive. A verb may be considered the control center of a sentence, since it determines the type of complements that may co-occur with it. As a parse is built, the system checks the lexicon to see whether, for example, a verb is transitive, and if so, whether it takes a simple noun phrase object, a prepositional phrase object, or some type of clausal object.

In building the lexicon we have taken the approach of coding all available grammatical information about a lexical item. We refer to a variety of general English and medical dictionaries to accomplish this. In addition, when coding lexical items, we consult sets of actual sentences containing those items. These sentences are drawn primarily from the UMLS test collection. Lexical items are entered using our menu-based, interactive program, called Lextool, which accepts as input either a file of lexical items or lexical items typed in from the keyboard. With the interactive aid of the user, it gives as output fully specified lexical frames. Lextool incorporates rules that dictate which slots are permissible for the frame in question, as well as what values they may have. These rules have been formalized in a grammar which includes all the allowable slots and values. The grammar serves to constrain the possible choices that must be considered when entering an item, and it also serves as an automatic check of the well-formedness of completed lexical records.

The syntactic/semantic component of SPECIALIST is an extended Definite Clause Grammar.* The grammar includes context-free BNF (phrase structure) rules together with context-sensitive restrictions which constrain the structures that are actually built. An example parse is reproduced below. The sentence processed is: "This drug treats the most severe symptoms of congestive heart failure."

```
OPS:    present
VERB:   treat
SUBJ:   this drug (sing)
OBJ:    the symptom (pl)
        L_MOD: adj:  severe (quality)
                     MOD:  intensifier: most
        R_MOD: pp:   of
                     obj: congestive^heart^failure
                          (sing)
```

The parser has first looked up all lexical items in the SPECIALIST lexicon, and then it has used the syntactic and semantic information associated with each lexical item, together with the appropriate rules, to construct the structured representation shown above. The operator in this sentence is the present tense. The verb and subject are represented in their base forms with accompanying inflectional information. The object of the verb is a noun phrase which has further internal structure. The head noun "symptom" is premodified by the adjective phrase "most severe", and it is

*See [3] and the references cited there for a description of this formalism. Since the time of that paper, which described the system we were developing, we awarded a contract to the Paoli Research Center of the Unisys Corporation. As a result of this successful collaboration of our two research groups during the academic year 1988-1989, the syntactic component of the system is extremely robust. See [7-8] for discussion of the Paoli system.

postmodified by a prepositional phrase. Note that "congestive heart failure" is considered a single, multi-word, lexical item here. This is because the lexicon treats this term as a single concept, in accordance with its use by the biomedical community.

We are in the process of augmenting our lexicon with the Meta-1 vocabulary. This has raised a number of methodological issues. There are over 43,000 reviewed main concepts in Meta-1 together with a large number of lexical variations for these concepts. For example, there are some 4,000 singular forms related to plural main concepts and some 8,000 plural forms related to singular main concepts. Many of these singulars and plurals were originally generated algorithmically. These variants are additional entry points into Meta-1, but in some cases they represent overgenerations, since they have not been reviewed. The variations that we can, however, include directly in our lexicon are those that were reviewed by the developers of the various vocabularies. These include irregular variations such as "louse" and "lice", "dermatitis" and "dermatitides", and "exanthema" and "exanthemata".

There is additional lexical information included in Meta-1 that is of interest for natural language systems. Some 2,800 terms are marked as acronyms (e.g, GABA, LAV-HTLV-III, RNA) or as having embedded acronyms (e.g., Amino Acid-Specific tRNA). Some 1,000 terms are marked as eponyms (e.g, Charcot's arthropathy, Barr Bodies, Wallerian Degeneration), and another 3,000 are marked as being trade names (e.g., Adicin, Fanasil, Motrin). The acronyms are generally linked as synonyms to their fully expanded forms, and the eponyms and trade names are also often linked to their non-eponymic forms and generic forms, respectively. The total number of reviewed synonyms found in Meta-1 is approximately 15,000.

There are about 17,000 single word Meta-1 reviewed concepts and about 26,000 reviewed multi-word concepts. The multi-word concepts present an interesting challenge. We would like to be able to include all multi-word concepts in the lexicon so that terms that are in common use in biomedical writing are available to the parser, but in some cases the multi-word terminology represented in Meta-1 is such that it would be unlikely to appear in natural, written text. The issue here is the following. Multi-word concepts which are in common use usually represent meanings that are not the sum of their constituent parts. Thus, even though, for example, "congestive heart failure" appears to be transparent in meaning, it has a very specific clinical meaning. A parsing system should recognize this as a single lexical item and not attempt to decompose it further. On the other hand, many of the multi-word terms in Meta-1 which are not in common use and, thus, would not be added directly to the lexicon, do have constituent parts which would be found in biomedical texts.

The approach we have taken to the augmentation of our lexicon is to generate partial lexical templates based on the information found in Meta-1. These templates are then reviewed and edited online. For many multi-word terms this means that they are decomposed if the entire phrase cannot be found in a medical dictionary. For example, the Meta-1 term "whooping cough due to bordetella pertussis" is entered as two lexical items, "whooping cough" and "bordetella pertussis". We have added approximately 7,000 lexical items from Meta-1 to our lexicon using this method. We will continue this work, but in the interim, we intend to use the rest of the terminology (including the 35,000 unreviewed main concepts and their variants) as a source of partial lexical information. That is, if the parser lexical look up routines are unable to find an item in the lexicon, they next look in Meta-1 to see if the item is there. If so, and if the item is a noun, certain agreement restrictions are relaxed, and a full parse can be built.

Biomedical Knowledge

An important component of a natural language understanding system is knowledge of the relevant domain. We originally evaluated the MeSH® structured vocabulary as a source of such knowledge for our parsing system. In order to provide us with a computational environment for exploring the MeSH structure, we developed Meshtool, a MeSH browser that runs on the Sun. It runs in both batch and interactive mode and, like the rest of the SPECIALIST system, is implemented in Prolog and C. Work on the MeSH file has led to the development of a suite of C functions for fast and efficient indexing using the prefix B-tree algorithm. We have built several optimizing features into this library, including a multi-user environment with mutually exclusive writes and a dynamic delete capability. We have profitably used this indexing approach in many of our applications, including the project lexicon.

Recognizing that one of the limitations of the MeSH vocabulary as a source of domain knowledge was its lack of explicit relationships between the terms in the structure, we developed Meshlink, an application to help domain experts make these relationships explicit. The domain expert is presented with a child and parent MeSH pair and is then prompted to choose from the available set of relationships. Meshlink was first used on an experimental basis by several visiting medical students. It was subsequently refined and used by our

UMLS collaborators at the University of Pittsburgh and Yale to label major portions of the MeSH hierarchy for inclusion in Meta-1. Over 9,000 of the approximately 16,000 MeSH terms have been labelled (including MeSH sub-trees for anatomy, diseases, physiology, genetics, and psychiatry and psychology).

The MeSH vocabulary represents a significant portion of the Meta-1 terminology. Value has been added to the vocabulary as it appears in Meta-1 by the addition of interterm relationships, by linking the vocabulary to other biomedical nomenclatures such as SNOMED, ICD, and CPT, and by assigning semantic types to each of the concepts. In order to make the Meta-1 content accessible to the rest of the SPECIALIST system, we have undertaken to develop a browser for Meta-1 which is similar in functionality to Meshtool. The browser is implemented in C only and runs on Sun workstations. In its current form the browser serves as an online reference tool in the SPECIALIST environment. We are working on an application that will allow the parser to reason with the information in Meta-1 and the Semantic Network. The top-level menu of the current browser is shown below.

Meta-1 Concept Retrieval System
---

a. Concept Name
b. Concept Definition
c. Lexical Variants
d. Lexical Tags
e. Syntactic Categories
f. Synonyms
g. Semantic Types
h. Related Concepts
i. Associated Expressions
j. Contexts
k. Global Search
l. Help
m. History
n. Exit

The application allows users (or programs) to search for Meta-1 terminology, reporting the term and its source vocabulary; its definition, lexical tags and variants; its synonyms, semantic types, related terms, other associated terms, or contexts as specified by the user. The global search capability allows the user to find all concepts in Meta-1 with a particular characteristic; e.g., all concepts that have a particular semantic type, or all concepts that are labelled as acronyms, etc. Sample output for some queries about "Parkinson disease" is shown below. [QT = Query Term, CN = Concept Name, SY = Synonym, VOC = Vocabulary, STY = Semantic Type].

Query Synonyms [return to quit]: Parkinson disease

QT: Parkinson disease
CN: Parkinson Disease
SY: Shaking palsy
SY: Paralysis Agitans

Query Ancestors [return to quit]: Parkinson disease

QT: Parkinson disease
CN: Parkinson Disease
VOC: MeSH

Diseases (Non MeSH) [C]
Nervous System Diseases [C10]
Central Nervous System Diseases [C10.228]
Brain Diseases [C10.228.140]
Basal Ganglia Diseases [C10.228.140.79]
Parkinson Disease [C10.228.140.79.804]

Query Semantic Types [return to quit]: Parkinson disease

QT: Parkinson disease
CN: Parkinson Disease
STY: Disease or Syndrome

We have carried out initial tests of the feasibility of using the UMLS semantic types for expressing selectional restrictions. Selectional restrictions establish what may sensibly co-occur with an item. For example, a verb such as "administer" takes an agent as a subject and may take a therapeutic substance as one object and a body region as a second object. This is illustrated by the following sentence from the test collection: "Nitrogylcerin ointment (12.5 to 50 mg) was administered in randomized fashion to three skin sites..." Here the agent is understood (someone); the therapeutic substance is "nitrogylcerin ointment"; and the body region is "three skin sites". The use of selectional restrictions such as these can help reduce the number of spurious parses that are generated by a parser that has only grammatical information, and it can give an indication of the meaning of the major concepts in a sentence.

We are taking two approaches to our initial investigations. The first involves an analysis of highly frequent verbs and their nominalizations as they occur in the test collection. The sentences containing these verbs are analyzed and their complements are studied to see if a match can be made to an existing semantic type. We have studied approximately thirty verbs to date. The second approach has involved identifying the semantic types of all the nouns that are in our current

lexicon and that are also in Meta-1. At last count this was over 10,000. We have added an additional look up step to our parser, so that if a semantic type exists for any of the lexical items in the sentence, it is reported as part of the final parse. This allows us to see the semantic types in context as part of our normal development work. A simple example illustrates. The sentence parsed is "Penicillin treats endocarditis".

OPS: present
VERB: treat
SUBJ: penicillin (sing, (Pharmacologic Substance..))
OBJ: endocarditis (sing, (Disease or Syndrome))

This sentence serves, too, to illustrate the kind of selectional restrictions one would want to establish for the verb "treat". This verb generally takes an agent (either human or pharmacologic) as a subject and a disorder or individual as an object.

Evaluation

Evaluation is an ongoing concern for us in the development of the parser. On a regular basis we produce new "releases" for use by the project group. With each new release we run a test suite of sentences against the parser looking for changes and improvements. Occasionally we uncover an error that has been introduced since the last release, but more often we see a change in the number or quality of the parses produced. The lexical information available in Meta-1 has allowed us to add a large number of biomedical terms to the lexicon in a relatively short time. This will continue to lead to improvements in the coverage of the parser. We are hopeful that the use of the UMLS semantic types in establishing selectional and other co-occurrence restrictions will improve the parsing capability of the system. The test suite will serve as our benchmark.

Since a primary goal of our work is to develop methods for improved access to biomedical information, we have, as noted above, implemented a database retrieval facility as part of the SPECIALIST system. The purpose of this facility is to allow us to test and evaluate the role of natural language processing techniques in information retrieval. The current implementation of the facility allows for Boolean retrieval of index terms. We are currently designing experiments that will assess the value of parsing queries and free text in titles and abstracts compared with standard retrieval using index terms only. The role of the UMLS knowledge sources is both indirect (in improving the overall quality of the parser) and direct (in providing additional search terminology). After a query has been parsed, its major noun phrases are identified. At this point the phrase and any of its grammatical permutations are added to the list of possible search terms.

The incorporation of UMLS knowledge in a natural language processing system serves, too, as a test of the correctness and adequacy of that knowledge in the context of a fairly complex application.

Acknowledgements

I would like to acknowledge the work of Alan Aronson, Allen Browne, Brandon Brylawski, Amir Razi, Suresh Srinivasan, and Theresa Waldspurger. They have contributed in many ways to the SPECIALIST project.

References

[1] Alexa T. McCray and Suresh Srinivasan, Automated Access to a Large Medical Dictionary: Online Assistance for Research and Application in Natural Language Processing, Computers and Biomedical Research, Vol. 23, No. 2, 179-198 (1990).

[2] Alexa T. McCray, Allen C. Browne, and Dorothy L. Moore. The Semantic Structure of Neo-Classical Compounds, in: Proceedings of the Twelfth Annual SCAMC, ed. Robert A. Greenes, pp. 165-168 (IEEE Computer Society, Los Angeles, 1988).

[3] Alexa T. McCray, Jeffrey L. Sponsler, Brandon Brylawski, and Allen C. Browne; The Role of Lexical Knowledge in Biomedical Text Understanding, in: Proceedings of the Eleventh Annual SCAMC, ed. W. Stead, pp. 103-107 (IEEE Computer Society Press, 1987).

[4] Donald A. B. Lindberg and Betsy L. Humphreys, The UMLS Knowledge Sources: Tools for Building Better User Interfaces, in: Proceedings of the 14th Annual SCAMC, ed. R.A. Miller, pp. 121-125 (IEEE Computer Society Press, 1990).

[5] Alexa T. McCray and William T. Hole, The Scope and Structure of the First Version of the UMLS Semantic Network, in: Proceedings of the 14th Annual SCAMC, ed. R.A. Miller, pp. 126-130 (IEEE Computer Society Press, 1990).

[6] Peri L. Schuyler, Alexa T. McCray, and Harold M. Schoolman, A Test Collection for Experimentation in Bibliographic Retrieval, in: Barber, B., Cao, D., Qin, D., Wagner, G; eds. MEDINFO 89, Amsterdam: North-Holland, 1989; 910-912.

[7] Lynette Hirschman, Conjunction in Meta-Restriction Grammar, Journal of Logic Programming 3, (1986) 299-328.

[8] M.S. Palmer, D. Dahl, R. Passonneau, L. Hirschman, M. Linebarger, J. Dowding, Recovering Implicit Information, in: Proceedings of the 24th annual meeting of the Association for Computational Linguistics, 1986, pp. 10-19.