Database Capture of Natural Language Echocardiographic Reports: A Unified Medical Language System Approach

*K. Canfield, **B. Bray, *S. Huff, *H. Warner
*Department of Medical Informatics
**Division of Cardiology
University of Utah, School of Medicine
Salt Lake City, UT

Abstract

We describe a prototype system for semi-automatic database capture of free-text echocardiography reports. The system is very simple and uses a Unified Medical Language System compatible architecture. We use this system and a large body of texts to create a patient database and develop a comprehensive hierarchical dictionary for echocardiography.

Introduction

A major goal of the Unified Medical Language System (UMLS) project is the development of a comprehensive relational database of medical terms. This database would allow a unified definition of medical concepts. It would supply the facts and relationships needed for natural language processing, translation of terms and data across systems, remote database access, complex inference for medical information and indexing of medical literature.

This paper describes a prototype system that is used to create a structured database of findings from free-text reports of echocardiographic image interpretations. The prototype is designed for compatibility with the UML Metathesaurus version 1, which is currently under development [1]. Our system has a similar relational structure to the UMLS metathesaurus, but with a more comprehensive treatment of a small domain. This investigation explores the possibilities for application of the UMLS concept to a patient database.

We will describe the system file structure, content, procedures, and performance on unedited texts. The system is called ECHODB and is implemented using a relational database on networked microcomputers.

Methods

The system prototype is described below.

File Structure

A table called LEX is the front end lexicon to the system. It has the following columns:

num	a unique ID for every term
text	the text for the term
alias	a marker for processing synonyms, and creating canonical terms
ot	object type - a marker for semantic class

Example:

num	1
text	left
alias	103
ot	6

Several other tables are needed to process the alias marker and are not described here. These include terms that have a one to many and a many to one relationship with the single word terms in the lexicon. The code 103 marks "left" as a term that participates in multiword combinations such as "left to right". Morphological variations are handled using a method from a program by David Evans' group at Carnegie-Mellon [2]. The input to the procedures connected with this table is output as a list of canonical atomic terms used in the dictionary text.

The table called DIX contains the dictionary for the system. It has the following columns:

num	link to LEX
num2	unique ID for dictionary terms
status	mark for hierarchical reasoning
text	text for dictionary entry
hstring	hierarchical code for dix entry
ot	semantic marker for object class
level	marker for depth in hierarchy

Example:

num	4
num2	3
status	t
text	left ventricle wall
hstring	0101010101
ot	1
level	5

The dictionary is hierarchical. The table called PTDB contains the patient data findings. It has the following columns:

texid	code for identifying text
ptid	link to demographic table for patients

anatomy	link to DIX (num2)
location	link to LEX (num)
pathology	link to DIX
pmodifier	link to LEX
severity	link to DIX
smodifier	link to LEX
procedure	link to DIX
probability	link to DIX
exist	flag for positive or negative finding

Example:

texid	t1
ptid	p1
anatomy	39 (mitral valve)
pathology	38 (insufficiency jet)
severity	40 (mild)
procedure	77 (doppler flow image)
probability	nl (null)
exist	1 (positive finding)

The relational columns can be thought of as slot names in a frame data structure. The procedures for this table treat the finding slots as if they had the following frame structure:

anatomy
    location
pathology (structure or function)
    pmodifier
severity
    smodifier
procedure
probability.
exist

These finding slots are based loosely on the SNOMED multi-axial model [3]. The parsing (understanding) procedure looks for the lower level slots in conjunction with their higher level parent. For example, "locations" are processed at the same time as "anatomies". The pathology slot contains the information from either the structure hierarchy or the function hierarchy (ot=2). Each finding frame represents all the information in one diagnostic finding clause. There may be multiple frames per sentence in a text.

All these tables reside in a single database. The hierarchy in the DIX table is detailed below.

Dictionary Hierarchy

The dictionary hierarchy is the core of our system and similar to the semantic network of the UMLS project. It supplies the knowledge that drives parsing. Each of the finding slots has a hierarchy that is labeled with the semantic marker - ot (object type). The following is an abbreviated example from the text field of the anatomy hierarchy:

.
.
.
cardiac valves
    mitral valve
        mitral valve annulus
        mitral valve orifice
        mitral valve leaflet
        mitral valve chordae
    aortic valve
        .
        .
        .

The hierarchical dictionary term is linked to LEX (the column "num") by the unique term in the string. For example, "mitral valve leaflet" is labelled with "leaflet" because it is the unique term down the hierarchy at that point. The marker called - status - contains information about whether the dictionary term is terminal or non-terminal. The uses of the various markers will become clear below in examples of the parsing process. The system currently contains the following object types (ot):

anatomy	1
pathology	2
severity	3
procedure	4
probability	5
location	6
clause marker	7
smodifier	8
pmodifier	9

Relationship to UMLS Meta-1

ECHODB has a similar relational structure to Meta-1 [6]. Some of the fields overlap as follows:

UMLS	ECHODB
semantic type	object type
concept name	text
contexts	hcode
synonyms	alias

There are some obvious differences. ECHODB requires a detailed hierarchy while Meta-1 has an incomplete one. ECHODB also requires fields that Meta-1 does not offer. Although the current version of Meta-1 does not have a detailed hierarchy, work under UMLS is addressing this issue [4].

These connections between ECHODB and Meta-1 allow many advantages to both. The carefully constructed ECHODB dictionary for echocardiography can be added to Meta-n because of the similar relational structure. Meta-1 can be used as a resource to create other limited domain comprehensive systems for the same reason. This structural similarity allows incorporation and/or transformation of columns (slots) between systems.

Parsing Process

The parsing process takes a sentence from a text and transforms it into one or more finding frames that it stores in the PTDB. The process is best described in terms of a detailed example of the database capture method. It consists of the following major steps:

1. preprocessing
2. slot filling procedure
3. heuristics for discourse understanding
4. saving a database record

Sample lines from an echocardiographic text:

LV wall motion is abnormal with significant septal hypokinesis. Posterior, lateral and apical segments appear to function normally.

Preprocessing includes:

1. mapping "LV" to "left ventricle"
2. morphological analysis to canonical terms
3. clause analysis

The first two items are accomplished in a straight-forward manner using the alias marker and the morphological program mentioned above. The third uses a simple heuristic. The parser looks for lexical clause markers such as: comma, period, "and", "or", "with". For example, the first line would parse into two arrays as follows:

#1	#2
left	significant
ventricle	septal
wall	hypokinesis
motion	
abnormal	

The first parse would result in a single array including "with" in position 6 and the period in position 10. This allows the split into #1 and #2. The clause heuristic simply starts a new array after each defined clause marker. This method is simplistic and may have to be modified in the face of a larger sample size of texts. The parser then concentrates on array #1. All anatomy terms (ot=1 and including location ot=6) are tested against the dictionary as follows:

slot filling procedure:

1. take the lowest numbered level (this is the highest hierarchical level) of the group - in this case "ventricle")
2. check remaining words for inclusion in the context subtree (defined below)
3. take the highest numbered level (lowest hierarchical level)
4. select dictionary term if step3 exists else put on wait stack
5. check for modifiers (in this case- location)
6. check status marker for terminal terms
7. save anatomy slot if step6=true else put on wait stack

This procedure tells us that "left ventricle wall" is indeed a valid slot filler for anatomy. This process is repeated for the pathology slot and the procedure slot. Pathology is set to "abnormal motion" and procedure is set to the default. The system then checks to see if we have a valid frame for an echo finding. This is defined as a frame containing an anatomy and a pathology. Since the frame includes a valid anatomy and a valid pathology, the frame flag=OK and the frame is saved.

The context subtree is defined differently for each object type. For example, the anatomy context subtree is at the level of "left ventricle". This system uses the numerical level of a term as a way of defining the level of granularity across terms in an object type tree. This is detailed in the discussion section.

The parser then moves to array #2 using the slot filling procedure. "septal" is a location with no corresponding anatomy term. It is put on a wait stack and fix procedure is called. This procedure goes back to the previous context subtree and gets "left ventricle wall". In this way it finds "left ventricle septal wall" as a terminal anatomy term. "hypokinesis" is added as a valid pathology with "significant" as a severity. This results in the following two findings in PTDB. Unfilled (null) slots are not shown.

#1
anatomy=left ventricle wall
pathology=motion
severity=abnormal
procedure=default

#2
anatomy=left ventricle wall
location=septal
pathology=hypokinesis
severity=significant
procedure=default

The parser proceeds to the next sentence and creates the following arrays:

#3	#4	#5
posterior	lateral	apical
		segments
		function
		normal

This sentence uses the slot filler procedure with the following discourse heuristics:

1. inside then outside the sentence
2. forward then backward within the sentence

The procedure moves through the entire sentence without finding a valid terminal anatomy. Since inside the sentence failed it goes outside to array #2 (the previous sentence). It uses that context subtree to create "left ventricle posterior wall segment" for the anatomy slot. This entails searching below the level of "left ventricle" in its subtree. This context tree propagates forward for selection of anatomy in #4 and #5.

Arrays #3 and #4 have no pathology terms. The valid pathology in #5 propagates backward to fill in #3 and #4 (heuristic number 2). This leaves us with three more valid frames:

#3
anatomy=left ventricle wall
location=posterior
pathology=function
severity=normal

#4
anatomy=left ventricle wall
location=lateral
pathology=function
severity=normal

#5
anatomy=left ventricle wall
location=apical
pathology=function
severity=normal

The two simple discourse heuristics handled almost all the interpretation problems in our sample. Linguistic reasons for this are given in the discussion section. Detailed results of the parsing process are given in the results section.

Performance

In order to test this database capture method, we took five texts at random from the large number of texts and refined the procedure by solving the problems they generated. Then, we

took five more of the texts at random and let the system try to handle them. We continue this incremental process until the dictionary is comprehensive. All the texts were generated by one person before this project was ever conceived. We may run into additional problems later due to idiolectal differences between cardiologists. An example text is given in the appendix.

Results

The method discussed above is used to refine our database capture system and create a dictionary. We have completed the following steps:

1. develop basic method
2. use 5 texts to refine the method
3. use 5 additional texts to test progress
4. use 5 additional texts to verify progress

Steps 1 and 2

We developed the understanding system by addressing the dictionary and parsing problems in the texts in a general way. The emphasis was on the general dictionary development and discourse problems. With the general method in hand we systematically improved the system to handle all the problems in a 5 text randomly selected corpus. The problems were of two major types:

1. dictionary deficits
2. exceptions to the heuristic interpretation rules

Dictionary deficits are the most easily solved problems. We add the terms to the lexicon and the dictionary. One such problem that arose was the term "upper limit" or "lower limit" (of normal). These terms were added to the dictionary and placed in the lexicon as complex terms. Complex terms are searched for as a unit. For example, if the term "upper" is hit, the co-term "limit" is searched for. If "limit" is not found, the term "upper" is considered to be atomic (as a location).

Quantifier-like terms were another modifier problem. For example, "All cardiac chambers" or "other cardiac chambers". These were solved by adding the terms directly to the dictionary as locations. We plan to make the meaning explicit with "word experts" or procedures that define the terms. For example, "other cardiac chambers" would call a special procedure that determines what the "others" are.

An exception to a heuristic interpretation rule is shown with "Pulmonary valve appears normal". In this case, the rule would be to go back to the previous context subtree and get an acceptable (terminal) pathology. This would not be correct because the "normal" here refers to a global normal for structure and/or function. We solved this problem by defining a special case where if "normal" appears by itself in the context of a sentence, it can be terminal and refer to all applicable pathologies.

Parses #3-#5 in the example in the last section show a problem. The pathology=function obscures the fact that a specific function of "motion" is most certainly intended. One suggested fix would be to create an index of anatomy to possible functions and infer the correct one from other textual clues. The general context subtree procedure would not be robust enough for this case.

Another failure of the heuristics is seen in the sentence: "(1)LV chamber dimensions are normal with (2)normal segmental and (3)global wall motion." The heuristic "forward then backward" fails here for the pathology in clause 2. This is due to the system making no distinction between types of coordination. The system parses clause 2 as: LV wall size normal. There are two approaches to fixing this problem. One is syntactic analysis for type of coordination. The other is adding semantic knowledge. We have chosen the later approach. We are adding modular knowledge that can override the heuristics. For example, in the case above, before filling the pathology slot in clause 2, the system would check a file that listed allowable relationships between anatomies and pathologies. The heuristic "forward then backward" would be overridden and "motion" would be propagated backwards.

Errors such as these constituted 11/49 sentences in our first 5 text sample. An error was counted for any mistaken group of records for a sentence. A mistake in any clause disqualifies the whole sentence. This is a 23% error rate.

Step 3

The next 5 texts were analyzed to test our system. Almost all the errors were of the simple dictionary deficit type. For example, "right-sided valves" need to be added to the dictionary. This improved the error rate to 14% (6/43 sentences). Most of the errors (5/6) were dictionary problems.

A more complex error occurred with "The IVC was poorly visualized due to a large polycystic liver present in the abdomen". We have chosen to ignore such findings at this time because they bring in outside concepts. Our system will handle such findings through an interactive user interface. ECHODB outputs the records in an easily edited ASCII format. A cardiologist reviews and edits the file before it is automatically incorporated into the patient data file.

The trend in the error rate is downward (23% to 14%) and more importantly, the interpretation problems are reducing even more quickly. The dictionary deficit problems were only half of the errors in step 2 (5/11) while they were the majority of the problem in step 3 (5/6). It seems that the number of problems are declining and that the remaining problems are becoming simpler. A larger sample of texts is needed to confirm this trend.

Step 4

An additional 5 texts also showed a 14% error rate (6/41). All of the errors were dictionary deficit type except one involving scope of negatives. This sentence "The aortic valve shows minimal sclerotic change, but no significant stenosis or insufficiency." missed the implied negative in clause 3. We fixed this by expanding an ad hoc rule that applies to clauses conjoined with "or".

If we change the stringent way of reporting errors, the numbers improve. If we look at proportion of correct clauses (records as opposed to sentences) the error rate is 12%. If we look at the proportion of correct slots (fields) the error rate is 5%. We are continuing to build the dictionary by analyzing increments of n texts where n is increasing as the error rate drops.

Discussion

The nature of these echocardiographic texts is in large measure responsible for the success of this system. The medical terminology of this area is somewhat narrow with little outside knowledge required. This would not be true of history findings for example. The task that generates the texts is consistent and well formed since the cardiologist interprets findings using a predictable routine.

Several practical advantages of this system are noted. The use of free text input of echocardiography findings minimizes the restrictions on language use by the physician and gives access to a large amount of data previously recorded in text format. The structured database format allows easy data retrieval for clinical research and teaching purposes. The system permits a unified structure for data capture which enhances data transfer and pooling from different laboratories.

The basis for this system is a context free semantic grammar. Other strategies are used, however, in the spirit of a flexible, multi-strategy system [5]. Construction specific pattern matching and domain specific rules are used. The interpretation heuristics of "inside then outside" and "forward then backward" capture some basic facts about discourse pragmatics. They include the fact that anatomy terms almost always propagate forward and pathology terms frequently propagate backward. This is because the anatomy terms are often subject or topic to the predicate-like pathology.

The efficacy of the semantic grammar is enhanced by our well behaved notions of context and focus. Context is strictly defined in terms of subtrees within an object type. These subtree levels provide a way of representing level of abstraction or granularity. For example, the dictionary makes the claim that "abnormal wall motion" and "abnormal cardiac flow" are at the same level of abstraction. This adds complexity to the dictionary development process. Focus is provided by the context mechanism combined with the marker processing. The markers - alias, ot, status - serve to mark nodes in the hierarchy for differential processing. Modular knowledge base tables provide semantic information that can override simple heuristics.

The absence of syntactic information processing limits the system performance. We did not include any (save for our simple clause rule) because it increases computational complexity and development time. Our main concern is to create a dictionary. Syntactic analysis can be added to the system as a layer of processing. Syntax can answer two kinds of questions that a semantic grammar has trouble with:

1. clause analysis
2. putting the correct modifier with the correct head

If we run across a sentence where the simple clause rule does not work, we will need to use more syntactic information or allow user interaction. Sentences with semantically unresolvable head-modifier relations require syntactic analysis.

We are in the process of adding an interactive interface to handle uninterpretable findings and adapt the parsing process to a query system for the resulting patient database. This approach emphasizes the system as a tool for rapid creation of a database from texts. It is not meant to provide complete and error free performance in its task.

Conclusion

The ECHODB system provides a simple method of database capture with an acceptable level of performance. It uses easily implemented database and linguistic techniques to create a structured patient database for echocardiography. Together with an interactive user interface to handle spurious or failed parses, this system can speed the creation of the database and provide a query option for the final product. An additional benefit of this project was to produce a comprehensive dictionary for echocardiography that was developed from actual clinical reports. ECHODB shows a UMLS approach to medical concept representation allowing flexible input to a structured database.

Appendix

A sample fragment of an echo report:

Left ventricular chamber dimensions are increased. LV wall motion is abnormal with significant septal hypokinesis. Posterior, lateral and apical segments appear to function normally with an estimated global ejection fraction in the lower limits of normal. Left atrium is severely enlarged. The right atrium and ventricle are at the upper limits of normal. The aortic valve is grossly normal in structure. Doppler flow images demonstrate the presence of a moderate-sized aortic insufficiency jet. The mitral valve demonstrates mild thickening, primarily involving the posterior annulus and leaflet with fixed posterior leaflet and mild doming of the anterior leaflet. The mitral valve orifice area appears to be only mildly reduced. There is no significant increased gradient across the mitral valve. [...]

Acknowledgments

This work was supported in part by NLM contract award number: N01-LM-8-3515.

References

[1] Tuttle M.S. et. al., Toward a Bio-Medical Thesaurus: Building the Foundation of the UMLS, Proceedings SCAMC, 1988.

[2] Evans D.A., The MedSORT-II Project, CMU-LCL-87-3 1987.

[3] SNOMED, Systemized Nomenclature of Medicine Introduction, Cote R. A. ed., College of American Pathologists 1979.

[4] Evans D.A., Pragmatically-Structured, Lexical-Semantic Knowledge Bases for Unified Medical Language Systems, Proceedings SCAMC, 1988.

[5] Carbonell J.G. and Hayes P.J., Robust Parsing Using Multiple Construction-Specific Strategies, in Bolc L. ed., Natural Language Parsing Systems, Springer-Verlag 1987.

[6] Personal Communication 1988, Mark Tuttle.