Validation of Clinical Problems Using A UMLS-Based Semantic Parser

Howard S Goldberg MD, Charles Hsu, Vincent Law, Charles Safran MD, MS
Center for Clinical Computing, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA

The capture and symbolization of data from the clinical problem list facilitates the creation of high-fidelity patient resumes for use in aggregate analysis and decision support. We report on the development of a UMLS-based semantic parser and present a preliminary evaluation of the parser in the recognition and validation of disease-related clinical problems. We randomly sampled 20% of the 26,858 unique non-dictionary clinical problems entered into OMR (Online Medical Record) between 1989 and August, 1997, and eliminated a series of qualified problem labels, e.g. history-of, to obtain a dataset of 4122 problem labels. Within this dataset, the authors identified 2810 labels (68.2%) as referring to a broad range of disease-related processes. The parser correctly recognized and validated 1398 of the 2810 disease-related labels (49.8±1.9%) and correctly excluded 1220 of 1312 non-disease-related labels (93.0±1.4%). 812 of the 1181 match failures (68.8%) were caused by terms either absent from UMLS or modifiers not accepted by the parser; 369 match failures (31.2%) were caused by labels having patterns not recognized by the parser. By enriching the UMLS lexicon with terms commonly found in provider-entered labels, it appears that performance of the parser can be significantly enhanced over a few subsequent iterations. This initial evaluation provides a foundation from which to make principled additions to the UMLS lexicon locally for use in symbolizing clinical data; further research is necessary to determine applicability to other health care settings.

INTRODUCTION

Symbolization services—mapping external concept representations to machine-readable, symbolic representations—are a desirable feature for emerging terminology servers. The process of symbolization usually occurs at the man-machine interface where some natural language representation of a clinical concept is transformed into a symbolic representation amenable to processing by other components in a clinical information system. This process may occur interactively through a user-interface or automatically through medical language processing.

The capture and symbolization of data from the clinical problem list facilitates the creation of high-fidelity patient resumes for further use in aggregate analysis and decision support. Our experience with the Online Medical Record (OMR), a heavily-used ambulatory computer-based patient record, has shown that providers are willing to enter limited amounts of clinical data into a computer-based patient record when the data is of high value for delivering care or provider communication.[1] Health care providers frequently employ medical sublanguage when documenting the delivery of care.[2] While the stereotypic patterns found in medical sublanguage facilitate medical language processing, this is complicated by the richness and variety of qualifiers and modifiers used in describing clinical processes.[3] Our anecdotal experience with a variety of context-sensitive graphical user interfaces has revealed that providers are less willing to undertake a prolonged interaction with an interface for the purposes of capturing finely-granular clinical problems.

Symbolization involves concept recognition—mapping to a machine-readable representation—and validation—ensuring that the representation is appropriate to the requested context. Within the domain of clinical problems, for example, it is useful to recognize that a concept lies outside the domain in order to facilitate the capture of a more relevant concept. Therefore, answers to the following questions are important in the development of robust symbolization services:

1. What is the nature of the data that health care providers enter into the problem list?

2. Do health care providers use stereotypic language to describe clinical problems or do they employ more robust "natural language"?

3. Can recognition and validation of entries into the problem list be mediated by a limited set of medical language processing techniques or does this require a generalized natural language processor?

1091—8280/98/$5.00

In this report, we present preliminary findings in the development of a semantic parser for clinical problems employing Lexical Technology's Metaphrase toolkit.[4] In addition to the questions posed above, we are interested in evaluating whether UMLS with its very large lexicon of clinical terms and wealth of semantic information is useful in implementing symbolization services. Specifically, we look at the performance characteristics of the parser in recognizing and validating free-text, disease-related clinical problems originating in OMR as a precursor to symbolic representation.

METHODS AND MATERIALS

OMR Clinical Problem List

The Online Medical Record (OMR) is an ambulatory computer-based patient record system in active use within 37 of 76 outpatient clinics on the East Campus of the Beth Israel Deaconess Medical Center.[5] In 1997, health care providers entered 270,890 notes, 203,979 medications, and 38,836 clinical problems into OMR. The clinical problem list is an interactive application within the OMR application suite that allows the health care provider to maintain a running list of acute and chronic medical conditions. The clinical problem list is used to provide a working patient summary in order to facilitate provider communication; the problem list has also been used to trigger decision support activity in the domain of HIV disease.

An entry in the clinical problem list consists of four fields: a problem label, a problem description, an activation date, and an inactivation date. Entry into the problem list is through free-text. A problem dictionary consisting of 1270 commonly used problem labels is searched when the problem label string is entered to provide initial suggestions.[6] The application does not enforce any particular style for the problem list, so that the format of the problem list—i.e., use of the label and description fields—is left to the discretion of the individual provider.

Problem List Data Set

All non-dictionary clinical problem labels entered into the OMR between 1989 and August 1997 were collected as an initial dataset. The problem labels were randomly assigned to one of five subsets for further analysis. The following qualified problem label types were subsequently eliminated from the dataset: historical problems and problems attributable to family members, prior procedures and events, probabilistic problems, and contemplated problems (e.g. HISTORY OF, FAMILY HISTORY OF, STATUS POST, QUESTION OF, RULE OUT, and related variants).

All problems in the subset analyzed in this study were classified into the categories listed in Table 1 by two of the authors (HSG, CH). We employed the following definition for clinical problems: *A clinical problem is a finding or process that requires consideration for diagnostic evaluation or therapeutic intervention.*[7] We defined *disease-related* to refer to a broad-range of organic and behavioral pathology, including physical signs and symptoms. Problem labels such as organism names, which were ambiguous in their representation of a finding or process, were classified as "Fails Definition of Clinical Problem". Problem labels that were ambiguous as to representing either a pathologic process or an abnormal test result (e.g., elevated serum cholesterol") were classified as "Test Result".

Table 1 Categorization of clinical problem labels

Problem Category	Number Of Labels	% N=4122
Disease-Related	2810	68.2
Fails Problem Definition Miscellaneous	487	11.8
Test Result	287	7.0
Procedure	214	5.2
Allergy	114	2.8
Social And Care Issues	61	1.5
Categorical Label	58	1.4
Protocol	47	1.1
Gravid Para Status	44	1.0

Semantic Parser for Disease-Related Clinical Problems

A semantic parser was developed to identify disease-related clinical problems. The parser identifies all problems labels meeting the following pattern as disease-related clinical problems: [<Modifier>]* [<Body Location>]* <Disease Process>, e.g., one disease process with zero or more body locations or modifiers. The parser's processing is based on UMLS semantic types; the semantic types used by the matcher are listed in Table 2.[8] Lexical Technology's Metaphrase Toolkit (UMLS version 8) is used to provide a semantic digest of every problem label to the parser. Metaphrase also provides a spell-

checking capability. The current parsing algorithm is limited to problem labels of five words or less.

Table 2 UMLS semantic types accepted for parser tokens

Type Codes	UMLS Semantic Types
	Disease Process
19	Congenital Abnormality
20	Acquired Abnormality
33	Finding
37	Injury or Poisoning
46	Pathologic Function
47	Disease or Syndrome
48	Mental or Behavioral Dysfunction
49	Cell or Molecular Dysfunction
55	Individual Behavior
184	Sign or Symptom
190	Anatomic Abnormality
191	Neoplastic Process
	Body Location
22	Body System
29	Body Location or Region
31	Body Substance
82	Spatial Concept
	Modifiers
79	Temporal Concept
80	Qualitative Concept
81	Quantitative Concept

The parser may identify matches as "dictionary matches" or "perfect matches". A problem label may exactly match a UMLS concept label with an appropriate semantic type and be labeled a dictionary match. Otherwise, the parser will further process the label to see if it is a perfect match to the disease-based problem pattern. Metaphrase may return matching fragments containing additional terms not analyzable by the matcher, e.g., "biceps tendon" for "biceps". These are reported as "perfect matches with extra information", but classified as match failures. Specific match failures reported by the parser include labels in which a disease process can not be identified and one word labels absent from the UMLS lexicon.

The performance of the parser is reported in terms of sensitivity, specificity, and predictive values with 95% confidence intervals. We report problem labels correctly identified as disease-related as *validated*; problem labels mapped correctly to a semantically identical pattern are reported as *recognized.* Performance is reported in terms of problem labels that have been accurately recognized and validated.

Match failures were analyzed to qualify reasons for failure. The following broad categories were identified: unmatched fragments either absent from UMLS or having inappropriate semantic types, wrong matches, labels containing more than one disease process, semantic matches unrecognizable by the parser, "complex" problems with embedded qualifiers, incorrect semantics where the phrase *in toto* implies a disease process, but no single disease process is identifiable by semantic type.

RESULTS

26,856 unique clinical problem labels representing 60,065 total non-dictionary problem labels were entered into OMR between 1989 and August, 1997 (36% of all OMR problems). Labels were randomly assigned to five subsets, the first of which was arbitrarily chosen for use in this analysis. From this subset, a series of qualified labels—historical and family history (404), probabilistic (249), status post events (362), and rule-out (34)—a set of nonsensical strings (43), and the set of labels with greater than five words (130) were all eliminated to leave a subset of 4122 problem labels.

These 4122 problem labels were classified into 9 categories listed in Table 1. Of the 2810 disease-related problem labels, 1629 labels (784 dictionary matches, 845 perfect matches) (58.0%) were correctly validated by the parser as disease-related. 1398 of these 1629 labels (85.8%) were accurately recognized by the parser. Of the 1312 non-disease-related problem labels, the parser correctly excluded 1220 labels (93.0%). Therefore, in terms of performance characteristics, the parser exhibited a sensitivity of 49.8±1.9%, a specificity of 93.0±1.4%, a positive predictive value of 93.8±1.2%, and a negative predictive value of 46.4±1.9%.

Match failures for disease-related problems as classified by the parser are listed in Table 3. Match failures as classified by the authors are listed in Table 4. Of the 92 false positive matches from the set of non-disease-related problem labels, 41 are classified as the UMLS semantic type <Pathological Function> or <Disease or Syndrome> and 30 are classified as the UMLS semantic type <Finding>.

Table 3 Match failures as classified by the parser

Match Failure According To Parser	Number Of Labels	% N=1181
Wrong Match	797	67.5
Extra Information	139	11.8
Absent From UMLS	129	10.9
Absent Disease Process	116	9.8

Table 4 Match failures as classified by the authors

Match Failure According To Authors	Number Of Labels	% N=1181
Unmatched Fragment	434	36.7
Wrong Match	378	32.0
More Than One Disease Process In Label	121	10.2
Match Unrecognized By Matcher	120	10.2
Complex Problem	81	6.9
Incorrect Semantics	47	4.0

Discussion

Several observations from the problem label classification are notable. A majority of the labels (64%) do refer to a broad range of pathology. Interestingly, the second largest subset (11%) refer to a wide variety of miscellaneous issues or represent clinical processes which would require significant inferential capabilities to symbolize automatically. The use of categorical problem labels, e.g., *dermatologic*, implies a particular style of problem list maintenance; unfortunately, this style buries clinical detail in a comment field which would require additional language processing capabilities to scavenge. The great diversity of data found within the problem list suggests that an alternate form of clinical resume might be more useful for segregating data for both delivery of care as well as symbolizing data.[9]

In identifying disease-related problems, this initial version of the parser performed with a high positive predictive value, but a low sensitivity. Use of the large UMLS lexicon in place of the native OMR problem dictionary would increase dictionary matches by approximately 8%. The analysis reveals two major deficits in the parser: an absence of lexical entries in UMLS and the presence of additional patterns for disease-related problems. A major feature of provider-entered OMR problems is the extensive use of abbreviations and acronyms. Because no current vocabulary within UMLS is used in a manner similar to the OMR problem dictionary, it is not unexpected that these kinds of lexical entries are missing from UMLS. A number of unmatched fragments also appear within wrong matches as well as false positive matches where the Metaphrase spell-checker has massaged a term into a valid modifier or process. Additionally, there are a significant number of disease processes and modifiers that only occur within the context of a postcoordinated phrase and not in an atomic form.

Empirically, the simple pattern used by the parser appears to be the predominant disease-related pattern. Subsequent iterations of the parser will need to address additional disease-related patterns revealed in the analysis, but these labels comprise a limited percentage of the total (22%). It is also interesting that the simple parser performs with such high specificity, given the inability of the parser to recognize nonsensical labels fitting the pattern. These pathologic cases do not appear to occur in the documentation of clinical problems; the more important issue seems to be the quality of the semantic information associated with the reference lexicon.

It should be emphasized that the task in which we achieved a modicum of success was in recognizing a problem label as *some disease process*. We employed a broad semantic filter that appears to have a high sensitivity and specificity in identifying clinical pathology for terms found in UMLS. However, because of existing heterogeneity in UMLS semantic typing, e.g., because a disease process may be classified as a <pathological process> or a <disease or syndrome> or a <finding> without a principled explanation, there still exists potential pitfalls when symbolizing a label even after it has been recognized. The arbitrary nature of distinctions between <disease or syndrome> and <pathological process> have previous been reported. [10] We add to this the difficulty in discriminating between <disease or syndrome> and <laboratory or test result>.

Finally, problem labels with a length greater than five words comprise 3.1% of all eligible problem labels. These labels can not be processed by the current version of the parser and were excluded from the analysis. A majority of these labels are disease-related, but generally describe more than one disease process or represent complex problems. If all of these labels are counted as unmatched disease-related labels, sensitivity is reduced to 47.6±1.8% and negative predictive value is reduced to 44.2±1.9%.

This initial analysis provides a foundation from which to make principled additions to the UMLS lexicon to enhance its use for medical language processing. We have begun this process by integrating the OMR problem dictionary (BI96), which contains a large set of abbreviations and acronyms, into UMLS through editing tools being developed for Metaphrase. It will be interesting to re-evaluate the performance of the parser with the addition of the BI96 vocabulary and terms identified through this analysis.

CONCLUSION

We have described an initial version of a semantic parser for clinical problems implemented on top of Lexical Technology's Metaphrase™ Toolkit. Using semantic information from UMLS, this version of the parser is designed to recognize and validate disease-related clinical problems. An evaluation of the parser involving a randomly selected subset of clinical problems from the OMR problem list revealed a sensitivity of 49.8±1.9%, a specificity of 93.0±1.4%, a positive predictive value of 93.8±1.2%, and a negative predictive value of 46.4±1.9%. 68.8% of match failures were caused by terms either absent from UMLS or modifiers not accepted by the parser, while the remaining 31.2% were caused by problem labels having patterns not recognized by the parser.

A great diversity of data is present in the OMR problem lists, although the majority of labels do refer to clinical pathology and test results. A majority of disease-related problem labels are reported using very stereotypic patterns. This initial evaluation provides a foundation from which to make principled additions to UMLS that should dramatically affect performance over relatively few iterations of the parser. While the semantic classification of the UMLS lexicon appears sufficient for the recognition of disease-related entities, further investigation is required to determine the adequacy of the classification system for creating symbolic representations of recognized entities. Further investigation is also necessary to determine the applicability of these results to other health care settings.

ACKNOWLEDGEMENTS

This work was supported in part by a cooperative agreement award from the Agency for Health Care Policy and Research and the National Library of Medicine (HS08749). Special thanks to Daniel Sands, MD for providing the OMR non-dictionary problem set and OMR usage statistics.

References

[1] Safran C, Rury C, Rind DM, Taylor WC. A computer-based outpatient medical record for a teaching hospital. MD Comput 1991;8(5);291-299.
[2] Sager N, Lyman M, Bucknall C, Nhan N, Tick LJ. Natural language processing and the representation of clinical data. JAMIA. 1994;1:142-160.
[3] Chute CG and Elkin PL. A clinically derived terminology: qualification to reduction. JAMIA Suppl. 1997;4:570-574.
[4] Tuttle MS, Olson NE, Keck KD, et. al. Metaphrase: an aid to the clinical conceptualization and formalization of patient problems in the healthcare enterprise. Presented IMIA WG6, Jacksonville, FL. January 1997.
[5] Sands DZ, Rind DM, Vieira C, Safran C. Going paperless: can it be done? JAMIA Suppl. 1997;4:887.
[6] Zehlinger J, Rind DM, Caraballo E, Tuttle MS, Olson NE, Safran C. Categorization of free-text problem lists: an effective method of capturing clinical data. SCAMC 1995: 416-420.
[7] Goldberg HS, Goldsmith D, Law V, Keck K, Tuttle M, Safran C. An evaluation of UMLS as a controlled terminology for the problem list toolkit. MEDINFO '98. In Press.
[8] UMLS Knowledge Sources. 8th Edition. National Library Of Medicine. January 1997.
[9] Claus PL, Carpenter PC, Chute CG, Mohr DN, Gibbons PS. Clinical care management and workflow by episodes. JAMIA Suppl. 1997;4:91-95.
[10] Steve G, Gangemi A, Pisanelli DM. Integrating medical terminologies with ONIONS methodology. IOS-PRESS. In Press. See http://saussure.irmkant.rm.cnr.it/onto/index.html.