Identifying UMLS Concepts In Emergency Department Terms

Using

Domain Knowledge and Natural Language Processing Techniques

ABSCESS TOOTH|Dental abscess|C0518988|
ABD DISTENTION/VOMITING|Term not found.
LIPS SWELLING|swollen lips|C0240211|
Side pain - left|PAIN-LEFT SIDE|C0542135|
Side pain - right|Term not found.

July 27, 2001

NLM Medical Informatics Fellows Summer Rotation Project
Student: Debbie Travers
Mentor: Olivier Bodenreider

ABSTRACT

Objective: The objective of this pilot project was to apply and evaluate methods for processing Emergency Department (ED) Chief Complaint (CC) terms, in order to identify the concepts that comprise the ED CC domain.

Materials and Methods: A corpus of CC data was collected from three EDs representing urban, rural and suburban academic medical centers. For the pilot project, the corpus included all CC terms recorded for ED visits during January 2001. There were 6900 visits, and 6054 unique CC terms recorded during the study period. ED terms were initially mapped to the Unified Medical Language System ® (UMLS ®) Metathesaurus® in order to identify corresponding concepts. We evaluated how many CC terms exactly matched a UMLS concept. We then performed a normalized match on those terms that had not matched a UMLS concept exactly, and again calculated a match rate. The terms that still did not match a UMLS concept were then examined by the investigator, who used domain knowledge to identify patterns in the data. Customized NLP techniques were then applied to address the use of punctuation, acronyms, and modifiers in the CC terms. The NLP techniques were applied in three successive rounds, starting with simple techniques and using more aggressive techniques. After each round, the resulting CC terms were again compared to the Metathesaurus to determine how many terms matched a UMLS concept.

Results: Prior to round one, 22% of the terms matched a UMLS concept. After three rounds of processing, UMLS concepts were identified for 35% of the CC terms that could not originally be mapped to the UMLS. Through the pilot project, we identified 1855 unique CC terms with corresponding UMLS concepts.

Discussion: Through application of domain knowledge and simple NLP techniques, we identified concepts that provide a partial representation of the ED CC domain. Additional review by domain experts and further customized NLP is needed to identify concepts for the non-matched ED CC terms. The techniques will then be applied to a sample of CC terms from one year, in order to provide a more complete representation of the ED CC domain by accounting for seasonality and less frequent CC terms.

Conclusion: Domain knowledge was a key factor in increasing the number of UMLS concepts identified from the ED CC terms over that from standard UMLS NLP resources.

INTRODUCTION

George Hripscak and colleagues (1995) described the challenges of trying to unlock clinical data from narrative reports, such as hand-written Emergency Department (ED) records. Most EDs in the U.S. have yet to implement automated electronic medical records. Notes are becoming more available in electronic form, but the richness of such clinical information can't be fully utilized with current free text entry and machine interpretation systems. Clinicians often use abbreviations and punctuation to document patient findings, treatment, and progress in narrative notes. In an effort to address these problems, the Centers for Disease Control and Prevention are sponsoring an ongoing national effort to establish standards for ED data (Pollock et al, 1998). The first release of Data Elements for Emergency Departments (DEEDS) 1.0 was in 1997. 156 data elements were included

in this first edition of DEEDS, which recommended existing vocabularies for those data elements with such standards (such as ICD for diagnosis). Data element 4.06 is "Chief Complaint." Since no standard vocabulary exists for documentation of chief complaint (CC), the DEEDS authors recommended adaptation and evaluation of established terminologies as a solution to the need for an ED CC system. Researchers in North Carolina are working to develop an ED CC Thesaurus, and my NLM project is a part of that effort.

Uses for an ED CC Thesaurus include clinical guidelines, electronic surveillance on the local level, as well as public health reporting on the regional and national level. Other uses involve information retrieval, to select cases for operational issues such as specialty room tracking and identification of patients for clinical research. In the absence of standards, chief complaints are recorded in many different ways. There are variations in different hospitals as well as between nurses at the same hospital. For example, allergic reaction might be recorded as: *all rxn, allerg reac, allergic reaction, or allg rxn*. There are also constraints pertaining to data entry systems, such as the fact that the computerized data entry system at UNC Hospitals has a 12-character limit on the entry field for CC.

Initial analysis of ED CC data for the ED Thesaurus project was conducted in 1999 (Schoeffler et al., 1999). In that study, researchers examined all CC terms recorded in 1998 at the UNC Hospitals emergency department. The goal of the project was to process the CC terms into a standard format and then hopefully match those standardized terms to a concept in the UMLS, which could be incorporated into a CC thesaurus. Processing included the standard UMLS normalization tools, which abstract away case, inflection and word order, as well as remove stop words, possessives and replace punctuation with spaces. Additional processing developed by Ms. Schoeffler included expansion of abbreviations and truncated words, and spelling correction. The processed CC terms were compared with the UMLS strings and resulted in a 14% match rate with existing UMLS concepts. Characteristics of the data that contributed to the low match rate included multiple concepts in one CC term, and various uses of punctuation that were stripped away by the UMLS tools and resulted in a loss of meaning.

In a second study, we specifically addressed the use of the oblique stroke ("/") in a corpus of ED CC text from three hospitals (Travers, Haas & Downs, 2001). We found that 19% of all CC terms contained one or more slashes. We developed rules to address the most common uses of the slash, and found that 58% of the terms with slash matched a UMLS concept after the rules were applied and the data were normalized. None of the terms with slash matched a UMLS concept when only the UMLS normalization process was applied. We also developed a rule to process the comma and semi-colon, which were used in place of the word "and". Eight percent of the ED CC terms that didn't match a UMLS concept contained a comma or semi-colon. Using that rule, we split terms containing the comma or semi-colon into two terms. This additional step brought the total match rate up to 62%.

GOALS AND OBJECTIVES

In this pilot study I aimed to improve the match rate with UMLS concepts over that from the previous studies, and evaluate various NLP approaches for handling ED text. These techniques will subsequently be applied to a larger sample of ED CC terms in order to more completely identify the concepts that comprise the ED CC domain. The objective and goals of the study were as follows:

Objective

To apply and evaluate methods for processing Emergency Department (ED) Chief Complaint (CC) terms, in order to begin to identify the concepts that comprise the ED CC domain.

Specific Goals

a. Use domain knowledge to identify patterns in the CC terms that don't match a UMLS concept.
b. Apply NLP techniques to address issues identified in a.
c. Evaluate the NLP techniques for handling ED text.
d. Provide the UMLS researchers with a list of words found in CC's and that are not in the SPECIALIST Lexicon.
e. Provide the UMLS researchers with a list of pairs of ED terms and misspellings of those terms.

MATERIAL AND METHODS

Materials

The sample for this project included ED CC entries from the three North Carolina EDs that are participating in the development of an ED CC thesaurus. The three EDs are at academic medical centers and include a rural site (East Carolina University Hospital), an urban site (Carolinas Medical Center) and a suburban site (University of North Carolina Hospitals). Included were the ED CC terms entered for all patients visiting the three EDs during January 2001. Triage nurses entered the CC terms directly into the hospital information system upon patient arrival to the ED at all three sites. Carolinas Medical Center (CMC) gave nurses the option of selecting a CC from a locally developed, controlled list of 238 terms, or entering the CC as free text. At CMC, 73% of the CC terms were from the controlled list. At the other sites, all CC terms were entered as free text.

In order to begin the experiment with an accurate sample of unique CC terms, we first excluded some cases, and cleaned the data to eliminate duplicate terms due to minor differences in CC terms. Excluded from the sample were terms for patients who presented to the ED but were immediately triaged to Labor and Delivery for pregnancy-related conditions, since the nature of the CC terms varied significantly from all other patients' complaints. Then we cleaned the data to standardize case, eliminate leading spaces and symbols, and eliminate trailing spaces.

Methods

We began the experiment by mapping the CC terms to the UMLS Metathesaurus in order to identify corresponding concepts. This was accomplished in two increasingly aggressive steps using the Knowledge Source Server (NLM, 2001). First, we evaluated how many CC terms exactly matched a UMLS concept. Next, we performed a normalized match on those terms that had not matched a UMLS concept exactly. The normalization tools abstract away case, inflection and word order, as well as remove stop words and possessives, and replace punctuation with spaces (McCray, Srinivasan, Browne, 1994). After the normalized match, we again calculated the match rate with Metathesaurus concepts. These steps were repeated after each round of the experiment.

During the next phase of the study, I examined the remaining non-matching terms. I served as a domain expert, with 20 years' emergency nursing experience and certifications in emergency nursing and informatics nursing. I was able to identify the semantic content of 99% of the CC terms. I then used various sorting techniques, frequencies and word counts to identify common patterns in the non-matching terms. I found three major patterns in the terms: multiple uses of punctuation, frequent acronyms, abbreviations and truncations, and several modifiers and qualifiers. We then applied NLP techniques to address each pattern, by starting with the simplest one first, and moving to more aggressive approaches in successive rounds. The goal was to maximize the match rate with existing Metathesaurus terms. After each round of processing, we first evaluated how many CC terms exactly matched a UMLS concept, and then performed a normalized match on

those terms that had not matched exactly. This allowed us to quantify the gain in concepts matched with each round. Then, I manually inspected the terms that matched a UMLS concept, in order to judge the accuracy of the match based on my knowledge as an ED clinician. At the end of each round, I looked for patterns in the terms that didn't match. Each of the NLP techniques is described below.

1. Punctuation processing

In Round 1, we processed the commonly used punctuation in an attempt to improve the UMLS concept match rate. Using techniques developed in the previous study (Travers, Haas & Downs, 2001), we used sed and awk scripts to remove slashes, commas and semi-colons and replace them as shown in Table 1.

TABLE 1- PUNCTUATION PROCESSING

Processing Step	Input CC	Output CC
Replace	1. h/a 2. b/p elevated	1. headache 2. blood pressure elevated
Expand coordinate constructions, split into 2 terms	1. hip/thigh pain 2. tingling feet/hands 3. testicle pain/redness	1a. hip pain 1b. thigh pain 2a. tingling feet 2b. tingling hands 3a. testicle pain 3b. testicle redness
Retain body location with sign, symptom, finding or injury	1. rash/groin 2. laceration/forehead	1. groin rash 2. forehead laceration
Eliminate unnecessary abbreviations	1. c/o earache	1. earache
Delete slash, comma, semi-colon & split into 2 terms	1. dizzy/nausea 2. fall, rib pain 3. fever; cancer	1a. dizzy 1b. nausea 2a. fall 2b. rib pain 3a. fever 3b. cancer

As shown in Table 1, many terms were split into two terms through the application of the punctuation rules. Some duplicate terms also resulted from this process, which were then eliminated. Thus, the number of terms pre-Round 1 cannot be related to the number of terms post-Round 1.

2. Expansion of acronyms, abbreviations and truncated words

In Round 2, we took the unmatched terms remaining from Round 1, and addressed acronyms, abbreviations and truncated words (AAT). Using the lexical tool for acronym expansion and a manual review of the remaining unmatched CC terms, I identified frequently used acronyms, abbreviations and truncated words. Four domain experts (one nurse and one physician from two of the participating EDs) reviewed the list and identified one or more expansions for each AAT. For those with more than one meaning, I developed rules for disambiguation. Examples of the expansions are shown in Table 2. A perl script was used to apply the expansions.

TABLE 2- EXPANSION OF ACRONYMS, ABBREVIATIONS, AND TRUNCATIONS

Processing Step	Input CC	Output CC
Expand acronym	FB	Foreign body
Expand abbreviation	Rx	Reaction (if after all, allerg, else prescription)
Expand truncation	Diarr	Diarrhea

3. Deletion of qualifiers and modifiers

In the last round, we took the unmatched terms remaining from Round 2, and addressed modifiers and qualifiers. We tokenized the unmatched terms into individual words and performed word counts. I then compared the words to lists of modifiers and qualifiers identified in previous research. As described by Chute and Elkin (1997), modifiers are words that alter the severity, location, or acuity of a clinical term, such as *acute* myocardial infarction. Qualifiers are words or phrases that qualify the meaning of a clinical term, such as *history of* a condition. Previous researchers have found that concept matching was improved when common modifiers and qualifiers were removed (Bodenreider, 2000; McCray & Browne, 1998; Chute & Elkin, 1997). I identified those modifiers and qualifiers present in two or more CC terms, and a perl script was used to delete them from the terms. Examples of the altered terms are shown in Table 3, and a complete list of the deleted modifiers and qualifiers is shown in Table 4.

TABLE 3- DELETION OF MODIFIERS AND QUALIFIERS

Processing Step	Input CC	Output CC
Delete modifier	Right leg injury Severe chest pain	Leg injury Chest pain
Delete qualifier	History of seizure Headache since 5 am	Seizure Headache

TABLE 4- DELETED MODIFIERS AND QUALIFIERS

Modifiers	Qualifiers
Possible Severe Left Right Small Recent Multiple Mid Posterior Multiple Both Sided Inferior	History History of Status post Rule out Wants Since (and delete everything in CC term after "since") After (and delete everything in CC term after "after") X# (for example, X1, X2, etc. and delete everything in CC term after "X#)

RESULTS

1. Quantitative Steps

There were 6900 visits, and 6289 unique CC terms recorded during the study period. Included were 6041 unique free text entries and 248 controlled terms from CMC in a 103 kb text file. Excluded were 76 CC terms for patients who presented to the ED but were immediately triaged to Labor and Delivery for a pregnancy-related condition. The remaining 6289 terms were cleaned to eliminate duplicates that differed only by case. Leading spaces and symbols such as the question mark were also removed. The final sample included 6054 unique CC terms.

Figure 1 shows the results from rounds 1-3 of the study. Prior to Round 1, the ED terms were compared with UMLS concepts, and 944 (16%) of the terms exactly matched a UMLS concept. After normalization, an additional 383 terms matched a UMLS concept, for a total of 1327 (22%) for the pre-Round 1 phase.

In the first round, we applied processing rules to the 4727 non-matching terms, to address the various types of punctuation used in the ED terms. After application of the rules and elimination of duplicates, the remaining sample included 4578 unique CC terms. 485 (11%) of the remaining terms exactly matched a UMLS concept. The remaining terms were again normalized and an additional 114 more matched a UMLS concept, for a total of 595 (13%) for the punctuation phase of the study.

In the second round, we expanded acronyms, abbreviations, and truncated words. We found that 237 (6%) of the remaining 3983 terms exactly matched a UMLS concept. The non-matching terms were again normalized and 119 more matched a UMLS concept, for a total of 358 (9%) for the expansion phase of the study.

The final step involved deletion of 21 modifiers and qualifiers, after which 439 (12%) of the remaining terms exactly matched a UMLS concept. The non-matching terms were again normalized and 227 more matched a UMLS concept, for a total of 666 (18%) for the deletion phase of the study.

During rounds 1-3 combined, we were able to map 35% of the ED CC terms to one or more UMLS concepts, in addition to a 22% match rate using exact and normalized string matching only.

2. Qualitative Aspects

We identified a total of 2946 entry CC terms that matched one or more UMLS concepts, of which 1855 CC terms were unique. 68% of the matched terms were identified with one UMLS concept only, and 32% were identified with two or more UMLS concepts. I then evaluated the accuracy of those matches. Of those terms that matched only one UMLS concept, I took a random sample of the results and manually examined all the matched concepts for accuracy. I found that 72 of the 77 matches (92%) were accurate. Of the 72 that matched, many didn't match exactly but the match was semantically accurate. For example, the input CC term *arm laceration* matched UMLS concept C0432974, *laceration of upper arm*. A small number of matches (8%) were not accurate, for example, the input CC term *stepped on by sibling* matched UMLS concept C0337504, *step sibling*.

Of those CC terms that matched more than one UMLS concept, I identified the semantic group for each CC term and the matching UMLS concepts (McCray, Burgun & Bodenreider, 2001)). 85% of the CC terms matched at least one UMLS concept from the same semantic group.

DISCUSSION

Using my domain knowledge supplemented by a low-technology approach, we were able to obtain a higher match rate with UMLS concepts than that obtained using the standard UMLS

matching and normalization tools. Our NLP techniques involved simple pattern matching. Though our more aggressive approach introduced more risk for altering CC terms from their original representation, my domain knowledge was useful in facilitating the appropriate processing of terms containing patterns such as punctuation and acronyms. We found that the CC terms are close to clinical terms, but have a level of granularity that is finer than the standard vocabulary terms found in the UMLS. By deleting selected modifiers and qualifiers, we were able to broaden the terms and increase the match rate significantly. We learned that modifiers and qualifiers are frequently used in ED CC terms, and thus identified the need to include them in the final ED CC thesaurus.

Through my evaluation of the accuracy of the UMLS matches, we learned that there will be a need for manual review of the matched concepts as this project continues. The accuracy rates of 92% for single concept matches, and similar semantic groups for 85% of the multiple concept matches is encouraging. However, the final ED CC Thesaurus will require a higher rate of accuracy. The techniques we have developed during this project certainly provide a semi-automated approach to concept identification for the ED CC domain, which is superior to manual review and mapping of each ED CC term.

We were able to identify 2946 CC terms that matched one or more concepts. 1855 of those terms were unique. We only achieved a match for 22% of the initial CC terms, and 35% of the remaining fully processed sample. With 2959 non-matched terms remaining after Rounds 1-3, we have at best a partial representation of the ED CC domain from the January 2001 visits to the three hospitals. Additional review by domain experts and further customized NLP is needed to identify concepts for the non-matched ED CC terms. The techniques will then be applied to a sample of CC terms from one year, in order to provide a more complete representation of the ED CC domain by accounting for seasonality and less frequent CC terms.

CONCLUSIONS

In conclusion, during the summer rotation, I completed pilot work toward the identification of concepts that comprise the ED CC domain. I met my goal of identifying patterns in the non-matched ED CC terms. With the assistance of my mentor, Dr. Bodenreider, I applied NLP techniques to address punctuation, acronyms/abbreviations/truncations, and modifiers/qualifiers in the ED CC terms. Through this processing, we increased the match rate with UMLS concepts. The most aggressive NLP technique was the deletion of modifiers and qualifiers, which resulted in the greatest gain in UMLS concept matching. Finally, I evaluated the UMLS matches and found them to be relatively accurate.

Future directions for this work include further identification of patterns in the non-matched terms and application of the NLP techniques to a more comprehensive sample of ED CC. I will also assemble a list of any words that are not in the SPECIALIST Lexicon and a list of pairs of ED terms and misspellings of those terms, for the UMLS researchers. Finally, I will begin to build the foundation for the ED CC Thesaurus by assembling concepts and other information identified from the ED data into a database. Information in the database will include:

1. Metathesaurus concept.
2. Metathesaurus concept properties (e.g., acronym).
3. Which Metathesaurus source vocabularies contain the concept.
4. All entry terms, including acronym and abbreviation information, and misspellings.

REFERENCES

Bodenreider O. Using UMLS semantics for classification purposes. Proc AMIA Symp. 2000; 86-90.

Chute CG & Elkin PL. A clinically derived terminology: Qualification to reduction. Proc AMIA Symp. 1997; 570-574.

Hripcsak, G., et al., Unlocking clinical data from narrative reports: A study of natural language processing. Ann Int Med, 1995. 122: p.681-688.

McCray AT, Burgun A & Bodenreider O. Aggregating UMLS semantic types for reducing conceptual complexity. MEDINFO 2001 (to appear).

McCray AT, Browne AC. Discovering the modifiers in a terminology data set. Proc AMIA Symp.1998; 780-784.

McCray AT, Srinivasan S, Browne AC. Lexical methods for managing variation in biomedical terminologies. Proc Annu Symp Comput Appl Med Care 1994; 235-9.

National Library of Medicine Knowledge Server, http://umlsks.nlm.nih.gov/, accessed July 25, 2001.

Pollock DA, Adams DL, Bernardo LM, Bradley V, Brandt MD, Davis TE, et al. (DEEDS Writing Committee). Data elements for emergency departments systems (DEEDS), release 1.0: A summary report. J Emerg Nursing 1998; 24: 35-44.

Schoeffler KM, Travers D, Hales JW, Waller AE, Tintinalli J. Evaluating emergency department chief complaint data. Proceedings of the 1999 National Library of Medicine's Medical Informatics Trainees' Annual Meeting; 1999 Jul 8-9; New York, New York.

Travers D, Haas S, Downs S. Multiple uses of the oblique stroke (slash, '/') in emergency department text. Proceedings of the 2001 National Library of Medicine's Medical Informatics Trainees' Annual Meeting; 2001 Jul 17-18; Bethesda, Maryland.

Figure 1- Results Summary