Mapping Clinically Useful Terminology to a Controlled Medical Vocabulary

Randolph C. Barrows Jr., MD, James J. Cimino, MD, Paul D. Clayton, Ph.D.
Center for Medical Informatics
Columbia University
New York, NY 10032

We have mapped clinically used diagnostic terms from a legacy ambulatory care system to the separate controlled vocabulary of our central clinical information system. The methodology combines elements of lexical and morphologic text matching techniques, followed by manual physician review. Results of the automated matching algorithm before and after partial manual review are presented. The results of this effort will permit the migration of coded clinical data from one system to another. Output from the system after the term review process will be fed back to the target vocabulary via automated and semi-automated means to improve its clinical utility.

INTRODUCTION

The Columbia-Presbyterian Medical Center (CPMC) has a dataset including coded clinical diagnoses and medications pertaining to ambulatory patients that was captured from primary care practitioners using the outpatient Clinical Profile system [1]. The base set of terms in the system dictionary, from which providers choose diagnoses, came from the Systematized Nomenclature of Medicine (SNOMED), but users of the system are free to define new terms as needed, and have frequently done so. New dictionary terms are assigned a unique code by the system, which is used to reference the term in the database. This system uses a hierarchical database and is isolated from the CPMC Clinical Information System (CIS) with its relational database and decision support capabilities [2].

Ongoing work at CPMC aims to develop a clinical system for ambulatory care that functions as part of the CIS and uses the central patient database with its own controlled medical vocabulary and data dictionary called the Medical Entities Dictionary (MED) [3]. In order to prevent the loss of clinical information to users of the new system, and redundant data entry efforts on the part of health care providers, it is desirable to map coded terms from the outpatient Clinical Profile system to coded entities in the MED and migrate patient data from the Clinical Profile system into the CIS. Manual mapping is feasible but time consuming for a medical domain expert, even given the presence of reasonable tools for browsing the target vocabulary. It would presumably, however, have the highest degree of accuracy.

METHODS

Given the importance of accurately representing real data on real patients, the approach taken is that of an initial enhanced lexical matching algorithm followed by manual review of the match results by a medical domain expert.

The match algorithm utilizes a unix-based implementation of the MED that is maintained to be concurrent with the CIS version. The MED is conceptually organized as a semantic network of about 35,000 medical entities. Each entity has a unique numerical identifier ("medcode") and a number of term attributes ("slots"). Some slots, such as SUBCLASS-OF, are link attributes (have medcodes as values), and other slots, such as the NAME and SYNONYM slots, are literal attributes (have string values). Some slots, such as SYNONYM, may have multiple uniquely-valued instances. For example, medcode 9624 is a SUBCLASS-OF 9623 ("Disease of Upper Respiratory Tract"), has a NAME value of "Acute Upper Respiratory Infection", has a SYNONYM value of "Acute URI", and another SYNONYM value of "Common Cold". Development efforts (by Dr. Barry Allen and Nilesh Desai) in the Center for Medical Informatics have resulted in an implementation of the MED that is optimized for speed of queries and compactness of representation, and resides in shared-memory on an IBM RS/6000. A library interface allows complex queries to be performed on the contents of the MED from application programs, and a command line interface, a menu-driven interface, and an X Window graphical browser have been developed as well.

0195-4210/94/$5.00

The match algorithm is dependent upon preprocessing information from the MED available as part of the unix implementation. In preprocessing, all string-valued slots of all MED entities are tokenized. Each token (a word or term fragment separated by non-alphanumeric characters) is indexed to the medcodes of the entities in which it is found to occur in some string-valued slot, and redundant token occurrences are eliminated. Also available in preprocessed form is a set of Word Groups (WG), consisting of groups of medically synonymous tokens that were originally obtained from lexical variants and synonyms in the Unified Medical Language System (UMLS). The base set includes medical morphemes such as "hepatic" = "liver" = "livers" in one set, and "cardiac" = "heart" = "hearts" in another, and was enhanced based on deficiencies noted after manual review and algorithmic use.

In the matching algorithm, a term from the Clinical Profile system is tokenized, and each token is mapped to a WG (which has the index token as a member). Each member of a WG maps to 0 or more medcodes where the member token occurs in some string valued slot. The union of the medcodes associated with each WG member comprises a set of MED entities, each of which might contain a conceptual match for the token in some literal-valued slot. Such a set of medcodes is determined for each token comprising the Clinical Profile term, and the intersection of these sets is then taken to yield a solution set of possible MED matches for the Clinical Profile term. If this solution set is empty, the intersection is relaxed so that, of n sets of medcodes, one for each of the n tokens comprising a Clinical Profile term, any medcode need only be common to n-1 of the sets, rather than all n, for inclusion in the solution set. If only a single medcode exists in the solution set, that match is returned as a "SOLO"; otherwise the MED entities in the solution set are ranked in order of their likeness to the Clinical Profile term according to the Longest Common Substring (LCS) algorithm [4]. An LCS scoring is performed between the Clinical Profile term and the NAME slot, as well as each SYNONYM instance, of each MED entity in the match solution set.

Results of the automated match are output to a file, and another program is run to analyze the results and generate a statistical characterization. The output of the automated match is reviewed manually by a clinical vocabulary domain expert (RCB and Dr. Olveen Carrasquillo). The manual review process identifies five types of matches: true positive matches (in which the algorithm proposes a correct match); true negative matches (in which the algorithm correctly identifies that no accurate match exists); false negative matches (in which the algorithm falsely states that no accurate match exists in the MED); false positive matches in which the algorithmically proposed match is not accurate, but another MED term exists which is an accurate match; and false positive matches in which the algorithmically proposed match is not accurate, and no accurate match exists in the MED. Another program is then run on the reviewed file and it generates 3 output files: one file maps the code of each successfully matched Clinical Profile term to a MED code; another file maps each unmatched Clinical Profile term (no accurate MED match exists after manual review) to the closest more general concept that can be identified in the MED; and a third file maps newly identified synonyms (Clinical Profile terms judged as useful by the reviewer) to existing MED terms.

RESULTS

Clinical profile terms were divided into two groups: those SNOMED-derived terms actually used to describe patients, and those user-defined terms actually used to describe patients. Unused dictionary terms were not matched. An LCS match value of 0.75 or greater was used to identify possible matches between Clinical Profile terms and Med terms. This number was chosen heuristically as a reasonable cutoff, based on earlier matching work which showed no improvement in the sensitivity of the match below this level, but significantly less specificity. The algorithm matched 65% (674/1045) of the SNOMED-derived Clinical Profile terms to at least one MED term (371 failed to match at an LCS cutoff of 0.75). Thirty seven percent (387/1045) of these terms matched to only a single MED entity, including 15% (159/1045) that matched "SOLO". Ninety seven percent of the terms that matched (647/674) did so to fewer than 10 MED terms, but one term matched to 48 MED terms. Thirty percent of the terms matched "exactly" (each was either a "SOLO" match or matched with a perfect LCS score of 1.00). Of SNOMED-derived terms with more than one potential match (after Term Group intersection), 81% (421/515) matched to the NAME of a MED entity, and 19% matched via SYNONYM instances.

The algorithm matched 51% (631/1225) of the user-defined terms to at least one MED term. Thirty one percent (377/1225) matched to a single MED term. Of user-defined terms that matched, 97% (613/631) matched to fewer than 10 MED terms, and 1 term matched to 31 MED entities. Twenty five percent of the terms matched "exactly", as defined above. Of user-defined terms with more than 1 potential match, 77% (348/453) matched to the NAME of a MED entity, and 23% to a SYNONYM instance.

Examples of match results and reviewer actions follow. In the first example, the Clinical Profile term matched "SOLO" (uniquely), so no LCS scoring was invoked. Here the reviewer need do nothing.

SRC|D0121001|ERYSIPELAS|
TARG|6458|L|Erysipelas|SOLO

In a second example, the term matched perfectly (LCS score == 1) to one MED entity, and with 81% agreement to another MED entity (the textual qualifier NOS --"Not Otherwise Specified" was removed in preprocessing). Here a reviewer need only delete the second, less exact, match.

SRC|D0102001|VENEREAL DISEASE, NOS|
TARG|6828|P|Venereal Disease|1.000
TARG|6984|P|Other ICD9 Venereal Disease|0.815

In the third example, the term matched with a high degree of accuracy to a very different appearing MED term via one of its synonyms. It also matched with lesser accuracy to the NAME of another MED term, and to another synonym (but only the NAME of the MED entity is displayed). Here a reviewer need only verify the accuracy of the top match, and delete the less accurate matches.

SRC|D0131001|GONORRHEA|
TARG|6929|S|Acute Gonococcal Infection of Lower Genitourinary Tract|0.909
TARG|6946|P|Chronic Gonorrhea|0.794
TARG|6929|S|Acute Gonococcal Infection of Lower Genitourinary Tract|0.750

In a fourth example, the term does not match with good agreement to any MED term (LCS score < 0.75), but the closest 10 matches according to LCS scoring are listed.

SRC|D0525001|HEPATITIS DISEASE OR SYNDROME|
PROB|no matches >= 0.750
POOR|33838|P|Hepatitis D (Delta Agent)|0.683
POOR|14747|P|Hepatitis|0.650
POOR|14750|P|Hepatitis in Viral Disease|0.629
POOR|6706|P|Hepatitis A|0.621
POOR|6709|P|Hepatitis B|0.621
POOR|33837|P|Hepatitis C|0.621
POOR|33839|P|Hepatitis E|0.621
POOR|14751|P|Hepatitis in Nonviral Infectious Disease|0.508
POOR|6701|P|Viral Hepatitis|0.483
POOR|6720|P|Mumps Hepatitis|0.483

Here a reviewer notes that the MED term "Hepatitis" in the list of sub-threshold matches is actually the appropriate conceptual match, and labels the algorithmic match as an FN (False Negative) while indicating the correct match and deleting unused alternatives:

SRC|D0525001|HEPATITIS DISEASE OR SYNDROME|
FN|14747|P|Hepatitis|0.65

In the following example, the Clinical Profile term matches with above-threshold agreement to a MED term, but this a false positive match due to lexical similarities:

SRC|D0521001|VIRAL HEPATITIS, TYPE A|
TARG|6701|P|Viral Hepatitis|0.826

In this case, the correct matching entity had to be identified by manual search and browsing of the MED, so the algorithmically suggested match is labeled as an FP (False Positive), and the correct (COR) match subsequently listed. In addition, the reviewer indicated that the Clinical Profile term should be added as a synonym to the MED term:

SRC|D0521001|VIRAL HEPATITIS, TYPE A|
FP|6701|P|Viral Hepatitis|0.83
COR|6706|M|Hepatitis A
ADDSYN

In a final example, the term was an algorithmically false positive match, and after manual review it was determined that no conceptually accurate match exists in the MED. However, the closest more general term was identified and listed in the NOMATCH line:

SRC|D2262001|HYPERPARATHYROIDISM, PRIMARY|
FP|3322|P|Hyperparathyroidism|0.839
NOMATCH|3322|Hyperparathyroidism|0.839

Algorithmic matches for the 1045 SNOMED-derived terms have been manually reviewed and edited as indicated in the preceding paragraphs. The edited version was then processed to generate the final match output and related files including performance statistics. Algorithm performance is as follows:

403 TP + 103 FN + 73 FP(match found) = 579 matches

285 TN + 181 FP(no match found) = 466 no-matches

With the current MED content, the best the matching algorithm could have done was to match 55% (579/1045) of Clinical Profile terms. It actually matched 39% (403/1045) of the terms, or 403 of 579 possible matches for a match recall of 70%. It was accurate in 403 of 657 (403+73+181) algorithmic matches for a match precision of 61%. Forty five percent (466/1045) of Clinical Profile terms could not be matched to the MED. The algorithm correctly identified 285 of 466 possible no-matches for a no-match recall of 61%, and was accurate in 285 of 388 (285+103) algorithmic no-matches for a no-match precision of 73%. In all, the algorithm correctly classified 688 (403+285), or 66%, of the 1045 terms.

DISCUSSION

The matching algorithm described above performs with reasonable recall and precision compared to other standard techniques[5] in the 1045 clinically used SNOMED-derived terms. It is likely that the algorithm would perform slightly less well on the collection of user-defined terms that has yet to be reviewed, just based on the automated analysis of that match.

The algorithmic mapping saves considerable time for medical domain experts who are ultimately responsible for the equivalency of medical concepts expressed in the two separate vernaculars used to represent patient states. The manual review process provides for maximal accuracy and permits statistical reporting on the accuracy of the automated match compared to "gold standard" experts. This process has proved to be, as expected, the principle bottleneck due to limited person-power and (until recently) adequate vocabulary browsing tools.

The immediate value of this work is two fold. First it will allow coded patient data to be migrated from one clinical information system to another despite complete disparity in their respective data dictionaries. Second, it provides valuable feedback to the MED regarding deficiencies and possible improvements that can be made toward the support of ambulatory patient care activities.. This is an important aim at CPMC, where there is strong interest in creating a truly useful controlled clinical vocabulary for electronic medical record systems and decision support. Output from the described system lists newly identified synonyms that can be automatically incorporated into the MED. The 45% of instances where target MED concepts are missing and new entities need to be defined in the MED will require some non-automated effort, but these instances are valuable increments towards constructing a general and clinically useful vocabulary for the storage of coded patient data.

Future directions of this work include comparing the results experimentally with other mapping techniques. Current methods for mapping text phrases that represent medical concepts include lexical techniques, such as the string matching techniques used by UMLS developers [6]; morphologic text analysis [7]; statistical techniques [8]; and semantic indexing [9,10]. In addition, a novel least squares fit mapping technique using large collections of human-assigned matches as training sets has been reported to outperform other available methods [5].

Although the MED is conceptually organized as a semantic network of medical entities, terms from the Clinical Profile have no associated structure, so semantic mapping techniques were not employed. However, much of the term-content for diagnoses and procedures in the MED was derived from ICD9 [11], so the post-review results reported above approximate those expected in a mapping of 1045 SNOMED terms to ICD9. This suggests that another interesting experiment would be to utilize the UMLS metathesaurus to translate SNOMED-derived Clinical Profile terms to ICD9, and then map these into the MED via the ICD9 code (which is retained as another term attribute, or slot, in the MED). Also, initial efforts toward implementing a statistical n-gram algorithm (considers occurrence of 2 letter "2-gram" sequences, 3 letter "3-gram" sequences, etc.) to accomplish the same mapping task are nearly completed, and there is interest in implementing the least squares fit mapping technique as reported by Yang.

References

1. Shea S, Clark AS, Clayton PD. Columbia-Presbyterian Medical Center Integrated Academic Information Management System (IAIMS) outpatient clinical information system implemented in a faculty general medicine practice. *Proceedings of the 14th SCAMC*, Washington, DC, 1990; 730-734.
2. Hripcsak G, Clayton PD, Cimino JJ, Johnson SB, Friedman C. Medical decision support at Columbia-Presbyterian Medical Center. IMIA Working Conference on Software Engineering in Medical Informatics, Amsterdam, Netherlands, 1990.
3. Cimino JJ, Clayton PD, Hripcsak G, Johnson SB. Knowledge-based approaches to the maintenance of a large controlled medical terminology. *JAMIA* 1994; 1(1):35-50.
4. Friedman C, Sidelli R. Tolerating spelling errors during patient validation. *Comput Biomed Research* (October 1992); 25(5):486-509.
5. Yang Y, Chute C. An application of least squares fit mapping to clinical classification. *Proceedings of the 16th SCAMC*, Baltimore, MD. McGraw-Hill, Inc., 1992; 460-464.
6. Sheretz DD, Tuttle MS, Olson NE Erlbaum MD, Nelson JS. Lexical mapping in the UMLS Metathesaurus. *Proceedings of the 13th SCAMC*, Washington, DC, 1989; 494-99.
7. Wingert F. An indexing system for SNOMED. *Meth Inform Med* 1986; 25:22-30.
8. Kimbrell RE. Searching for text? Send an n-gram. *Byte* (May 1988):297-312.
9. Cimino JJ, Barnett GO. Automated translation between medical terminologies using semantic definitions. MD Computing 1990; 7(2):104-9.
10. Chute CG, Yang Y, Evans DA. Latent semantic indexing of medical diagnoses using UMLS semantic structures. *Proceedings of the 15th SCAMC*, Washington, DC, 1991; 185-189.
11. Cimino JJ, Barrows RC, Allen BA. Adapting ICD9-CM for clinical decision support. (abstract) American Medical Informatics Association Spring Congress, 1992.