Mapping Medical Vocabularies to the Unified Medical Language System

Qing Zeng M.S., James J. Cimino, M.D.
Department of Medical Informatics
Columbia University, New York, New York

This paper presents our work in automated mapping of medical vocabularies to the National Library of Medicine's Unified Medical Language System (UMLS). We used the UMLS Knowledge Source (KS) tool to map terms from several sources to UMLS Metathesaurus concepts. We compared performance of the KS tools with our own Minimal Representable Units Method (MRUM). The KS tools were able to map terms from 13% to 54% of the time, depending on the term set and the KS options used. Our MRUM method mapped between 96% and 99% of the terms. Based on our experience, we believe that questions remain about the best method by which the UMLS can be used to achieve automated term translation.

INTRODUCTION

Automated transfer of information between medical systems is a current area of medical informatics research. Our own work has focused on the linking of various on-line information resources with the clinical information system (CIS) at the Columbia Presbyterian Medical Center (CPMC). Other work from our center has linked the CIS to Medline[1] and DXplain.[2,3] Current work is directed at linking patients' x-ray reports to on-line information sources that may be relevant to the information needs of caregivers reading those reports, including a text database describing radiographic findings (the Chorus project at the Medical College of Wisconsin[4]) and to a database of pathology images (the Internet Pathology Lab for Medical Education (IPLME) at the University of Utah[5]).

We are currently working with x-ray reports that have been processed through a natural language processor[6] into terms coded in our Medical Entities Dictionary (MED).[7] The MED is a controlled vocabulary used locally at CPMC, whereas the vocabularies used by Chorus and IPLME are local non-controlled vocabularies of their respective institutions. The use of differing vocabularies is a well-known barrier to communication between applications. To address this, we are exploring the use of the National Library of Medicine's (NLM) Unified Medical Language System (UMLS)[8] to help translate terms from one vocabulary to another. A first step in this translation is mapping the diverse vocabularies into the UMLS. Algorithms for specific translations have been developed in the past;[9,10] however, no general solution to translation using the UMLS has been described. The NLM now provides Knowledge Source Server (KS) Application Programming Interface (API) tools to retrieve information from the UMLS.[11] In this paper, we describe our work using the KS tools and our own method for mapping terms to the UMLS.

METHODS

Sets of terms were obtained for each of the three vocabularies. CPMC terms were obtained from the MED; Chorus terms were obtained by conducting a search of their interface search engine using stemming ("a*", "b*", ... "z*"); and IPLME terms were obtained by traversing their index and selecting all hyperlinks. All terms were preprocessed to standardize their form. This preprocessing consisted of: 1) conversion to all lower case, 2) removal of all possessives ("'s"), 3) replacing all punctuation with single space, and 4) reducing all white space to single space. All duplicate terms from each set were removed after preprocessing, and we produced a version of each term in which prepositions ("of", "with", "in") and conjunctions ("and", "or") are removed.

The UMLS KS provides API programs that perform queries on UMLS knowledge sources, including the Metathesaurus (Meta), Semantic Network, Specialist Lexicon and Information Source Map. The API client programs can be obtained through anonymous ftp to "lhc.nlm.nih.gov". In this study we mostly used the Metathesaurus function. The Metathesaurus API is given a term or a file of terms with various query options. It returns requested information such as concept name, concept unique identifier, etc.

The options we used are:

-c looks for concepts with exact matched name. A search for "atelectasis" returns: "Query Term: atelectasis ; Concept Name: Atelectasis; UI: C0004144"

0195—4210/96/$5.00 © 1996 AMIA, Inc.

-cv looks for concepts which have name or variants of the name that match the term. A search for "atelectasis" returns: "Concept Name: Atelectasis; Lexical Variant: ATELECTASIS; Lexical Variant:: atelectasis"

-tv looks for concepts which match the term name or variants of the term name. A search for atelectasis returns: "Term Name: Atelectasis; Lexical Variant: ATELECTASIS; Variant Tag: ; Source: COS92/PT/U000054"

-ns uses normalized string index (see UMLS documentation for details).[8,12] A search for "atelectasis" returns: "Atelectasis".

-nw uses word index and looks for concepts which have a name which contains all the words in the term. A search for "atelectasis" returns: "Atelectases; Atelectases, Congestive; atelectasis; Atelectasis neonatorum; Atelectasis, complete; Atelectasis, compression; Atelectasis, Congestive; Atelectasis, discoid"

We used the KS API to obtain matching Meta concepts for each term from the CPMC chest x-ray reports, Chorus, and IPLME. We used the preprocessed terms with option -c, -cv, -tv, -ns, and -nw. All searches were carried out using terms without removal of prepositions and conjunctions. In addition, we searched for the terms after preposition and conjunction removal using the -ns option. We also searched to the terms which could not be found using the -ns option by using the -nw option.

We also developed our own mapping method using UMLS resources combined with our own keyword synonym table. Specifically, it uses the UMLS to represent terms by a Minimal Representable Units Method (MRUM). To implement MRUM: 1) the UMLS MRCON[8] table (The UMLS file that contains information of concept names) is preprocessed in the same manner as the terms to be mapped; (e.g. , all letters are converted to lower case, all punctuation is replaced with single space, etc..); 2) the words in each MRCON concept name are sorted alphabetically for normalized search, e.g. "renal infarction" became "infarction renal"; 3) a local keyword synonym table was used to normalize MRCON and the source terms (This synonym table is a collection of the synonym information extracted from UMLS and some local expert knowledge of synonyms); After 1) 2) 3), MRCON was transformed to MRCON'; and 4) a lexeme table was extracted from UMLS Specialist Lexicon as a supplement to MRCON.

MRUM breaks down a term into a minimal number of units. (The units are the largest representable units.) Each unit can be mapped to a UMLS concept or lexicon. The algorithm is:

Given an input string A, let X = A,

1) Generate an alphabetic sort of X which is called X'.

2) Search MRCON' for a corresponding concept for X'. If found, output the concept and go to 3. If not found, search the lexicon table for a corresponding term for X. If found, output the term and go to 4.

3) Remove the last word from the right end of X (not X'), if X is not empty, go to 1, otherwise quit.

4) Let X = A - X, If X is empty, quit. If not, go to 1.

For example, "acute kidney infarct" after normalization becomes "acute renal infarction". The alphabetically sorted form of this term ("acute infarction renal") does not match any normalized Meta concept or lexeme (an entry in Specialist Lexicon). "Infarction" is removed, but the alphabetically sorted form of "acute renal" still can't be mapped to a Meta concept or a lexeme. "renal" is removed, leaving "acute" which can't be mapped a Meta concept, but it can be mapped to the lexeme "Acute" (E0007127). "renal infarction" is left and its alphabetically sorted form ("infarction renal") is mapped to the Meta concept "Renal Infarction" (C0035085) by seraching MRCON'. So "acute kidney infarction" is represented with the pair of codes <E0007127><C0035085>. A variation of this step of the algorithm is to remove the first word from the left end. The result, for the above example, is identical. In fact, the results for both methods are generally similar.

We used MRUM for CPMC, Chorus, IPLME terms. We also implemented MRUM without our local keyword synonym table and used for CPMC, Chorus, IPLME. All conjunctions and prepositions were removed from the terms.

RESULTS

The results of mapping are shown in Table 1. As expected, successively more permissive KS options provided more matches for each source (nw>ns>tv>tw>c). Removal of prepositions and conjunctions is more permissive but had very little absolute effect on retrieval. Similarly, more permissive options also returned more concepts per match. Except in "-nw", it is ignorable—normally less than 1.01 concepts per term.

Table 1: Results of mapping by various methods

No. of Terms	CPMC	Chorus	IPLME
Raw files	1472	1122	1058
Preprocessed	652	793	1055
Mapped using the KS -c	196 (30.06%)	261 (32.91%)	133 (12.61%)
Mapped using the KS -tv	210 (32.21%)	292 (36.82%)	153 (14.50%)
Mapped using the KS -cv	210 (32.21%)	292 (36.82%)	153 (14.50%)
Mapped using the KS -ns	218 (33.44%)	275 (34.68%)	190 (18.01%)
Mapped using the KS -nw	352 (53.99%)	324 (40.86%)	216 (20.47%)
Mapped using KS -ns after prepositions are removed	226 (34.66%)	277 (34.93%)	202 (19.15%)
Mapped using MRUM	629 (96.47%)	791 (99.74%)	1050 (99.53%)

Table 2: Results of mapping using the KS -nw option

No. of Concepts per Matched Term	CPMC	Chorus	IPLME
Using -nw alone	87.90	4.00	25.60
Using -ns and -nw	50.83	5.22	5.48

Searching using "-nw" returns significantly more concepts per search term. This problem can be reduced by matching with an exact-match method first (e.g. "-ns") and then using the "-nw" method for those terms which do not match. The results are shown in Table 2.

The MRUM method matched substantially more terms than any KS method, including the -ns/-nw combination method. Table 3 shows the breakdown of MRUM matches with respect to number of term concepts needed to represent source terms.

Table 4 shows the results of using the MRUM method. Here, the data are shown based on the number of concepts or lexical variants (from MRCON) needed to represent each term. "No Match" corresponds to zero lexemes or concepts found. Because MRUM uses a non-UMLS resource (our keyword synonym table), we also conducted the match without using the table. The results, as shown in Table 4, indicate that the synonym table adds only a small increment to performance.

Table 3 The breakdown of MRUM (using keyword synonym table) matches with respect to the number of concepts or lexemes per term

Represented by:	CPMC	Chorus	IPLME
0 lexeme or concept	23	2	5
1 lexeme or concept	348	351	209
2 lexemes or concepts	186	269	355
3 lexemes or concepts	64	136	354
4 lexemes or concepts	23	29	136
5 lexemes or concepts	5	6	65
6 lexemes or concepts	2	0	23
7 lexemes or concepts	1	0	7
8 lexemes or concepts	0	0	1
No. of lexemes/ concepts for matches	1.67	1.82	2.61
Average number of words per term	2.08	2.70	3.99

Table 4 The breakdown of MRUM (without using keyword synonym table) matches with respect to the number of concepts or lexemes per term.

Number of terms Represented by	CPMC-CXR	Chorus	IPLME
0 lexeme or concept	23	2	5
1 lexemes or concepts	347	342	204
2 lexemes or concepts	185	259	344
3 lexemes or concepts	64	150	265
4 lexemes or concepts	26	33	139
5 lexemes or concepts	4	7	64
6 lexemes or concepts	2	0	26
7 lexemes or concepts	1	0	7
8 lexemes or concepts	0	0	1
Average no. lexemes or concepts returned for terms that can be represented	1.67	1.86	2.64
Average number of words per term	2.08	2.70	3.99

DISCUSSION

Our ultimate goal is to automate the translation between vocabularies of clinical applications and medical information resources. We want to avoid the need for human intervention in the translation process if possible. Mapping to a common vocabulary such as the UMLS is a logical first step in this process. We explored two mapping methods: the NLM's KS API and our own MRUM.

The UMLS KS search provides a method usable by anyone to map source terms into the UMLS. We only used KS to map terms to Meta concepts. As the results of our study showed, some methods perform better than others. However, the results of methods are not significantly different. Except for the "-nw" method, the mapping is accurate and requires no human intervention. The major limitation of the KS method is the amount of knowledge UMLS contains, e.g. the number of concepts and lexical variants.[13]

MRUM provided improved recall and precision over KS tools. There are several reasons for this. First, we map one term to multiple Meta concepts and lexemes instead of mapping one term to one concept. This enables us to map a term even when there is no single concept in the UMLS which matches the meaning of the term. Second, we used our own keyword synonym table. The knowledge of synonyms enables our program to map the various names of a concept to the concept. Results in Table 3 and 4 show that less concepts or lexemes are needed to represent a term when using keyword synonym table because more concepts names are mapped to single concepts (e.g., "kidney infarction" is mapped to <renal infarction> instead of <kidney><infarction> because in the synonym table "kidney" and "renal" are synonyms). Third, we transformed all terms names and concept names to a uniform format to increase the mapping rate (e.g., "X ray" and "x-ray" are transformed to "x ray"). Fourth, MRUM allows partial mapping. For example, even if "acute" was not mapped, "acute renal infarction" would be mapped to "renal infarction". Partial mapping lowers accuracy but allows us to capture as much information as possible. Low accuracy could be a potential problem. However, we feel that in this case capturing as much meaning from a term as possible provides us with a base for further mapping or searching.

There are a number modifications which might provide further improvement in MRUM's performance. One apporach involves using the semantic and syntactic knowledge in the UMLS to split terms into representable units and representing terms using UMLS concepts or lexemes. E.g. "acute renal infarction" can be represented as <<modifier: acute><finding: <renal infarction>> instead of <acute><renal infarction>.

The comparison of the two mapping methods also provides an evaluation of UMLS content. When maping directly with the KS tools, one might come to the conclusion that the UMLS lacks breadth of content - that is, it seems to lack the majority of terms found in three different medical information sources. However, the MRUM results show that the problem is not so much breadth as it is depth: the granularity of the UMLS concepts does not match that which is found in medical applications. One way to address this situation is to continue the lexical matching process which has been used thus far to construct the Metathesaurus. However, a more expedient approach might be to direct research toward solving a problem which we believe will continue to exist no matter how much content is added to the Metathesaurus: mapping between vocabularies of different granularities.

Although our mapping results show that MRUM is a useful method, the ultimate evaluation will be to use MRUM to map CPMC terms to Chorus and IPLME terms. Our initial attempts seems to be promising, however, further research has yet to be conducted.

CONCLUSION

The ability to use the UMLS to translate between controlled medical vocabularies requires the initial step of mapping an external vocabulary into the UMLS. The Knowledge Source API from the UMLS is the first publicly-available tool for this purpose. Our study shows that the addition of some simple algorithms to the tool set has the potential for improving such mapping.

Acknowledgments

This work was supported by UMLS contracts from the National Library of Medicine (NLM 95-053/VMS, 001 LM05857-01). The authors thank Steve Johnson, Ph.D. for his advice and for graciously sharing various utility programs he developed, Carol Friedman, Ph.D. for her advice, and John Olyarchuk, M.D. for editorial assistance.

References

1. Cimino JJ. Johnson SB. Aguirre A. Roderer N. Clayton PD. The MEDLINE Button. Frisse ME, ed. *Proceedings the Annual Symposium on Computer Applications in Medical Care* :81-5, 1992.

2. Barnett GO, Cimino JJ, Hupp JA, Hoffer EP: DXplain. An evolving diagnostic decision-support system. *Journal of the American Medical Association*; July 3, 1987; 258(1):67-74.
3. Elhanan G, Socratous SA, Cimino JJ Integrating DXplain into a Clinical Information System using the World Wide Web. In Cimino, JJ, ed.: *Proceeding of the AMIA Annual Fall Symposium*; Washington, DC; October, Hanley and Belfus, Philadelphia, 1996:this volume.
4. Kahn CE Jr. CHORUS: a computer-based radiology handbook for international collaboration via the World Wide Web. *Radiographics*. 15(4):963-70, 1995 Jul.
5. Klatt EC. The "Electronic City" in the Laboratory. *Laboratory Medicine* 1996; 27(2):117-121.
6. Friedman C. Alderson PO. Austin JH. Cimino JJ. Johnson SB. A general natural-language text processor for clinical radiology. *JAMIA*. 1994 ; 1(2):161-74.
7. 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.
8. National Library of Medicine. *UMLS Knowledge Sources - 6th Experimental Edition Documentation*. Bethesda, Maryland: The Library, 1995 Apr.
9. Barrows RC Jr. Cimino JJ. Clayton PD. Mapping clinically useful terminology to a controlled medical vocabulary. In Ozbolt JG, ed.: *Proceedings of the Eighteenth Annual Symposium on Computer Applications in Medical Care*; Washington, DC; November, McGraw-Hill, New York, 1994:211-215.
10. Cimino JJ, Johnson SB, Peng P, Aguirre A: From ICD9-CM to MeSH using the UMLS: A How-to Guide. In Safran, C, ed.: *Proceedings of the Seventeenth Annual Symposium on Computer Applications in Medical Care*; Washington, DC; November, McGraw-Hill, New York, 1993:730-734.
11. McCray AT, Razi, A. The UMLS Knowledge Source Server. *In Kaihara S, Greenes RA, eds. Proceedings of the World Congress on Medical Informatics - Medinfo '95*; Vancouver, Canada; Healthcare Computing and Communications Canada, Edmonton, Alberta, 1995: 144-147.
12. McCray AT. The Unified Medical Language System. Lindberg DA. Humphreys BL. *Methods of Information in Medicine*. 32(4):281-91, 1993 Aug.
13. Friedman C. The UMLS coverage of clinical radiology. In: Frisse ME, ed. *Proceedings of the Sixteenth Annual Symposium on Computer Applications in Medical Care*. New York: McGraw-Hill, 1992:309-13.