Utilizing the UMLS for Semantic Mapping between Terminologies

Kin Wah Fung MD, MSc, MA and Olivier Bodenreider MD, PhD

National Library of Medicine, Bethesda, Maryland

{kwfung|olivier}@nlm.nih.gov

An algorithm was derived to find candidate mappings between any two terminologies inside the UMLS, making use of synonymy, explicit mapping relations and hierarchical relationships among UMLS concepts. Using an existing set of mappings from SNOMED CT to ICD9CM as our gold standard, we managed to find candidate mappings for 86% of SNOMED CT terms, with recall of 42% and precision of 20%. Among the various methods used, mapping by UMLS synonymy was particularly accurate and could potentially be useful as a quality assurance tool in the creation of mapping sets or in the UMLS editing process. Other strengths and weaknesses of the algorithm are discussed.

INTRODUCTION

Despite decades of work in the medical informatics community, a universally accepted standard medical terminology remains an elusive goal. This has been cited as one of the greatest impediments to the widespread development of electronic medical records. [1] In view of the range of functions that standard terminologies have to serve, including direct patient care, billing, statistical reporting, automated decision support and clinical research etc., it is doubtful that a single terminology will ever be deemed suitable for all purposes. [2] Whenever information captured in one terminology is reused for another purpose that requires coding in a different terminology, there is a need for inter-terminology mapping. Mapping between terminologies is a labor intensive process and there is strong incentive to automate as much of it as possible. Broadly speaking, automatic mapping between terminologies can be divided into lexically-based and semantically-based methods. [3-8] This study explores the use of semantic information in the Unified Medical Language System® (UMLS®) to map between two clinical terminologies.

BACKGROUND

One important aim of the UMLS is to establish connections between disparate biomedical terminologies. [9, 10] This is achieved by incorporating them into a Metathesaurus organized on the basis of a 'concept' – a unit of meaning. Concept names (more commonly referred to as terms) from various terminologies that represent the same meaning are encompassed by the same UMLS concept. In the 2004AA version of the UMLS (the version used in this study), there are over one million concepts, 2.8 million distinct strings from over 100 source terminologies.

Generally speaking, creating a mapping between two terminologies is to find, for each term in one terminology (the source terminology), the term that has the closest meaning in the other (the target terminology). Provided both terminologies are in the UMLS, there are several ways in which candidate mappings can be discovered.

Synonymy. The most direct way is through synonymy. If both the source term and any of the terms in the target terminology are in the same UMLS concept, a mapping is found.

Explicit mapping relations. Another UMLS resource that can be utilized for mapping is the explicit mapping relations provided by some source terminologies. Inter- and intra-terminology mappings are created by some source terminologies for various purposes and they are incorporated into the UMLS. Some examples are: SNOMED CT to ICD9CM mappings provided by SNOMED CT and ICD9CM to ICD10AM mappings provided by ICD10AM. These mappings can be found in the MRREL file in the Metathesaurus and most of them can be identified by their relationship attributes (e.g. mapped_from/to, primary_mapped_from/to, other_mapped_from/to). The source vocabularies contributing explicit mapping relations are shown in Table 1.

Source terminology	Number of explicit mapping rel.
Medical Subject Headings (MeSH)	* 467,581
SNOMEDCT	67,513
SNOMED International	32,348
Medical Dictionary for Regulatory Activities	26,296
Intl. Classification of Diseases (ICD10AM)	23,955
ICPC2E-ICD10 relationships	12,384
CRISP Thesaurus	11,743
COSTART	6,509
Others	15,028
Total	663,357

* these are mappings between chemical entities within MeSH

Table 1. Sources of explicit mapping relations in the UMLS (refer to UMLS documentation for source name abbreviations)

When an explicit mapping relationship exists between two terms T_1 and T_2 in two UMLS concepts C_1 and C_2, it is likely that all terms in C_2 are also potential mapping targets for all terms in C_1. In other words, specific mappings between two terminologies can be 'reused' for other terminologies by means of the UMLS concept structure.

Ancestors expansion. A third UMLS resource that can be utilized in mapping is hierarchical relationships. Quite often mappings are done from a more granular to a less granular terminology. In such cases, sometimes a term and its children are mapped to the same term in the target terminology. If one fails to find mapping for a source term, it is possible that the mapping of its ancestors may be a candidate. By using the UMLS hierarchical relationships, one can walk up the hierarchy from a source term to extend the search for candidate mappings. Note that in walking up the UMLS hierarchy (as opposed to walking up the hierarchy in a source terminology), one is not restricted to concepts containing terms from the source terminology. In effect, this increases the number of ancestors used in searching for a potential map.

Children and siblings expansion. Further extension of the search base using hierarchical relationships is possible. For example, one can expand the pool of the ancestors by including ancestors of the children of the source concept (i.e., first walking down the hierarchy before walking up). Analogously, ancestors of the siblings of the source concept can also be included. However, the wider the net that is cast, the higher the level of noise and some restriction mechanisms will be necessary (see below).

METHODS

Mapping algorithm

One of the authors (OB) has developed earlier an algorithm for semantic mapping from UMLS concepts to MeSH terms. [11] This algorithm was generalized to map between any two terminologies in the UMLS. Further flexibility was added to allow inclusion or exclusion of specific relationships according to their type and source. The search algorithm is carried out sequentially as follows (search will stop if a candidate mapping is found):

1. A target term is in the same concept as the source term.
2. A target term is in a target concept linked to the source concept by an explicit mapping relationship.
3. A target term is found through any of the ancestors of the source concept (either by synonymy or by explicit mapping relations of the ancestors). In this process, ancestors are required to be semantically compatible with the source concept. Moreover, target concepts that are ancestors of other target concepts are excluded to ensure that the closest target term is found.
4. Finally, the graph of the ancestors can be seeded not with the source concept itself, but with its children (or siblings). In this case, candidate target concepts are required to be linked to at least 75% of the children (or siblings) of the source concept in order to prevent semantic drift.

Since a directed acyclic graph is required for computing lists of ancestors for Metathesaurus concepts (otherwise infinite looping may occur), we use a slightly modified version of the Metathesaurus from which the links responsible for the cycles in hierarchical relations have been removed. [12]

Evaluation

To evaluate the algorithm, we used it to map SNOMED CT terms to ICD9CM. The 2004AA release of the UMLS was the first to include SNOMED CT. In addition to concepts and their interrelations, SNOMED CT also provides a set of mappings from SNOMED CT to ICD9CM. This would serve as our gold standard. All the SNOMED CT terms with one-to-one ICD9CM mappings in the gold standard were fed through the mapping algorithm. The mapping was done with specific instructions to ignore the explicit mapping relations to ICD9CM provided by SNOMED CT. The resulting mappings were compared to the gold standard. Precision is the percentage of suggested mappings that were correct. Recall is the overall percentage of correct mappings that were found. As some applications (e.g. statistical grouping) might require a lower degree of accuracy than others, the results were also evaluated up to the 3rd and 4th ICD9CM digits level. Additionally, a small randomly selected subset was examined manually in greater detail.

RESULTS

Quantitative results

A total of 66,382 SNOMED CT terms had one-to-one mappings to ICD9CM terms in the gold standard. Our algorithm successfully found candidate mappings in 56,830 of them (85.6%). Altogether 137,134 distinct mappings (pairs of SNOMED CT and ICD9CM codes) were generated, with an average of 2.4 mappings per SNOMED CT term.

Table 2 summarizes the results. The bulk (97%) of mapped SNOMED CT terms was mapped through synonymy, explicit mapping relations and ancestors expansion (i.e., steps 1-3 in our algorithm). While the number of mapped SNOMED CT terms increased in this order, the quality of the mappings decreased. Mapping by synonymy constituted 12.6% of mapped SNOMED CT terms and had the highest category

precision (74.5%). Ancestors expansion mapping accounted for the highest proportion of mapped SNOMED CT terms (56.8%) but had low category precision (8.7%). Mapping through explicit mapping relations was somewhere in between,. Overall, the algorithm had recall of 42% (27,850/66,382) and precision of 20%. The last two rows of Table 2 show how the percentage of correctly mapped SNOMED CT terms would increase if a lower degree of accuracy was required.

	Method of mapping					Overall
	Synonymy	Explicit Mapping	Ancestors expansion	Children expansion	Siblings expansion	
SNOMED CT terms mapped (% of total mapped)	7,148 (12.6%)	15,639 (27.5%)	32,286 (56.8%)	458 (0.8%)	1,299 (2.3%)	56,830 (100%)
Mappings per SNOMED CT term (category precision = correct mappings/total mappings in category)	1.1 (74.5%)	1.5 (55.1%)	3.1 (8.7%)	2.7 (10.9%)	2.8 (9.0%)	2.4 (20.3%)
Correctly mapped SNOMED CT terms (% within category)	6,022 (84.2%)	12,527 (80.1%)	8,835 (27.4%)	134 (29.3%)	332 (25.6%)	27,850 (49.0%)
Correctly mapped SNOMED CT terms up to 4th ICD9CM digit (% within category)	6,543 (91.5%)	13,662 (87.4%)	12,169 (37.7%)	149 (32.5%)	386 (29.7%)	32,909 (57.9%)
Correctly mapped SNOMED CT terms up to 3rd ICD9CM digit (% within category)	6,933 (97%)	14,481 (92.6%)	20,435 (63.3%)	220 (48%)	721 (55.5%)	42,790 (75.3%)

Table 2. Proportion and accuracy of mapping according to method

Detailed analysis of a subset

A randomly selected subset of 500 SNOMED CT terms and their mappings were analyzed in more detail. Among these, 431 terms (86.2%) were successfully mapped. The majority of mappings came from synonymy (13.7%), explicit mapping relations (29.0%) and ancestors expansion (55.5%). The numbers were very close to the full set, suggesting that this was a representative sample. The analysis focused on cases in which the algorithm failed to find the correct mappings. This we felt would be more informative than the successful cases most of which would not require further explanation.

Mapping by synonymy

Fifty-nine SNOMED CT terms (*sct*) were mapped by this method, of which 55 (93%) were mapped correctly. Four SNOMED CT terms mapped by synonymy did not match the gold standard. In some of these cases, our suggested mappings (*sm*) were in a sense more accurate than the gold standard mappings (*gs*). In the second and third examples below, *gs* was narrower in meaning than *sct*. This might be due to the possible requirement that *gs* mappings must be made to the lowest level of ICD9CM terms available (ICD9CM codes are shown in brackets):

sct: Phytanic acid storage disease
gs: Other disorders of lipoid metabolism (272.8)
sm: Refsum's disease (356.3) (this is a synonym of Phytanic acid storage disease) [13]

sct: Fetal and neonatal hemorrhage
gs: Unspecified hemorrhage of newborn (772.9) (this excludes fetal blood loss (772.0))
sm: Fetal and neonatal hemorrhage (772)

sct: Open wound
gs: Open wound (s) (multiple) of unspecified sites(s) without mention of complication (879.8) (this excludes limb wound because it is a child of Open wound of other and unspecified sites, except limbs (879))
sm: Open wound (870-879.99)

The remaining case was:

sct: Retinal defects without detachment
gs: Retinal defect, unspecified (361.30)
sm: Retinal defects without detachment (361.3)

Normally, 361.30 would be put in the same UMLS concept as 361.3 because they were not semantically different. This would have allowed our mapping algorithm to find the correct mapping. However, the mapping to 361.30 was missed because it was not placed in the same UMLS concept as 361.3, which was probably an error in UMLS editing.

Mapping by explicit mapping relations

A total of 125 SNOMED CT terms were mapped by this method, of which 106 (85%) were mapped correctly. Among the 19 SNOMED CT terms for which the gold standard mappings were not found, there were four cases in which *sm* might be considered better than *gs*, for example:

sct: Propionic acidemia, type I
gs: Unspecified disorder of metabolism (277.9)
sm: Disturbances of branched-chain amino-acid metabolism (270.3) (more specific mapping than *gs*)

In five cases, *gs* and *sm* were normally expected to be in the same UMLS concept but were found in sepa-

rate UMLS concepts, leading to *gs* being missed, for example:

sct: Submammary mastitis associated with childbirth
gs: Abscess of breast associated with childbirth, unspecified as to episode of care (675.10)
sm: Abscess of breast associated with childbirth (675.1)

In four cases, an explicit mapping relation to a *sm* less specific than *gs* existed in the UMLS, for example:

sct: Neoplasm of bone
gs: Neoplasm of unspecified nature of bone, soft tissue, and skin (239.2)
sm: Neoplasms of unspecified nature (239)

The mapping from 'Neoplasm of bone' to 'Neoplasm of unspecified nature' was provided by Coding Symbols for Thesaurus of Adverse Reaction Terms (COSTART) to map its terms to WHO Adverse Drug Reaction Terminology (WHOART). While this mapping was appropriate from COSTART to WHOART, it was not specific enough for mapping to ICD9CM, which has finer granularity than WHOART.

In three cases, *sm* was different from *gs* but was considered an acceptable alternative, for example:

sct: Irritability and anger
gs: Other mental problems (V40.2)
sm: Other ill-defined conditions (799.89)

In one case, a dubious explicit mapping relation resulted in this mapping:

sct: Hypertrophy of bile duct
gs: Other specified disorders of biliary tract (576.8)
sm: Spasm of sphincter of Oddi (576.5)

Finally, an explicit mapping relationship from a less specific concept to a more specific concept existed which resulted in an overly specific mapping:

sct: Neoplasm of pancreas
gs: Neoplasm of unspecified nature of digestive system (239.0)
sm: Malignant neoplasm of pancreas (157)

Mapping by ancestors expansion

Among the 239 SNOMED CT terms mapped by this method, 69 (28.9%) were correctly mapped. All but one of the correct mappings came from distance-1 (parent) and distance-2 (grandparent) ancestors. The average distance from source concept for correct mappings was 1.3 (2.2 for incorrect mappings). If we only used distance-1 and distance-2 ancestors, the category precision would increase from 10% to 14.5%.

Mapping by children and siblings expansion

These were not further analyzed as the yield from these categories was too small to be significant.

DISCUSSION

The demand for mapping between standard terminologies is on the rise. This is partly fueled by the increase in the electronic capture of healthcare information (which usually requires a granular clinical terminology like SNOMED CT), and the demand to reuse the same data for other purposes such as reimbursement (which often requires coding in a statistical terminology like ICD9CM). Using our algorithm we found candidate mappings for 86% of SNOMED CT terms to ICD9CM terms with recall and precision of 42% and 20% respectively. These numbers should be considered a lower bound of the real performance because of the existence of imperfections in the gold standard and alternative valid mappings, as revealed by our detailed analysis. In another run of the mappings, we also included MTHICD9, which contained entry terms to ICD9CM terms, as a target terminology. This increased the proportion of mappings by synonymy and decreased that by explicit mapping relations. However, the overall performance was very similar, with 86% of SNOMED CT terms mapped, and recall and precision of 43% and 22% respectively.

Some tweaking of the algorithm is possible. If we only used mapping by synonymy and explicit mapping relations, we could increase the precision to 60% but the recall would drop to 28% and we could only find candidate mappings for 34% of SNOMED CT terms. Mapping by ancestors expansion accounted for over half of all mapped SNOMED CT terms but its precision was low (8.7%). It seems that we could improve the precision without significantly affecting the recall if we restrict the expansion to close ancestors.

Another potential way to improve performance is by specifically allowing or disallowing certain explicit mapping relations according to the source terminologies they originated from. This is a feature of our algorithm and can be easily accomplished. Mappings are created for different purposes which affect the degree to which some mappings can be reused for other mapping tasks. Ignoring less specific mappings from some sources could potentially improve the precision in the context of our study.

The presence of explicit mapping relations from SNOMED International (SNMI, the precursor of SNOMED RT which was combined with Clinical Terms Version 3 to form SNOMED CT) to ICD9CM in the UMLS has certainly improved our results. To

quantify this, the mapping process was repeated without using the explicit mappings relationships from SNMI. We still managed to get mappings for 85% of SNOMED CT terms to ICD9CM terms. However, the overall recall and precision dropped to 29% and 11% respectively. Apart from the fact that SNMI is closely related to SNOMED CT, SNMI explicit mappings relationships also constituted 25% of all potentially useful explicit mapping relations among clinical terms (not counting those from MeSH which were between chemical entities and those from SNOMED CT which were always turned off in our study). Therefore, it was not surprising that ignoring them had significant impact on performance. However, for other mapping tasks between clinical terminologies, the number of potentially useful explicit mapping relations will be 50% more than what we used in our study (as SNOMED CT explicit mapping relations will be used), which is likely to have a positive impact on performance.

Apart from the automatic suggestion of inter-terminology mappings, our algorithm can potentially be used to supplement existing quality assurance tools in the creation of mappings or in the UMLS editing process. This is particularly true for mappings found by UMLS synonymy. Theoretically, no mapping can be better than one in which the source and target terms are considered synonymous. This is true to the extent that in our study, whenever there was discrepancy between mappings generated by UMLS synonymy and the gold standard mappings, there was a high chance of finding problems either in the UMLS assertion of synonymy or the gold standard mappings (including sub-optimal mappings). Even though mappings found by explicit mapping relations are not as accurate as those found by UMLS synonymy, they are probably still accurate enough to have a role in quality assurance. In our study, there was a significant proportion of cases (9 out of 19 cases) in which discrepancy between our mappings found by explicit mapping relations and the gold standard led to the discovery of either problems in UMLS editing or the gold standard mappings.

Finally, SNOMED CT and ICD9CM have a significant degree of structural and content similarities which make them easier to map to each other. In future work, we will study the performance of our algorithm in mapping between other terminologies.

CONCLUSION

Making use of the semantic knowledge in the UMLS, our mapping algorithm successfully found candidate mappings for 86% of SNOMED CT terms to ICD9CM terms, with recall and precision of 42% and 20% respectively. While not accurate enough to support automatic data mediation on its own, this method can be a useful adjunct in the creation of mappings between any pair of vocabularies within the UMLS.

Acknowledgements

This research was supported in part by an appointment to the NLM Research Participation Program sponsored by the National Library of Medicine and administered by the Oak Ridge Institute for Science and Education.

References

1. United States General Accounting Office. Automated medical records: Leadership needed to expedite standards development. USGAO-IMTEC-93-17. Washington D.C., 1993.
2. Cimino JJ. Review paper: coding systems in health care. Methods of Information in Medicine 1996;35:273-84.
3. Cimino JJ, Barnett GO. Automated translation between medical terminologies using semantic definitions. MD Comput 1990;7:104-9.
4. Barrows RC, Jr., Cimino JJ, Clayton PD. Mapping clinically useful terminology to a controlled medical vocabulary. Proceedings - the Annual Symposium on Computer Applications in Medical Care 1994:211-5.
5. Dolin RH, Huff SM, Rocha RA, Spackman KA, Campbell KE. Evaluation of a "lexically assign, logically refine" strategy for semi-automated integration of overlapping terminologies. Journal of the American Medical Informatics Association 1998;5:203-13.
6. Rocha RA, Rocha BH, Huff SM. Automated translation between medical vocabularies using a frame-based interlingua. Proc Annu Symp Comput Appl Med Care 1993:690-4.
7. Masarie FE, Jr., Miller RA, Bouhaddou O, Giuse NB, Warner HR. An interlingua for electronic interchange of medical information: using frames to map between clinical vocabularies. Comput Biomed Res 1991;24:379-400.
8. Rocha RA, Huff SM. Using digrams to map controlled medical vocabularies. Proc Annu Symp Comput Appl Med Care 1994:172-6.
9. Lindberg DA, Humphreys BL, McCray AT. The Unified Medical Language System. Methods of Information in Medicine 1993;32:281-91.
10. Humphreys BL, Lindberg DA, Schoolman HM, Barnett GO. The Unified Medical Language System: an informatics research collaboration. J Am Med Inform Assoc 1998;5:1-11.
11. Bodenreider O, Nelson SJ, Hole WT, Chang HF. Beyond synonymy: exploiting the UMLS semantics in mapping vocabularies. Proceedings / AMIA ... Annual Symposium 1998:815-9.
12. Bodenreider O. Circular hierarchical relationships in the UMLS: etiology, diagnosis, treatment, complications and prevention. Proceedings / AMIA ... Annual Symposium 2001:57-61.
13. National Institute of Neurological Disorders and Stroke website. http://www.ninds.nih.gov/.