Enhanced Biomedical Taxonomy Mapping Through Use of A Semantic Measure of Proximity

Jeffery L. Painter
Computer Science Department, North Carolina State University, Raleigh, NC, USA
GlaxoSmithKline, Statistical & Quantitative Sciences, RTP, NC, USA

Abstract— *By employing a notion of semantic closeness to create a multi-phase mapping between one biomedical taxonomy and another, we are able to determine various levels of proximity for mapping term based structures which may fail to map using traditional mapping techniques. The multi-phase approach allows for defining the appropriateness of applying the maps to various domain problems such as the investigation of high level system organ or class effects versus a problem that requires a higher level degree of specificity in the analysis.*

Keywords: semantic mapping, meaning hierarchy, ICD-10 codes, MedDRA codes, UMLS Metathesaurus, terminology alignment

1. Introduction

For many exercises in biomedical data analysis, the need arises to map one medical terminology to another. Over the years, several methods (1) have been developed which attempt to solve the problem of accomplishing this goal in some automated procedure to reduce the effort necessary to process these sometimes large taxonomies which would otherwise require hundreds if not thousands of man hours to complete and would be prone to human error and fatigue.

In another paper (2), it was shown that it might be possible to also introduce a structure to previously unstructured terminologies by making use of "extra" information found within already structured terminologies.

In this paper, the purpose is to demonstrate how taking a *multi-phase* approach; the mapping from one terminology to another may be drastically enhanced to help solve any number of inferential analysis problems by looking at broader meaning relations among various terminologies in order to capture more relationships among the codes they each contain.

It is a *multi-phased* approach in that we attempt various methods for aligning, or matching, one terminology to another where many previous attempts at ontology alignment seek to only exploit one method or another. It is by the combination of several different methods, and developing a model of measure for relative closeness of mappings, that this enhanced procedure produces much more effective results.

By assigning a weight to indicate a sense of "closeness" between two terms or concepts, the mapping which we produce could then be applied in a manner described by Wang, Gong and Zhou (3) to create a composite model for ontology mapping.

Mappings of these types are being developed to enhance the data analysis of large observational databases such as electronic health records, claims data and other medical history databases used in both public and private systems such as those being developed by the Observational Medical Outcomes Partnership (OMOP) [1] initiative and the SafetyWorks (4) project which was initiated by GlaxoSmithKline. A common data model (5) allows for a standardization of the records across disparate data sources. However, the first step in creating these models is to *normalize* the data to a common, or reference, terminology. It is only by making use of ontologies and methods such as those described in this paper that we are able to achieve any level of success in *fitting* the data to the common data model.

The problem first encountered with the Read/OXMIS codes in (2) is re-evaluated with these enhanced methods in addition to mapping the ICD-10 codes to the Medical Dictionary for Regulatory Affairs (MedDRA)[2].

Our goal for mapping ICD-10 stemmed from the fact that the latest effort relating to the SafetyWorks project is to incorporate the IMS Germany database which is itself coded in ICD-10, German language variant. In order to incorporate this new database into the system, a mapping between ICD-10 and MedDRA (our reference terminology) needed to be constructed.

2. Term Based Matching

In previous work, we have taken advantage of many techniques previously described for aligning two ontologies, taxonomies or terminologies based on the terms occurring in each of them – from here on we will just call them terminologies since we are typically making use of the terms identified as representing the concepts in each. However, *term-based* mapping can be prone to errors due to the nature

[1] See http://omop.fnih.org/

[2] By 'ICD-10' we mean to refer to the International Classification of Diseases, 10th Revision, which is maintained by the World Health Organization. MedDRA® (Medical Dictionary for Regulatory Activities) is a registered trademark of the International Federation of Pharmaceutical Manufacturers Association. The Clinical Terms Version 3 (Read Codes)© are maintained by the (UK) National Health Service Information Authority.

of the representation employed within any given terminology system.

When one investigates the meaning being represented by an individual term, there may be some contextual information found in the hierarchy which does not appear directly in the term itself. For example, ICD-10 has several codes in the Yxx family which indicate possible poisoning or adverse effect of a particular agent, however the code's term may only list the agent and not explicitly state that it should have any adverse indication in that term. If you were trying to align ICD-10 to a terminology which also included a drug or medicament hierarchy, matching solely on the term occurence could place ICD-10 codes which are meant to represent an adverse reaction in the wrong position in the target terminology where the same terms represent only the substance and no mention of reaction, positive or negative.

Therefore, while we still employ term-based mapping within our *multi-phase* mapping process, those codes that can only be linked by a term-based match alone will be given the lowest level of "closeness" in our *hierarchy for semantic proximity*.

2.1 Hierarchy for Semantic Proximity

There are essentially four levels of "closeness" in our model which include:

1) Conceptual Level
2) Boosted Level
3) Nearest Neighbor Level
4) Term Level
    a) Direct Match
    b) Fuzzy Match

The conceptual level is considered as having the highest (or most relevant) degree of closeness between any two given entries in the two terminologies we are attempting to align, followed by the boosted level, the nearest neighbor level, and finally the term level (possessing the lowest degree of closeness) discussed briefly above.

After creating the mapping file between any two terminologies, we retain the semantic proximity information embedded within the mapping to enable the end user to filter mappings based on the particular need of precision in a given analysis. To speed the multi-phase mapping process (and to insure the highest level of semantic proximity is held between any two given terms), we execute the mapping in order from highest level to lowest level, with the exception of the term level.

The term level maps are generated first and are used in addition to the other methods. If a mapping can be found at any higher level, then the term map will be eliminated in favor of one which has a higher degree of semantic proximity.

Once a term from the source terminology is mapped at one of the higher levels, it is then removed from the set of terms to be evaluated for the lower level maps to follow.

2.1.1 Conceptual Level

The conceptual level of the hierarchy takes advantage of the UMLS Metathesaurus[3] in order to create a mapping between two terminologies. If the source and target terminologies are both found within the UMLS, this process is accomplished quite easily, and we can associate any two term entries by the presence of a shared concept unique identifier (CUI) which indicates synonymy within the UMLS conceptual model. If one terminology is not found within the UMLS, then an attempt can be made to identify "potential" conceptual maps by string matching. If two terms have the same exact string, in most instances, these are in fact representative of the same concept.

One way we were able to improve the mapping of the ICD-10 and Read/OXMIS codes was by incorporating mutliple sources from the UMLS. It may be the case that a term found in one terminology is present in another, even if it is not present in the target terminology. By matching the terms to those known to exist in the UMLS, we are still able to find a CUI which would link back to the source terminology and give us the necessary information in order to make a conceptual level match to the target terminology (in this case MedDRA).

When mapping the IMS Germany database, which is coded in ICD-10 (German), to MedDRA, the first step is to link each of the ICD-10 code entries to the version of ICD-10 found in the UMLS. Only one code from the IMS Germany database was unable to be mapped directly to ICD-10 in this way.

After converting the IMS Germany codes to their ICD-10 equivalents. We looked at the UMLS entries for each ICD-10 code that was applicable and identified it's CUI in the UMLS. If there existed a corresponding CUI for MedDRA, then we subsequently link the IMS Germany code to the MedDRA code. Any given CUI may link to several MedDRA entries, and from those, we choose a single MedDRA code based on some simple rules related to the term type associated with a particular MedDRA code.

Those selection rules favor MedDRA codes at the PT/LT level and if none are found, proceeds to climb the MedDRA hierarchy until a match is found. We do take care to note the problems commonly attributed to the MedDRA hierarchy (6) by eliminating duplicate LT entries when a PT entry is found. The selection criteria also involves multiple passes of the potential MedDRA concepts identified by CUI association to identify term-type (TTY) matches at various levels of the MedDRA hierarchy. If a preferred term (PT) is found, then preference is first given to that concept (since it exists at a more specific level of the MedDRA hierarchy). If the PT level is exhausted, the code proceeds to look further at the potential matches by trying to associate to a high-level term

[3]UMLS Metathesaurus is a project of the (US) National Library of Medicine, Department of Health and Human Services. Available at: http://www.nlm.nih.org/research/umls/

(HT), followed by a group term (HG) and lastly by looking for obsolete terms found in MedDRA which may be useful for determining levels of semantic proximity at a lower level than the conceptual level.

The entries identified by this first step in the *multi-phase* match are given a CONCEPT_MAP identifier in the semantic proximity embedding of our mapping file. 44% of the IMS Germany codes are mapped to the MedDRA target by the conceptual level, while only 16% of the Read/OXMIS codes are mapped to MedDRA at the conceptual level alone.

2.1.2 Boosting

Boosting is one of the more contraverisal methods for mapping codes from one terminology to another. Many of the biomedical terminologies have some inherent structure identifiable in the manner which the codes are constructed or by some external hierarchical structure that is annotated by the codes and terms found within that system.

The idea of boosting is to simply take advantage of this structure in an attempt to incorporate knowledge about *surrounding* codes that may have been mapped by either the conceptual map or a term level map. This idea is modified for terminologies, but deeply rooted in the notion of discovering proximity by relatedness such as discussed in (7).

When looking at the IMS Germany database, each data reference to an ICD-10 code is typically in the form of a 5 digit code.

The ICD-10 codes themselves are arranged in a hierarchy such that the first 3 and 4 digits of each code form various levels of a "family" of related codes. Boosting will atttempt to extract from an unmapped 5-digit code both it's 3 and 4-digit family levels and attempt to search for a conceptual mapping which was found at those higher levels. If a match iss found, then that boosted node will have a CUI associated with it which also occurs in the overall MedDRA CUI set. The process is then to associate the unmapped 5-digit code with the boosted relative's CUI and map it into the target terminology.

An additional 775 ICD-10 codes were mapped using this method. From the Read/OXMIS codes, an additional 11,648 mappings (approximately 14.9%) were created to the MedDRA target by using this method.

The maps produced by *boosting* generally provide a broader concept map to a lower level term than one might generally hope for, but it still allows many codes to be included which might not have otherwise been captured using traditional methods. For many of the types of analysis used in data mining observational databases, it is often the case that we only care about a higher level of generality (such as *liver disease, diabetes, etc*) anyway, rather than searching for individual lower-level conditions or diagnoses.

In some cases, we were able to create additional links by climbing the hierarchy beyond the parent level to the grandparent or great-grandparent in order to find a link back to the target terminology. These links were only applied if the level of the target terminology was still at the PT or LT level of the MedDRA hierarchy. In the case of the Read/OXMIS codes, it is not advisible to proceed past the grandparent level for boosting. The mappings produced beyond this level possessed numerous inaccuracies due to the fact that the Read terminology is multi-hierarchical and the linkages between higher levels of Read and MedDRA lead to inconsistencies in annotating the maps between code instances at a lower level to those in the target terminology which are found at a higher level.

From the GPRD Read/OXMIS terminology, an example of a code captured by the boosting method is: "K44..00 - Female gonococcal pelvic inflammatory disease" which was succesfully mapped to the broader MedDRA code "10034254 - Pelvic inflammatory disease". And from IMS Germany, an example boost match includes "E02 - Subclinical iodine-deficiency hypothyroidism" mapped to the MedDRA code "10043709 - Thyroid disorder".

Each of the mappings produced by the *boosting* method were evaluated manually for relevance of match, and were all deemed successful aside from those at the great-grandparent level for the Read codes. These entries are identified with a BOOST_MAP identifier in the semantic proximity embedding of our mapping file.

2.1.3 Nearest Neighbor

The last method developed for the multi-phase mapping process is the concept of using a *nearest neighbor* match. Both the IMS Germany codes and those found in Read possess an inherent hierarchical structure used to organize the content in each terminology. The nearest neighbor level matching attempts to look at codes relatively "close" to an unmapped code, that is they share at least the first three characters in their code designation, or by their hierarchical structure reside as siblings in the tree structure. If any of the neighboring codes were successfully mapped to MedDRA, then those links are used to enhance the mapping process to create additional links to the unmapped siblings.

To illustrate the success of this method, an IMS Germany example is given:

"D61.3 - Idiopathic aplastic anaemia" had no direct term or conceptual match to any particular MedDRA code. But through the nearest neighbor level, the MedDRA code "10002037 - Anaemia aplastic" was associated with this particular ICD-10 code and given an embedding of a nearest neighbor match for the level of semantic proximity in the mapping file.

Again, the results were manually reviewed to determine goodness of fit in the overall map file. Most of the mappings produced via this level occur again at the PT or LT level of MedDRA. An additional gain of slightly more than 400 IMS Germany codes were mapped using this method while the GPRD Read/OXMIS maps were enhanced with an addition

355 mappings at this level.

2.2 External Sources

2.2.1 ICD-9 CrossMap

While mapping the IMS Germany database to MedDRA, our first step was to investigate the existence of any maps already created from ICD-10 to another terminology such as MedDRA. While we found no freely available crossmaps between these two coding schemes, the existence of crossmaps between ICD-10 and ICD-9[4] were readily available.

The UMLS itself possesses a much higher degree of concept coverage between ICD-9 and MedDRA than it does between ICD-10 and MedDRA. Therefore, the idea was to incorporate the use of the external crossmap files available for ICD-10 to ICD-9 in order to further enhance our mapping process to provide a crossmap between ICD-10 and MedDRA.

The algorithm gives mappings produced at this level the same degree of semantic proximity as that of the conceptual level. Since most of these external sources have been validated for collecting statistics for large national health services, the level of rigor in creating these maps is of similar caliber as that of the concept maps created for the UMLS itself. The algorithm still gives preference to a UMLS concept map above all others, then if a crossmap reference was found, it was given preference over any lower level mapping produced in the multi-phase approach.

This process did prove extremely useful in gaining an additional 2,684 code mappings between ICD-10 and MedDRA that the UMLS itself failed to reveal in the conceptual mapping phase.

The sources we made use of were:

1) New Zealand Health Information Service (8)
2) National Center for Health Statistics (GEMS) map (9)

The cross maps for the most part seemed highly correlated to the MedDRA mappings that were produced. There were however a small number of "Y" codes (e.g. Y51.1, Y54.5, and Y58.9) from the New Zealand file which were not correct (again, this may be attributed to the contextual placement of terms within ICD-10 that do not completely relate to those same terms in the MedDRA coding scheme).

3. Evaluation

The multi-phase approach to mapping terminologies provides an extremely diverse set of target matches between one terminology and another. It is highly improbable to ever produce exact one-to-one mappings (1) between any two biomedical terminologies due to the fact that most are developed independently in order to serve the needs of a particular domain problem such as recording surgical procedures, diagnostics and conditions for an electronic health records system or for medical claims and billing data. Each terminology has a specific context from which it was developed and meant to be applied toward. However, the hope still exists to be able to align these terminologies with one another to the greatest extent possible.

The UMLS goes a long way towards reaching this goal, by providing an overarching conceptual model through which concept synonymy is expressed by means of shared CUIs among terms in each vocabulary. However, it still is not inclusive of every possible biomedical terminology, and still suffers from a lack of resources to maintain and evaluate the changes which frequently occur in biomedical terminologies from one release to the next.

It is also still very much the case that the UMLS does not claim to have any conceptual hierarchy relating one concept to another with any distance measure that can be used with any consistency when moving between one terminology and another. Therefore, by providing a hierarchy for semantic proximity such as the one proposed in this paper, we hope to encourage the development of similar measures which can then be embedded within a system such as the UMLS.

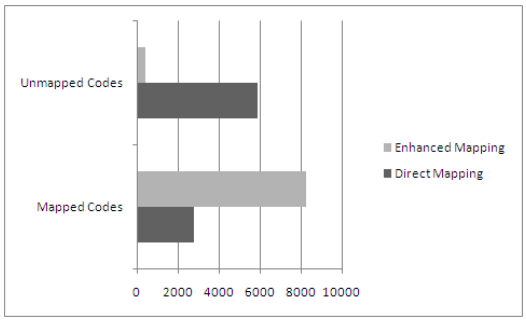

Figure 1: ICD-10 Code Map Coverage

Traditional mapping methods still fall short in producing as many maps are as possible with the multi-phased mapping approach demonstrated here. Term matching alone can produce a high number of potential map targets between any given two terminologies such as the *fuzzy* matching methods described in (2). However, term based matching alone is still subject to a lack of contextual information as noted in the examples stated previously.

The conceptual, boosted and nearest neighbor methods provide additional maps which can still be useful in many data mining activities. The content coverage between IMS German and the MedDRA terminology was drastically improved as shown in Figure 1. The enhanced multi-phase map reduced the total number of unmapped codes by almost 95%.

The number of maps produced between the Read/OXMIS and MedDRA terminologies using the multi-phase approach reduced the total number of unmapped codes from 38,825 to 34,878. While this may not seem all that significant, the

[4]By 'ICD-9' we mean to refer to ICD-9-CM, the International Classification of Diseases, 9th Revision, Clinical Modification, which is maintained jointly by the National Center for Health Statistics and the Health Care Financing Agency.

actual clinical data coverage that was increased by these mappings jumped from slightly less than 50% previously, to more than 60% in this update. The confidence in the mappings produced is also boosted by the fact that we now can provide the semantic proximity as an additional aid in determining the usefulness of these maps to the scientists who will be working with them in the future.

4. Conclusion

Through the use of a multi-phase mapping process, many more potential maps between two terminologies can be realized than by using traditional methods alone. The use of a semantic measure of proximity gives valuable insight into the mappings produced and discretion in how they may be applied to various data mining problems.

It is often the case that the mappings produced between two terminologies must be validated or verified before inclusion in applications such as those used to assess drug safety issues. But the results of this effort show that in many instances, we can reduce the overall volume of concepts which need manual review to those only occurring at lower levels of semantic proximity. Thus, saving countless man hours and valuable resources which would be better suited to the actual investigation of data rather than simply reviewing mapping files between one terminology and another.

Shortly after producing the map files between Read/OXMIS and MedDRA and the IMS Germany codes and MedDRA, we were tasked with creating yet another map between the MeSH (Medical Subject Heading) and Read/OXMIS terminology. The result of the work in this paper allowed us to leverage mappings produced between Read/OXMIS and MedDRA to produce a mapping to MeSH in relative short order achieving similar results and enabling scientists to proceed with mining literature by way of the Read/OXMIS terminology and the translation to MeSH. Again, the embedding of semantic proximity helps to provide valuable clues as to the level of specificity to be deemed necessary when conducting a literature review and being able to cast a broader or narrower net as necessary by means of the filtering now available in the maps this method is capable of generating.

Future work will investigate the possibility of refining our hierarchy for semantic proximity even further and to investigate the applicability of this method to general ontology matching methods.

References

[1] Y. Kalfoglou and M. Schorelmmer, "Ontology mapping: the state of the art", *Knowledge Engineering Review*, vol. 18, no. 1, pp. 1–32, 2003.

[2] Jeffery L. Painter, "Toward automating an inference model on unstructured terminologies: Oxmis case study", in *Advances in Computational Biology*, Hamid R. Arabnia, Ed., vol. 680, pp. 645–651. Springer New York, 2011.

[3] Ying Wang, Jianbin Gong, Zhe Wang, and Chunguang Zhou, "A composite approach for ontology mapping", in *Flexible and Efficient Information Handling*, David Bell and Jun Hong, Eds., vol. 4042 of *Lecture Notes in Computer Science*, pp. 282–285. Springer Berlin / Heidelberg, 2006.

[4] G.H. Merrill, P.B. Ryan, and J.L. Painter, "Construction and annotation of a UMLS/SNOMED-based drug ontology for observational pharmacovigilance.", in *Proceedings of the Intelligent Data Analysis for bioMedicine and Pharmacology*, Washington, DC, 2008.

[5] Stephanie J Reisinger, Patrick B Ryan, Donald J O'Hara, Gregory E Powell, Jeffery L Painter, Edward N Pattishall, and Jonathan A Morris, "Development and evaluation of a common data model enabling active drug safety surveillance using disparate healthcare databases", *Journal of the American Medical Informatics Association*, vol. 17, no. 6, pp. 671–674, November 2010.

[6] G.H. Merrill, "The MedDRA paradox", in *AMIA Annual Symposium Proc*, Washington, DC, 2008, pp. 470–474.

[7] M. A. Merzbacher, "Discovering semantic proximity for web pages", in *Proceedings of the 11th International Symposium on Foundations of Intelligent Systems*, London, UK, 1999, ISMIS '99, pp. 244–252, Springer-Verlag.

[8] "New zealand health information service", 2011, Available online: http://www.nzhis.govt.nz/moh.nsf/pagesns/254.

[9] "National center for health statistics (gems)", 2011, Available online: ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Publications/ICD10CM/2010/2010_DiagnosisGEMs.zip.