Towards Automating an Inference Model on Unstructured Terminologies: OXMIS Case Study

Jeffery L. Painter

GlaxoSmithKline, Research Triangle Park, NC, USA, 27709

jeffery.l.painter@gsk.com

Abstract

Most modern biomedical vocabularies employ some hierarchical representation that provides a "broader/narrower" meaning relationship among the "codes" or "concepts" found within them. Often, however, we may find within the clinical setting the creation and curation of unstructured custom vocabularies used in the everyday practice of classifying and categorizing clinical data and findings.

A significant and widely used example of this lies in the General Practice Research Database which makes use of the Oxford Medical Information Systems (OXMIS) coding scheme to represent drugs and medical conditions. This scheme is intrinsically unstructured, is generally regarded as disorganized, and is not amenable to comparison with other hierarchically structured medical coding schemes. In order to improve processes of data analysis and extraction, we define a semantically meaningful representation of the OXMIS codes by way of the UMLS Metathesaurus. A structure-imposing ontology mapping is created, and this process provides a complete illustration of a general semantic mapping technique applicable to unstructured biomedical terminologies.

1. Introduction

The construction of any ontology is in itself a grand challenge. Research relating to automating and proceduralizing this task continues to play a large role in the areas of ontology development, schema mapping, and the alignment of medical term systems. Our approach allows mapping the Oxford Medical Information Systems (OXMIS) [1] codes (used in the General Purpose Research Database[1], or GPRD) to multiple coding schemes and is facilitated in large part by the UMLS

[1] General Practice Research Database (GPRD) is maintained by the (UK) National Health Service Information Authority

Metathesaurus[2]. The result enables the OXMIS codes to be viewed from the same navigational structure defined in existing and familiar coding schemes. This is of substantial benefit because the GPRD is heavily used in epidemiological and health outcome studies.

The GPRD originally employed OXMIS for coding data, but was later augmented by use of Read codes. Identifying the correct set of codes representing a single medical concept in the GPRD is particularly problematic because of the mixture of the coding schemes that has evolved. As a result, some studies choose to ignore portions of the GPRD data in order to take a uniform approach toward analyzing patient records:

> "OXMIS codes used in the earlier years of the database are not hierarchically organized and do not map readily to equivalent Read codes. We therefore omitted practices which used OXMIS codes by selecting the 123 practices whose records included at least 100% Read codes in each year from 1987 to 2000." [2]

If the researchers working on this study had access to a knowledge representation of the Read-OXMIS codes[3] which could account for the partitioning of the GPRD data, they may have captured a more accurate account of patients' longitudinal records, regardless of which scheme the records were coded in. Unfortunately, with the mixed coding one finds in the GPRD, this is *not* currently possible without much manual work and the identification of corresponding codes between the Read and OXMIS coding schemes.

3. Methods

Our goal is to create a meaningful hierarchical structure of the Read-OXMIS codes as the basis for an informed retrieval model. The techniques employed aim for a high degree of meaning association among the codes. An additional goal is to enable the mapping of the Read-OXMIS codes to other coding schemes found within the UMLS. By making use of the UMLS Metathesaurus, we are able to create this structure and associate a concept hierarchy with the Read-OXMIS codes which will facilitate future mappings of Read-OXMIS to additional coding schemes.

The immediate problem with which we were presented was to decide on which UMLS source should serve as the "target" to which the Read-OXMIS code set

---

[2] 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/

[3] From now on, we will refer to the collective set of codes found in the GPRD as Read-OXMIS. The designation refers to the combination (OXMIS and Read version 2) of coding schemes found in this particular database's medical records.

would be mapped. For this case study, we restricted our attention to only those sources most often referenced by our epidemiologists (i.e. ICD-9, ICD-10, CPT, MedDRA, and CTV 3)[4].

Rahm and Bernstein [3] help illuminate the potential mechanisms by which one might automate the mapping of one schema into another. We approach the process of ontology mapping similarly by first attempting schema integration in their sense. The integration, even when schemata model similar domains (as in our case), first involves a matching process [3]. For our mapping process, we are interested in moving from one coding scheme (the "base") to another (the "target"). Our methodology aids in (1) identifying the appropriate target, and (2) generating an abstraction of the base which allows for imposing the hierarchical structure of the target.

One difficulty in mapping OXMIS codes to any other coding scheme is that there appears to be no comprehensive source of the OXMIS code set on electronic media. We extracted our Read-OXMIS code set using a dictionary listing of all the codes which appear directly in the GPRD data. The majority of these codes are linked to a verbatim string (the term) which assigns some meaning representation to the code itself. However, a few of the codes have no associated string, and this constitutes a problem in mapping them to another coding scheme.

We define two methods in our matching process. The first is a direct method employing exact string matching, while the second takes an associative (or indirect) approach to mapping the Read-OXMIS codes. The associative mapping is a process using a mathematical (probabilistic based) calculation to associate code/term pairs with one or more candidate referents in the target coding scheme.

Both methods convert Read-OXMIS code/term pairs into a concept node related to a concept found in the selected target scheme. We call this process "reification of the concept" represented by the code. The reification of the concept bears a certain similarity to the idea of semantic ascent as addressed by Willard [4]. The process thereby allows us to abstract the code/term pair associations to one or more concept unique identifiers (CUIs) in the UMLS Metathesaurus.

The concept nodes are then used in formalizing the structure-imposing mapping of the Read-OXMIS codes. However, not all codes in the base coding scheme can be

[4] SNOMED CT is copyrighted by the International Health Terminology Standards Organization (IHTSDO). ICD-9 refers to ICD-9, CM the International Classification of Diseases, 9th Revision, Clinical Modification. ICD-10 is copyrighted by the World Health Organization and developed by the National Center for Health Statistics. Current Procedural Terminology (CPT) is copyright the American Medical Association. The Clinical Terms Version 3 (Read Codes) are maintained by the (UK) National Health Service Information Authority.

successfully mapped by using our current methods and we reserve an unclassified "dummy" node category for these entries.

Note that we no longer look at the GPRD Read and OXMIS codes as separate coding schemes (as most users of GPRD previously have), but rather as a single Read-OXMIS vocabulary. This approach allows for the creation of a simpler – yet powerful – model, and ultimately aims to create a single view of GPRD which will improve data extraction and analysis.

3.1 Direct Mapping and Target Selection

In general, if two codes originating from two separate coding schemes are associated with the same term (modulo case differences), then it seems logical to assume that in fact those two codes are representations of the same concept – which is in keeping with the method of lexical alignment as demonstrated by Zhang, Mork and Bodenreider [6].

We refer to this method as "direct mapping", and by using it we found that we could easily determine which potential coding scheme provided the greatest level of coverage. Clinical Terms Version 3 (CTV3) – also known as Read version 3 – was chosen as the target model in order to provide the basis for our hierarchy-imposing representation with direct coverage near 68%.

3.2 Associative Mapping

The direct mapping approach is not sufficient for a complete integration of Read-OXMIS into the concept framework, and we therefore enhance it with a less direct approach of associative mapping. These additional maps allow for lexical variations between source and target terms increasing the likelihood of concept identification within the UMLS Metathesaurus.

The associative mapping procedure attempts to identify candidate strings in the target model that have a probability of semantic similarity to the code/term pairs found within the Read-OXMIS coding scheme. It first preprocesses all of the verbatim strings from the Metathesaurus by using our customized string normalization process similar to the approach described by Bodenreider [7].

1. Remove case differences, parenthetical plurals and contractions
2. Apply a standard stemming algorithm
3. Remove stop words (customized for our domain)

As noted in Mork and Bernstein [8], similarity of the normalized string form is appropriate for lexical matching of this sort. However, we deviate from their metric of similarity (which they confess was based on personal choice) in favor of bi-gram comparison. We chose the use of bi-grams since it provides a higher granularity of lexical comparison. A simple Bayesian calculation determines a proba-

bility of similarity between any two strings. By adjusting an arbitrary limit (which we define as the minimum match probability) these calculations must meet or exceed, we are able to balance between precision and specificity through re-iteration of our process. Although this method is computationally intense, Jensen and Martinez [9] outline clear advantages to it over more simplistic matching techniques.

After normalizing both the Read-OXMIS and target terms, we employ a bi-gram matching algorithm in order to generate candidate term matches in the target coding scheme. Typically, bi-gram matching yields not one, but multiple candidate target strings for any particular Read-OXMIS code entry. The target terms meeting the minimum match probability are then collected into a match list. The resulting match list is similar to the match matrix described by El-Nasan, et al. [10] used for word discrimination. Only the highest ranking match list item is selected for annotation by the Read-OXMIS code/term pair.

The idea is that given a certain level of probability in semantic similarity, lexically distinct terms should fall within the same or similar concept categories. The minimum match probability was set to 0.75 (based on observation) for this Read-OXMIS case. Mapping other coding schemes may require some adjustment to the minimum match threshold. Thus, associative mapping is characterized as a reproducible (and tunable) process that compares normalized versions of the base and target terms using bi-gram matching for the metric of similarity; and then reifying the base code to one or more concepts related to the highest ranking target term.

Adding the associative mapping procedure allowed us to map the Read-OXMIS coding scheme to the target (CTV3) vocabulary achieving 93% code coverage of the original Read-OXMIS codes (leaving only 8,665 Read-OXMIS codes with "dummy" nodes to be placed in our unclassified category).

4. Imposing Hierarchical Structure

The mapping accomplished in the previous steps gave us a method to annotate the existing CTV3 hierarchy. By annotation, we mean the association of a base Read-Oxmis code/term pair with a node in the target CTV3 hierarchy where they share a common concept unique identifier (CUI) found in the Metathesaurus. A Read-OXMIS code can therefore annotate one or more CTV3 nodes via the mappings described above.

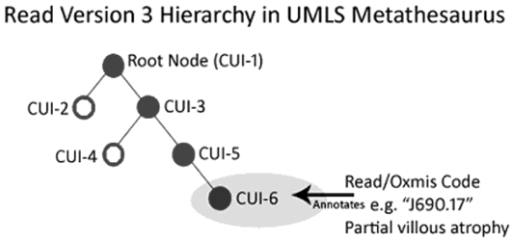

Figure 1: Annotated CTV3 Hierarchy

An illustration of the annotation of the target (CTV3) hierarchy appears in Figure 1. This simplified example demonstrates several of the issues we faced and the decisions we made to support imposing the structure on Read-OXMIS. In this example, we assume there is one node (CUI-6) in the hierarchy which contained a CTV3 term annotated by a Read-OXMIS code. The code "J690.17 – Partial villous atrophy" is then mapped into the hierarchical representation at this node (or category). If a Read-OXMIS code does not annotate any concept in the target representation, then it is placed under a "dummy" node for unclassified codes.

The principle we follow in imposing a foreign structure onto an otherwise unstructured coding scheme is that we keep only those nodes which have annotations in their downward ancestral chain. Therefore, the complete ancestral chain leading from the root to CUI-6 is preserved (denoted by solid circles). Since the nodes labeled CUI-2 and CUI-4 (open circles) fail to meet this criteria, they are subject to purging, thereby increasing the efficiency of the representation. We are then left with a single view of the Read-OXMIS codes as shown in Figure 2.

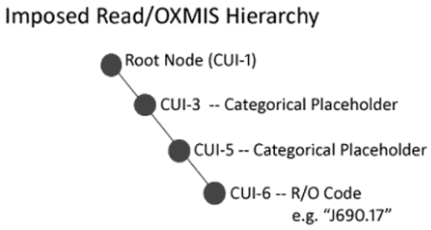

Figure 2: The Read-Oxmis Hierarchy after Purging Un-annotated Nodes

The resulting structure-imposing translation now exhibits more information relevant to the coding scheme than would be possible with a rudimentary (non-structure-imposing) translation. Just as the example above demonstrates, the actual placement of the code "J690.17" is located two levels deep from the root node, under the categorical placeholders of "Clinical findings" and "Morphology findings".

By retaining nodes from the more comprehensive target scheme as categorical placeholders, we associate contextually relevant information with the Read-

OXMIS codes themselves. This is sufficient reason for preserving the complete ancestral path, since (for example) it allows a broader set of searches to succeed. In the resulting hierarchy enriched with the additional categories from CTV3, a search for "morphology" will succeed. The user may then discover the code "J690.17" in this context, and such scenarios illustrate the exploratory paradigm we are seeking to support.

5. Conclusion

Our process creates a view of Read-OXMIS through Clinical Terms Version 3 colored glasses. In order to bring structure to the Read-OXMIS codes, we borrow from the Metathesaurus-based CTV3 hierarchy to provide a template for the placement of the Read-OXMIS codes within a broad and medically meaningful context.

The additional "knowledge" this model provides by way of the semantic structures leveraged from the UMLS concept model, is now imposed on our previously deprived list of code/term pairs providing a richer environment for data retrieval and analysis.

References:

1. Perry, J ed: OXMIS Problem Codes for Primary Medical Care. Oxford Headington. 1978.
2. Jones R, Latinovic R, Charlton J, Gulliford M (2006) Physical and psychological comorbidity in irritable bowel syndrome: a matched cohort study using the General Practice Research Database. Alimentary Pharmacology & Therapeutics 24 (5), 879-886.
3. Rahm E and Bernstein PA, A Survey of Approaches to Automatic Schema Matching. The VDLB Journal, vol. 10, pp. 334-350, 2001.
4. Willard, Dallas. Why Semantic Ascent Fails. Metaphilosophy, Vol. 14. Nos. 3 & 4. July/October 1983, pp. 276-290.
5. Zhang S, Mork P., and Bodenreider O. (2004); Lessons learned from aligning two representations of anatomy; In Proceedings of the First International Workshop on Formal Biomedical Knowledge Representation (KR-MED 2004), (U. Hahn, S. Schulz and R. Cornet, eds). pp. 102-108.
6. Soriano, Maier, Visick & Pride. Validation of General Practitioner-Diagnosed COPD in the UK General Practiec Research Database. European Journal of Epidemiology. Vol 17, No. 12 (2001), pp. 1075-1080.
7. Bodenreider O. Using UMLS semantics for classification purposes. Proc AMIA Symp. 2000: 86-90.
8. Mork, P and Bernstein PA. Adapting a Generic Match Algorithm to Align Ontologies of Human Anatomy. In: 20th International Conference on Data Engineering; 2004 March 30-April 2; Boston, MA: IEEE; 2004
9. Jensen LS, Martinez T. Improving text classification by using conceptual and contextual features. KDD-2000 Workshop on Text Mining, Boston, 2000, pp. 101-102.
10. El-Nasan A, Veeramachaneni S, Nagy G, Word Discrimination Based on Bigram Co-Occurrences. icdar, p. 0149, Sixth International Conference on Document Analysis and Recognition (ICDAR'01), 2001.