

Automatic Mapping ICPC2-PLUS Terms into SNOMED CT Terminologies

Jon Patrick¹, Yefeng Wang¹, Graeme Miller², Julie O'Hallaran²

¹*School of Information Technology, University of Sydney, Sydney, Australia*

²*Family Medicine Research Centre, University of Sydney, Sydney, Australia*

ABSTRACT

In this study, we have mapped the International Classification of Primary Care 2 Australian Version (ICPC2PLUS) terms into the SNOMED CT (SCT) terminology. We have developed a series of computerised mapping algorithms. The UMLS metathesaurus mapping which utilizes the links of ICPC2P and SCT terms in UMLS library has mapped 46.5% of ICPC2P terms to SCT. The lexical mapping explores the lexical similarities between terms in these two terminologies, and has mapped 60.3% of ICPC2P terms overall. Post-coordination of those unmapped terms has been performed, allowing one ICPC2P term to be mapped into composition with two SCT terms, which gives an increase of about 20% in mapped terms. Overall we had mapped 80% of ICPC2P terms to SCT. A manual review of the mapping shows that about 90% of the string-based mappings are accurate, the unmapped terms and mismatched terms are due to the different hierarchy between ICPC2P and SCT. Also terms contained in ICPC2P but not contained in SCT caused a large number of failures in the mappings.

Keywords:

Medical terminology, Terminology mapping, ICPC2 Plus, SNOMED CT

INTRODUCTION

Effective information retrieval within across a hospital's information systems is limited by the lack of semantic interoperability between terminologies used by different departments. The use of multiple terminologies and ad hoc modifications to standard schemes prevents users from cross searching multiple repositories, cross-sectoral resources and inter-disciplinary material. In order to overcome this, improved compatibility between terms is required [1]. The process of "terminology mapping" refers to an identification of identical concepts or relations between different terminologies. Terminology mapping is an important step to achieving knowledge sharing. There has been a large amount of effort spent on computerised terminology mapping. However the nature of this task makes it is very difficult to automate because heterogeneous terminologies may reflect fundamentally subtly different conceptualizations of domains by the creators of these terminologies [2]. In this research we aim to develop an algorithm to map ICPC2P terms into SCT concepts. This mapping process is semi-automatic, because it requires humans to verify the results at the end of mapping, but it transforms the time consuming searching and mapping task into an easier selection and validation task.

BACKGROUND

ICPC-2 PLUS [3] is an interface terminology developed and maintained by The Family Medicine Research Centre (FMRC) of the University of Sydney. It provides a useable coding system for symptoms, diagnoses (problem labels), past health problems and the processes of care for use in age-sex disease registers, morbidity registers and full electronic health records in primary care. It

currently contains over 7,000 terms that are commonly used in Australian general practice. SNOMED CT (SCT) [4] is the most comprehensive biomedical terminology in the world. It is created owned and maintained by the College of American Pathologists, and widely used in the US and UK. It contains more than 360,000 concepts and those concepts are connected by rich relationship networks. It has been used for a long time in coding clinical records and the Australian government is proposing to adopt the SCT terminological system for describing all patient encounters. This decision creates a need to map ICPC2P to SCT and so to complete the task in a reasonable amount of time some computational based methods of matching concepts is needed to assist humans to do the job.

The main approaches to terminology mapping include lexical matching, concept matching, and structural matching. A number of linguists have attempted to make use of linguistic information such as lexical similarity and semantic similarities [5]. Other approaches have been developed recently using structural information to mapping between terminologies. Mork and Bernstein [6] modified a genetic terminology mapping algorithm for mapping human anatomy, using lexical similarities and structure similarities. Fung and Bodenreider [7] derived an algorithm to find candidate mappings between any two terminologies inside UMLS making use of synonymy, explicit mapping relations and hierarchical relationships. Post-coordination can be used to map terms to compositions of two or more concepts. Elkin and Brown [8] developed a technique for discovering and formalizing the implicit semantic relationships between SNOMED Reference-Terminology (SNOMED-RT) and International Classification of Disease Version 9 Clinical Modification (ICD9-CM)

METHODS

Mapping use UMLS Metathesaurus

The most direct mapping is to utilize the link provided by the UMLS (Unified Medical Language System) Metathesaurus which is organized by concepts, and one of its primary purposes is to connect different names for the same concept from many different vocabularies. Different terms in different vocabularies were implicitly connected by a unique concept identifier. The ideal of our approach is to find the terms in these two terminologies that share a common unique concept identifier (CUI) in UMLS. Every term in UMLS has been represented in a “concept structure”. The “concept structure” contains concept identifiers, concept names, their language, and vocabulary source. This information is organised in the Concept Names and Sources File (MRCONSO.RRF). We make use of the common CUI in this file to map terms.

String-Based Mapping

The intuition behind the string-based matching is that because most terminologies have lexical similarity in their vocabularies describing the same concepts when the natural languages underlying the vocabularies are the same. This linguistic connection exists naturally since all terms are developed by humans and are required to be understood by humans. Four string based mapping techniques are used.

Normalized Term Matching: Before comparing the string, the terms from both terminologies are normalized. First, the words within parenthesis are removed. Then the terms are broken into atomic forms and converted into lowercase. Stop words such as “a”, “the”, “of”, “NOS” etc. and punctuation are dropped from multi-word expressions. A morphological process is performed on the remaining terms to remove the inflections. Then some common lexical variations of the terms are generated. Finally the remaining words are sorted in alphabet order. Using this method, the term “*Disease of liver (disorder)*”, for example, is normalized to “*disease liver*” and can be mapped to “*Disease;liver*”.

Expanded Term Matching: There are two kinds of abbreviation found in ICPC2P terms. One is the acronyms such as IUCD which stands for “Intra-Uterine Contraceptive Device” and another is the abbreviation due to space limit in ICPC2P terms, e.g. “musculo” for “musculoskeletal”. These abbreviations are expanded and then exact string matching is performed

Substring Term Matching: To increase the matching coverage, substring matching is also performed. We match the pair of the terms if the normalized and expanded ICPC2 term is a substring of the SCT term. For example, the term *chronic pain* is a substring of *chronic back pain*.

Synonym Matching: Synonym matching uses a thesaurus to explore the semantic meaning of the word constituents. The WordNet synset [10] or the UMLS SPECIALIST Lexicon were used to provide the semantic information of the term. This allows, for example, to map “heart disease” into “cardiac disease” because “heart” and “cardiac” are synonyms. WordNet also provides the derivationally related terms for a given word. For example, the word “fever” is linked to its related adjective “feverish” and “feverous”. We search using the derivational words for the terms as well.

Post-Coordination Mapping

The post-coordination mapping process aims to map a pre-coordinated ICPC2P term to compositions of two or more SCT concepts. This algorithm consists of three steps. Firstly, we break the ICPC2 Plus term into atomic terms. This step includes term normalization, term expansion and break the text into separate words. Then we map each atomic term to the SCT atomic concepts. The atomic term mapping is based on a longest string match. Finally, we find the relationship between the SNOMED CT concepts by matching the relationship patterns we discovered in SCT. We aim to map two kinds of post-coordinations in SCT, the *Qualification* and *Combination*. The following table shows some examples of post-coordination.

Source Term	Post-coordination
Pain;mouth	Pain (Clinical Finding) + attribute = “Finding Site” + Entire mouth region (Body Structure)
Dislocation;knee;simple	Traumatic dislocation of knee joint (Clinical Finding) + attribute = “Onset” + Simple (qualifier value)

Table 1. Example of post-coordination mapping

RESULTS

This section reports the mapping results of each method. A large set of the UMLS mapping results and string-based mapping results had been evaluated by human experts, and only the best candidate mapping is considered as the correct mapping.

UMLS Mapping Result

A total of 3448 (46.53%) ICPC2P terms have been mapped to SCT terms using UMLS mapping. The mapping algorithm found 6557 mapping candidates. 3326 (50.72%) mapping candidates were best-fit mappings, and 96.49% of the mapping candidates have at least one best-fit mapping.

String-based Mapping Result

The string-based mapping results are tabulated in table 2.

Matching Method	Matched Term	Candidates	%age	Newly Mapped Term
Normalized String Matching	3266	3770	44.08%	-
Expanded String Matching	3570	4731	48.18%	304
Synonym Matching	3662	5321	49.42%	92
Substring Matching	4471	108953	60.34%	809

Table 2. String-based mapping results

A total of 3266 ICPC2P terms were mapped to SCT using normalized string matching. This matching method generated a total 3565 mapping candidates, on average, 1.2 matches per matched terms. The Expanded String Matching further mapped 304 terms, however, the average mappings per term increased to 1.33. Synonym Matching is not very effective and only gave a 1% increase in mapping coverage. Most of the substring matches were one to many, and the average number of matches per term increased to 24.88. Overall, the string-based term mapped 60.34% of ICPC2P terms to SCT.

The normalized string matching results were evaluated by an expert. 3031 (92.8%) terms have at least one correct mapping candidate. Among the 3770 mapping candidates, 3565 (94.25%) were correct mappings.

Post-coordination Mapping Result

We excluded the terms that had been mapped in the pervious mapping and performed post-coordination mapping. The remaining set consisted of 3840 terms.

Post-coordination Type	Number of Mapping	Percentage
Qualification	343	4.63%
Combination	902	12.17%
Undetermined	255	3.44%
Total	1500	20.24%

Table 3. Post-coordination results

Qualifications are the post-coordinations that have at least one qualifier value concept. Combinations are these post-coordinations except qualifications and the relationship between the concepts were identified, while undetermined post-coordinations are those with atomic concepts that had been matched to SCT concepts, but the relationship between the concepts were not determined. Overall there were 20.24% terms mapped using post-coordination.

DISCUSSION

As the number of medical terminologies increases, it increases the need for terminology integration. As a result, the demand for rapid and effective computer-assisted terminology mapping has arisen. Computerised mapping systems could reduce significant human effort, especially for mapping large terminologies, such as SCT. The proposed algorithms have found candidate mappings for 80% overall of ICPC2P terms.

The UMLS mapping requires the latest UMLS Metathesaurus version to get the best performance, since the content of UMLS is refined and updated constantly. Current experiments were conducted on the 2005AB version which contains the ICPC2P 2000 vocabulary and SCT 2002 vocabulary. The

ICPC2P 2000 only covers 87% of the newest ICPC2P terminology. We expect a larger number of mappings will be discovered when we use the latest version of UMLS.

On evaluation, the normalized string matching and expanded string matching were accurate and useful. The substring matching had broader coverage, but it results in a huge number of mapping candidates. Upon normal inspection, a lot of mappings were imprecise. Nevertheless, roughly 10% of the mappings were still accurate. One possibility for reducing the superfluous mapping candidates in string-based mappings was to use structural information in both terminologies to eliminate the irrelevant mappings. Although the hierarchy of ICPC2P and SCT were organised differently, it was still moderately effective.

The results of post-coordination mapping haven't been evaluated yet. Nevertheless, the system has demonstrated its ability for automated term composition using a combination of standard string-based mapping techniques. One important phenomenon in post-coordination is the identification of relationships between the mapped terms. This may require description logic generation and more detailed semantic analysis to make sure the composition of two concepts make sense. We believe that the post-coordination mapping is a way to solve the content completeness problem among different terminologies.

CONCLUSION

In conclusion, we have mapped about 80% of ICPC 2 Plus terms to SNOMED CT concepts with different levels of accuracy via three automated mapping approaches. This research demonstrated that the automated mapping has been able to perform different levels of terminology mapping. The results have shown that some of the mapping methods produce very reliable mapping, while some methods yield boarder coverage, but less convincing selections. The mapping results provide an opportunity to analyse the difference in these two different terminologies. Further refinement of the mapping methods could be done to reduce superfluous and incorrect mapping using structure and categorical information, for example, the elimination of synonym ambiguity. More sophisticated post-coordination mapping can be developed in order to provide more reliable mapping.

REFERENCE

- [1] M. Tsiknakis, C. E. Chronaki, S. Kapidakis, C. Nikolaou and S. C. Orphanoudakis, *An Integrated Architecture for the Provision of Health Telematic Services based on Digital Library Technologies*. Int J Dig Libr 1997;1(3):257-277.
- [2] A. Rector, *Clinical terminology: why is it so hard?* Methods Inf. Md. 38 (1999) pp 239-252.
- [3] ICPC2P International Classification of Primary Care 2 Australia Version <http://www.fmrc.org.au/icpc2plus/>.
- [4] K. Spackman, K. Campbell. Compositional concept representation using SNOMED: towards further coverage of clinical terminologies. Proc AMIA Symp 1998:740-4.
- [5] N.Noy, M.Musen. *Prompt: algorithm and tool for automated ontology merging and alignment*, in: Proc National Conference on Artificial Intelligence, 2000.
- [6]P.Mork, Bernstein PA. *Adapting a Generic Match Algorithm to Align Ontologies of Human Anatomy*. In: 20th International Conference on Data Engineering; 2004 .
- [7] K.Fung, O. Bodenreider Utilizing the UMLS for Semantic Mapping between Terminologies AMIA 2005.
- [8] P.Elkin, Brown SH Automated enhancement of description logic-defined terminologies to facilitate mapping to ICD9-CM J Biomed Inform. 2002 Oct-Dec;35(5-6):281-8.
- [9] National Library of Medicine, *UMLS Unified Medical Language System*, Website: <http://umlsks4.nlm.nih.gov>
- [10] Fellbaum, C., WordNet: An Electronic Lexical Database, MIT Press,1998.