7th Annual Conference and Exhibition

Developing SNOMED CT Subsets from Clinical Notes for Intensive Care Service

Jon Patrick, Yefeng Wang, Peter Budd
School of Information Technologies, University of Sydney, Sydney, Australia
{jonpat, ywang1, pbudd}@it.usyd.edu.au

Alan Rector, Sebastian Brandt, Jeremy Rogers
Department of Computer Science, University of Manchester, Manchester, UK
arector@cs.man.ac.uk, brandt@cs.manchester.ac.uk, jeremy.rogers@nhs.net

Robert Herkes, Angela Ryan
Intensive Care Service, Royal Prince Alfred Hospital, Sydney, NSW, Australia
roberth@mail.usyd.edu.au, angela@cs.usyd.edu.au

Bahram Vazirnezhad
Department of Biomedical Engineering, Amirkabir University of Technology, Iran
bahram@it.usyd.edu.au

Abstract

This paper describes the development of a SNOMED CT subset derived from clinical notes. A corpus of 44 million words of patient progress notes was drawn from the clinical information system of the Intensive Care Service (ICS) at the Royal Prince Alfred Hospital, Sydney, Australia. This corpus was processed by a variety of natural language processing procedures including the computation of all SNOMED CT candidate codes. There are about 13 million concept instances comprising about 30,000 unique concept types detected in the corpus. These instances have been processed by a tool which computes the closure of the minimal sub-tree of concept types in the SNOMED hierarchy thus inferring the complete subset of SNOMED CT that would be necessary for an intensive care unit. A subset of about 2700 concepts gives a coverage of 96% of the corpus and the transitive closure uses less than 1% of SNOMED concepts and relationships. Use of this subset will enable clinical information systems to efficiently deliver SNOMED CT terminology to the presentation interface.

1 Objectives

This study uses the contents of the clinical notes collected from an ICU's clinical information system, CareVue Classic (Philips Medical Systems, Andover, MA) to compute a suitable ICU subset. The notes were originally extracted, anonymised and analysed in a variety of ways for two objectives: to identify linguistic characteristics relevant to successful automatic processing of the narratives, and to understand the everyday use of written (typed) clinical language and assess where it might be improved. Subsequently it became apparent that the narratives contain the information needed to define the language of the ICU and hence define the concepts that needs to be expressed in an ICU clinical dialect. From this study it appears feasible to use the clinical notes to construct an ICU subset of SNOMED CT (SCT) as an alternative method to collecting a subset in a Delphi process using expert intensivists.

2 Background

The ability to extract meaningful fragments from ontologies is a key issue for ontology usage. In practice, a method is needed to extract a subset that includes just the information relevant to a specific domain of interest. Ideally, this subset should be as small as possible while still guaranteeing to capture the meaning of the terms used in the clinical specialty, that is, when answering arbitrary queries, importing the subset would give us exactly the same answers as if we had imported the complete SNOMED CT.

In the last few years, numerous techniques for extracting fragments of ontologies for knowledge reuse purposes, called modularisation, have been developed. Most of these techniques rely on syntactically traversing the axioms in the ontology and employing various heuristics for determining which axioms are relevant and which are not.

One example of such a procedure is the algorithm implemented in the PROMPT-FACTOR tool (Noy & Musen, 2003). Given a domain vocabulary and ontology, the algorithm retrieves a subset as follows: first, the axioms in the ontology that mention any of the terms and symbols in the vocabulary are added to the subset; second, the vocabulary is expanded with the other atomic concepts or roles of the subset. These steps are repeated until a fixed point is reached. Another example is the algorithm (Seidenberg and A. Rector, 2006) which was used for segmentation of the medical ontology GALEN (Rector and J. Rogers, 1999). Given a vocabulary and ontology, the algorithm adds all definitions of the atomic concepts of the terms and symbols in the vocabulary to the subset. Then this algorithm prunes irrelevant axioms by traversing the class hierarchy “upwards” and across existential restrictions, but does not detect other dependencies. These algorithms, are intended to extract “relevant parts” of the ontology which are “likely to be related” to a given domain. They do not guarantee conformity to a particular semantic subset.

Modularisation of ontologies has been studied in the context of collaborative ontology development and controlled integration (Cuenca Grau, et al 2007). They give a definition of a subset that guarantees that it will capture completely the meaning of a given set of terms, i.e. include all axioms relevant to the meaning of these terms. They also study the problem of extracting conforming minimal modules.

The method introduced in the work of Stuckenschmidt & Klein (2004) consists of automatically partitioning a large ontology into smaller partitions based on the structure of the class hierarchy taking into account the internal coherence of the concepts. The method can be broken down into two tasks; the creation of a weighted graph, and the identification of partitions from the dependency graph. The weights defined in the first step determine the results of the second step.

In the method introduced in Noy & Musen (2004) a user is allowed to state what relations are to be considered in the procedure of making the subset. This allows the user to run the method with different configurations to refine the results to their own requirements.

3 Methods

The Royal Prince Alfred Hospital (RPAH) Intensive Care Service (ICS) is a Level 6 tertiary referral service with 45 operational critical care beds exceeding 3400 admissions per annum. The service collocates the General Intensive Care (GICU), Neurosciences Intensive Care (NSICU), Cardiothoracic Intensive Care (CICU) and High Dependency Units (HDU). The service caters for patients from a wide variety of needs, including transplantation, trauma, neurovascular and cardiothoracic surgery, and multi-organ failure.

The RPAH-ICS has embarked on a partnership with the School of IT, University of Sydney to develop a range of technologies based around the automatic processing of language (Ryan, Patrick, Herkes, 2008). One of the needs of the service is for classification subsets to make the manual and automatic use of the classifications easier to implement in a Clinical Information System (CIS) and more convenient and productive for staff usage.

The RPAH-ICU corpus contains patients' daily medical measurements, conditions, progress notes and care plans recorded during the patient’s ICU stay. The corpus contains 44 million words from medical and nursing notes written from 2002 to 2006 amounting to 461,969 notes for 12,076 patients. Tables 1 and 2 give a profile of the lexical distributions and Table 3 gives the frequencies of SCT categories. The most frequent categories are Clinical Findings 23.5%, Substance 13.7%, Body structure 13.3%, Observable entity 12.3%, & Procedure 12.2%.

Table 1. The size of the total corpus and the token types

No. of word instances including punctuation	44,072,299
No. of unique word types - case sensitive	809,662
No. of unique word types - case insensitive	703,198

Table 2. Frequencies of purely alphabetic tokens and their lexical validation.

Token Type	**No. of token types**	**No. of tokens in corpus**	**%age**
Alphabetic strings	157,866	31,646,421	71.8%
Words in Moby English lexicon	32,081	28,095,490	63.7%
Words in SNOMED	22,421	29,008,594	65.8%
Words in UMLS (excludes SNOMED words)	25,956	27,893,156	63.3%
Words in SNOMED but not in Moby	5,005	1,985,391	4.5%
Words in either SNOMED or Moby	37,086	30,080,881	68.3%
Words in neither SNOMED nor Moby (Unknown words)	120,780	1,565,540	3.6%

Table 2 indicates that of about 160,000 words about thee quarters are not recognised as valid words. This occurs for a number of reasons: the word is misspelt, the word is a valid word but not recorded in our lexical sources (e.g. bibasal/bibasally) or the word is a neologism (e.g. Warfarinise). Some 25% of these words can be corrected with 1 or 2 spelling changes with a unique word. Another 25%, with 1 or 2 spelling changes, produce on average 3 possible valid words. Thus 50% of unrecognised words are not readily recoverable. This has consequences for other types of language processing, for example, parsing sentences will be seriously limited. With an average length of sentence of, say 10 words, about 1 in 4 sentences will have an unrecognizable word and thereby not be parseable. About 75% of the SCT concepts found in the corpus are represented by 5 categories: Clinical finding, substance, body structure, observable entity and procedure (Table 3).

A prototype strategy has been developed to assess the approach of inductively inferring the SNOMED CT subset for ICUs from the texts of the patient notes. The steps are:

1. Identify all the SCT candidates in the clinical notes - this is the *extracted set*.
2. Construct a histogram of all the concept codes in the extracted set and separate them into SCT categories.
3. Import the code frequency tables into one table to be a fair sample of concepts.
4. From that extracted set, select the codes that were used at least 100 times (an arbitrary cut off point) - this is the *reduced set*.
5. Using appropriate software, compute the minimal closure across the reduced set - this is the *closure set*.
6. Manually clean the closure set semi-automatically to remove anomalous concepts and to correct defective SNOMED modeling - this is the *clean closure set*.

Table 3. The frequencies of the SNOMED CT Categories.

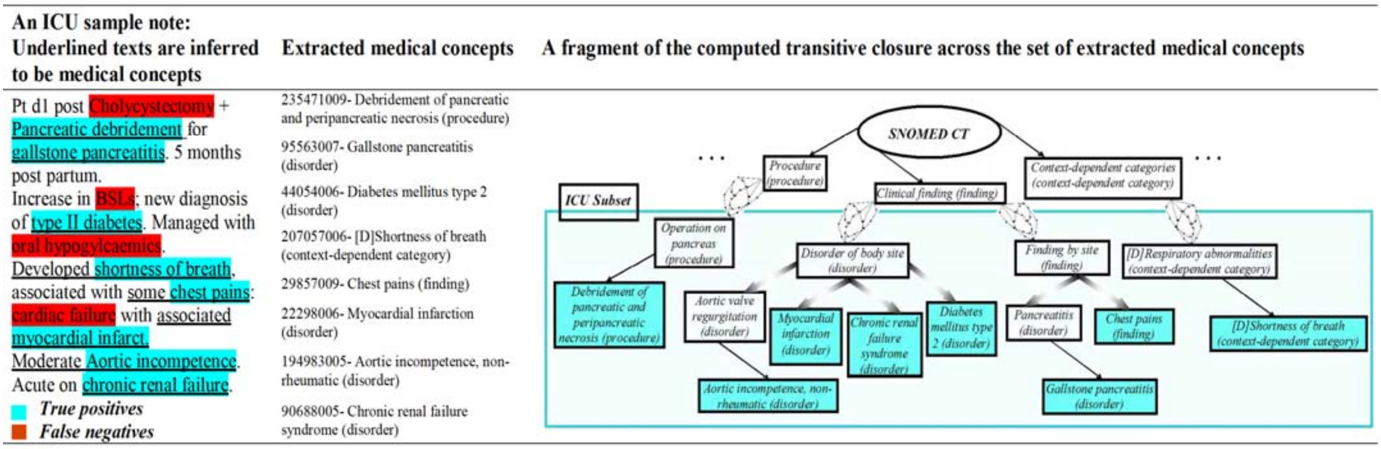
CLASS ID	CLASS NAME	# of Concepts	%age
2	Clinical finding (finding)	3085935	23.5%
19	Substance (substance)	1799003	13.7%
1	Body structure (body structure)	1742270	13.3%
7	Observable entity (observable entity)	1609937	12.3%
12	Procedure (procedure)	1600392	12.2%
11	Physical object (physical objects	968856	7.4%
9	Pharmaceutical / biologic product (product)	853474	6.5%
15	Social context (social concept)	592940	4.5%
18	Staging and scales (staging scale	474280	3.6%
3	Context-dependent categories (context-dependent category)	165741	1.3%
8	Organism (organism)	123683	0.9%
10	Physical force (physical force)	53439	0.4%
17	Specimen (specimen)	24204	0.2%
16	Special concept (special concept)	21525	0.2%
5	Events (event)	20343	0.2%

Figure 1 shows an example of a clinical note transformed through steps 1 to 5. There are 4 false negatives: “Cholycystectomy” has the orthography “Cholecystectomy” in SCT, a spelling error in “hypogylcaemics”, BSL is an unknown abbreviation in SCT, and “cardiac failure” is a phrase that is part of a total of 27 different concepts but is never a concept by itself. The search algorithm (Patrick, Wang, Budd, 2007) for matching text strings to SCT concepts uses a probabilistic matching strategy and in this example there are an insufficient number of words to match against any specific target concepts. The sample shows that notes are rarely fully formed sentences with correct grammatical structure.

The process of computing the closure of the minimal sub-tree involves identifying for every identified concept all the siblings of that concept and the sub-tree of concept nodes below that concept. The transitive closure of this minimal sub-tree is the process of including all the relationships the concept nodes have with other axes of SNOMED CT e.g. attributes such as Episodicity, Severity, etc.

4 Analysis

The *reduced set* comprises 2718 codes, and covers 6,177,077 of the 6,428,597 codable item instances summarised by the histogram reports, or 96.09% of all codable items. The remaining 3.91% of codable items (n=251,520) requires an additional 21,375 SNOMED codes. Restricting the set to only the top 1000 codes by usage would cover 89.25% of the codable items in the RPAH-ICU corpus.



An ICU sample note: Underlined texts are inferred to be medical concepts

Pt d1 post Cholycystectomy + Pancreatic debridement for gallstone pancreatitis. 5 months post partum.
Increase in BSLs; new diagnosis of type II diabetes. Managed with oral hypogylcaemics.
Developed shortness of breath, associated with some chest pains: cardiac failure with associated myocardial infarct.
Moderate Aortic incompetence.
Acute on chronic renal failure.

🟦 *True positives*
🟥 *False negatives*

Extracted medical concepts

235471009- Debridement of pancreatic and peripancreatic necrosis (procedure)
95563007- Gallstone pancreatitis (disorder)
44054006- Diabetes mellitus type 2 (disorder)
207057006- [D]Shortness of breath (context-dependent category)
29857009- Chest pains (finding)
22298006- Myocardial infarction (disorder)
194983005- Aortic incompetence, non-rheumatic (disorder)
90688005- Chronic renal failure syndrome (disorder)

A fragment of the computed transitive closure across the set of extracted medical concepts

SNOMED CT
. . .
Procedure (procedure)
Clinical finding (finding)
Context-dependent categories (context-dependent category)
. . .
ICU Subset
Operation on pancreas (procedure)
Disorder of body site (disorder)
Finding by site (finding)
[D]Respiratory abnormalities (context-dependent category)
Debridement of pancreatic and peripancreatic necrosis (procedure)
Aortic valve regurgitation (disorder)
Myocardial infarction (disorder)
Chronic renal failure syndrome (disorder)
Diabetes mellitus type 2 (disorder)
Pancreatitis (disorder)
Chest pains (finding)
[D]Shortness of breath (context-dependent category)
Aortic incompetence, non-rheumatic (disorder)
Gallstone pancreatitis (disorder)

Figure 1. A sample ICU note (left), Extracted medical concepts (centre) and computing the transitive closure for the ICU subset (right).

The list of 2718 codes was then run against a logical model version of SNOMED circa Dec-06, using a segmenter to extract the minimal SCT fragment containing all 2718 codes, and a plug-in to parse SCT concepts files for Fully Specified Names (FSNs), taken from the 2008 release. A small number of the 2718 codes relate to SCT codes that were added to the international release after the date (Dec06) of the most recent logic version we had on which to do the modularization.

5 Discussion

The coverage of this subset uses 1 percent of SCT. The SCT release (Jul 2007) has 310,311 active concepts and 1,218,983 relationships, so 1,529,294 rows from two database tables. The subset covering 96% of everything codable in the RPAH-ICU corpus plus a skeleton of ancestors to wrap it in contains 7540 classes, 5600 subclass axioms and 1939 equivalent class axioms. This corresponds to 15,079 rows of the 1.5 million, or about 1% of the content.

This result is not a definitive assessment of the ICU subset. It is subject to a number of error sources, namely

- An incomplete lexical retinue because of words that are unrecognized due to our limited lexical sources, such as neologisms, and local abbreviations. An example of this problem is the widespread use of the word “bibasal” in Australian hospitals whilst it is unrecorded in SCT.

- False Positives which include incorrectly deduced concepts from the SCT set due to weaknesses in the natural language processing methods.

- False negatives which are missing when forming the minimal sub-tree and the transitive closure. This may not be a very large problem as many such terms may be identified through the process of building the transitive closure because they exist within sub-trees of the ontology captured due to correctly recognised concepts.

- The full set of SCT categories has been included in this study.

- Issues with the internationalisation of spelling, for example anaemia versus anemia. orthographic errors, grammatical errors, and typing errors due to English as a second language.

6 Conclusions

Although still in a preliminary stage the work has established that it is feasible to construct a useful subset for SCT using the clinical notes. The full project will require:

- a more reliable extraction from the source corpus with improved orthographic corrections for more accurate lexical verification,

- identification of the transitive closure from the complete set of SCT categories,

- removing from the transitive closure inappropriate and poorly modeled concepts,
- implementation of the subset in an ICU clinical information system for testing its efficiency,
- testing the subset against ICU notes from another institution to assess its generalisability.

This resulting subset of SCT indicates that about 1% is needed for use in the ICU. This reduction in the amount of SCT that is used to analyse the clinical narrative will have at least four positive consequences. It will:

- increase the speed of computation and hence make real-time searching across the SCT logical model more practicable,
- decrease the amount of false hits and so make the user interfaces that exploits SNOMED more accurate and efficient saving staff time,
- permit the use of description logic software to process concept generalisations,
- lead the way in the development of a methodology for deriving SNOMED subsets for other clinical specialities using their clinical progress notes.

References

Cuenca Grau, B. Horrocks, I. Kazakov, Y. and Sattler, U. A logical framework for modularity of ontologies. In Proc. IJCAI-2007, pages 298–304, 2007.

Noy N. and Musen M. The PROMPT suite: Interactive tools for ontology mapping and merging. Int. Journal of Human-Computer Studies, 6(59), 2003.

Noy N. F. and Musen M. A.. Specifying ontology views by traversal. In international Semantic Web Conference, pages 713-725, 2004.

Patrick, J, Budd, P. & Wang, Y. &. An automated system for Conversion of Clinical notes into SNOMED CT. Health Knowledge & Data Mining Workshop, Ballarat, Research & Practice in Info Tech, 68, 195-202, 2007.

Rector A. and Rogers J. Ontological issues in using a description logic to represent medical concepts: Experience from GALEN. In Proc. of IMIA WG6 Workshop, 1999.

Ryan, A. Patrick, J & Herkes, R. (2008). Introduction of Enhancement Technologies into the Intensive Care Service, Royal Prince Alfred Hospital, Sydney. Health Information Management Journal Vol 37 No 1 2008 ISSN 1833-3583 (PRINT) ISSN 1833-3575 (ONLINE), pp39-44.

Seidenberg J. and Rector A. Web ontology segmentation: Analysis, classification and use. In Proc. WWW-2006, 2006.

Stuckenschmidt H. and Klein M. Structure-based partitioning of large concept hierarchies. In Proc. of the 3rd International Semantic Web Conference, Hiroshima, Japan, 2004.