Inductive Inference of a SNOMED CT Subset for Intensive Care Services

Jon Patrick[1], Yefeng Wang[1], Bahram Vazirnezhad[1], Alan Rector[2], Sebastian Garde[2], Jeremy Rogers[2] , Robert Herkes[3], Angela Ryan[3].

[1] Univ. of Sydney, [2] Univ. of Manchester, [3] Royal Prince Alfred Hospital

1. Abstract

A corpus of 44 million words of patient progress notes has been drawn from the clinical information system of the Intensive Care Service (ICS) at the Royal Prince Alfred Hospital, Sydney, Australia. The corpus has been processed by a variety of natural language processing procedures including the computation of all SNOMED CT candidate codes. The false positive error rate is estimated to be about 30%. The false negative rate is unknown but is believed to be of the order of 10% based on inspection of some of the texts. There are 13,136,022 concept instances making up 30,000 unique concept types detected in the corpus. These instances have been processed by a tool which computes the transitive closure of the concept types over the SNOMED hierarchy thus inferring the complete subset of SNOMED CT that would necessary for an ICS. A subset of 2718 concepts gives a coverage of 96% of the corpus but only needing to use less than 1% of all of SNOMED. This will give significant advantage to clinical information systems for efficiently delivering SNOMED terminology to the presentation interface.

2. Introduction

The RPAH-ICU corpus contains patients' daily medical measurements, conditions, care taken from the ward round. Notes were written by doctors and nurses. The notes were from year 2002 to 2006 amounting to a total of 461,969 notes for 12,076 patients. Tables 1 and 2 give a profile of the lexical distributions in the corpus and Table 3 gives the profile of SNOMED CT concepts identified in the corpus.

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

No. of token includes punctuation	44,072,299
No. of token type case sensitive	809,662
No. of token type 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 words	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 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 object)	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%

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 hypoglycaemics.
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

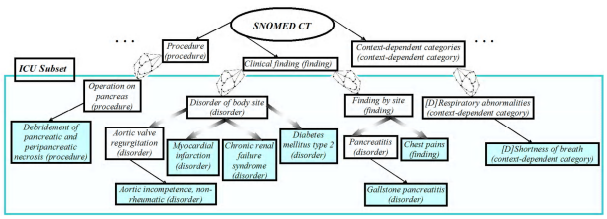

Figure 1. A sample ICU note (left), Extracted medical concepts (centre) and computing the transitive closure for the ICU subset (right)

3. Methods

A prototype strategy has been developed to asses the merit of the approach to the inductive inference of 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 and separate them into SCT categories.
3. Import the code frequency tables of body structure, clinical finding and procedures 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 transitive closure across the set of extracted codes - this is the *closure set*.
6. Clean the closure set semi-automatically to remove anomalous concepts and to correct defective SNOMED modeling - this is the *clean closure set*.

Figure 1 shows an example clinical note transformed through steps 1 to 5. There are 3 false negatives because Cholycystectomy is in SCT Cholecystectomy, BSL is an unknown abbreviation, and cardiac failure is an incomplete expression in SCT used in 27 concepts.

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.
The list of 2718 codes was then run against a KRSS version of SNOMED circa Dec-06, using a segmenter to extract the minimal SCT fragment containing all 2718 codes, and a plugin to parse sct_concepts files for FSNs, taken from the 2008 release.
Note: 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 KRSS version we had available on which to perform the segmentation.

5. Coverage of SNOMED CT

The subset uses 1 percent of SCT. The current SNOMED (Jul 2007 release) has 310,311 active concepts and 1,218,983 relationships, so 1,529,294 rows from two tables.
The subset covering 96% of everything codable in the RPAH_ICU corpus plus a skeleton of ancestors to wrap it in, contains only 7540 classes, 5600 subclass axioms and 1939 equivalent class axioms. This corresponds (more or less) to 15,079 rows from the full 1.5 million, or just less than 1% of the content.

6. Conclusions

It appears feasible to construct a useful subset for SCT using the clinical notes. This work is preliminary. The major project requires:
- a more reliable extraction from the source corpus with better orthographic corrections for better lexical verification.,
- identification of the transitive closure from the complete set of SCT categories,
- cleaning of the transitive closure of inappropriate concepts and poorly modeled concepts,
- implementation of the subset in an ICU clinical information system for testing its efficiency.

This research is sponsored by the Royal Prince Alfred Hospital - Intensive Care Service, Sydney, the Health Information Technology Research Laboratory, U Sydney, Australia <www.it.usd.edu.au/~hitru> and the Bio-Health Informatics Group, U. Manchester, UK <intranet.cs.man.ac.uk/bhig>

The University of Sydney