Informatics for Health: Connected Citizen-Led Wellness and Population Health
R. Randell et al. (Eds.)

doi:10.3233/978-1-61499-753-5-446

Building SNOMED CT Post-Coordinated Expressions from Annotation Groups

Jose Antonio MIÑARRO-GIMÉNEZ[a,1], Catalina MARTÍNEZ-COSTA[a], Pablo LÓPEZ-GARCÍA[a] and Stefan SCHULZ[a]
[a] *Institute for Medical Informatics, Statistics and Documentation.*
Medical University of Graz

Abstract. SNOMED CT supports post-coordination, a technique to combine clinical concepts to ontologically define more complex concepts. This technique follows the validity restrictions defined in the SNOMED CT Concept Model. Pre-coordinated expressions are compositional expressions already in SNOMED CT, whereas post-coordinated expressions extend its content. In this project we aim to evaluate the suitability of existing pre-coordinated expressions to provide the patterns for composing typical clinical information based on a defined list of sets of interrelated SNOMED CT concepts. The method produces a 9.3% precision and a 95.9% recall. As a consequence, further investigations are needed to develop heuristics for the selection of the most meaningful matched patterns to improve the precision.

Keywords. SNOMED CT, Post-coordination

1. Introduction

Encoding free-text clinical information with standard medical terminologies and ontologies is essential to improve many areas of healthcare: from semantic retrieval, to real-time decision support, cross-border data interoperability, and retrospective reporting for research and management [1]. SNOMED CT [2] is the largest medical terminology for coding clinical information, based on an ontological model of meaning. Its representational units (concepts) can be atomic ones such as *Bleeding*, *Stomach,* etc.) or pre-coordinated ones, which define complex expressions such as *Acute gastrointestinal hemorrhage* using logical axioms.

Both, pre- and post-coordinated expressions are based on the SNOMED CT Concept Model, although the meaning of some concepts is not always fully formalized (e.g. "severe" within *Severe pain*).

When annotating a given piece of information, e.g. a clinical text with SNOMED CT concepts by human annotators or NLP systems, there are several valid ways to encode the same meaning, depending on whether the user or systems tends to use pre-coordinated or primitive concepts. For example, the text "acute hemorrhage of the gastrointestinal tract" one annotator could have selected the primitive concepts C_1: *Acute hemorrhage* and C_2*: Gastrointestinal tract structure,* while another one could

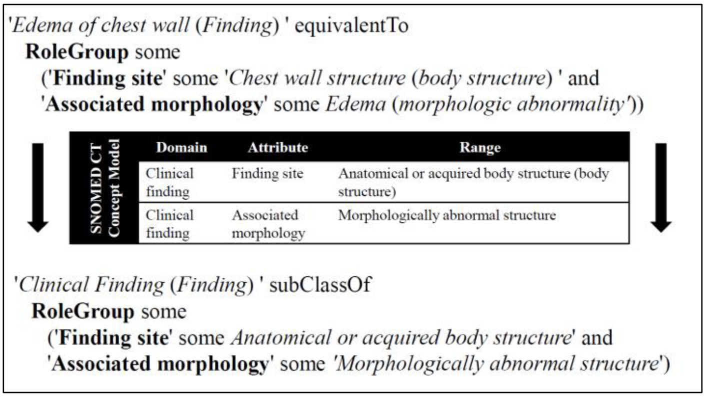
[1] Corresponding author, Jose A. Miñarro-Giménez, Institute for Medical Informatics, Statistics and Documentation, Medical University of Graz, Auenbruggerplatz 2, 8036 Graz, Austria; Email: jose.minarro-gimenez@medunigraz.at

have chosen the pre-coordinated concept C_3: *Acute gastrointestinal hemorrhage.* Although there is a full logical definition of C_3, referring to both C_1 and C_2, it is not possible to infer C_3 by post-coordinating C_1 and C_2, since this would require SNOMED CT relations for connecting body structure and morphologic abnormality concepts with a clinical finding concept, in this case the relations **AssociatedMorphology** and **FindingSite**, respectively. SNOMED CT relations, however, are missing in the annotations. Our work aims at exploring the possibility of inferring meaningful post-coordinated expressions out of set of SNOMED CT concepts. This would require guessing the missing relations and putting them in the right place. The question is whether a relatively shallow text processing is enough for inferring the correct post-coordinated expressions. The SNOMED CT text annotations were produced within the ASSESS CT project. They resulted from annotating a collection of medical texts by domain experts [3]. The experts had to identify cohesive text chunks and assign a set of SNOMED CT concepts to them. The results are based on a previous study that demonstrated that most pre-coordinated expressions within SNOMED CT share a limited number of structural patterns [4]. Here we extend these patterns with the SNOMED CT Concept Model.

2. Methods

2.1. SNOMED CT Pattern Extraction Based on the Concept Model

The list of patterns from [4] is extended taking into account information from the SNOMED CT Concept Model. Figure 1 describes the extraction of one of the patterns based on a pre-coordinated SNOMED CT concept. A pattern is described in terms of the SNOMED CT top-level categories (*Clinical Finding*, *Body Structure*, *Procedure, Qualifier Value*, etc.). The selection of the particular top-level categories is indicated by the SNOMED CT Concept Model.



'*Edema of chest wall* (*Finding*) ' equivalentTo
RoleGroup some
('**Finding site**' some '*Chest wall structure* (*body structure*) ' and
'**Associated morphology**' some *Edema* (*morphologic abnormality'*))

SNOMED CT Concept Model

Domain	Attribute	Range
Clinical finding	Finding site	Anatomical or acquired body structure (body structure)
Clinical finding	Associated morphology	Morphologically abnormal structure

'*Clinical Finding* (*Finding*) ' subClassOf
RoleGroup some
('**Finding site**' some *Anatomical or acquired body structure*' and
'**Associated morphology**' some *'Morphologically abnormal structure*')

Figure 1. Post-coordination pattern extraction method uses a pre-coordinated expression (here the concept definition of *Edema of chest wall*) and the constraints from the SNOMED CT Concept Model to obtain a pattern.

2.2. Generation of Post-Coordinated Expressions

The post-coordination method takes as input the list of patterns extracted from SNOMED CT and automatically produces as output a list of valid post-coordinated expressions (not necessarily meaningful) for a provided list of AGs.

An annotation group (AG) is an unordered set of SNOMED CT concepts that jointly represent or approximate the meaning of a cohesive piece of clinical discourse, supposed to be expressible as a SNOMED CT compositional expression. Our manual annotations yielded 169 AGs. For example, the text "Given the rapid extension of the subgaleal bleeding, coagulopathy workup was initiated" was annotated with the following AG {*Subgaleal area*, *Bleeding* and *Blood coagulation panel*}.

The post-coordination method is divided in four main steps (see Figure 2). The *AG filtering* step avoids processing AGs with less than two concepts because the minimal requirement is one focus concept (i.e. the root concept within a post-coordinated expression) and one modifying concept related to it. The *pattern selection* step goes through the list of patterns to gather those whose focus concept is compatible with the focus concept in the AG. The *pattern matching* step obtains all concept combinations within an AG to match the selected patterns. Here, wherever there are two partial matched patterns with the same content, the lower frequency is discarded due to redundancy. Patterns for which all relations in the expression have a valid target concept from the AG are always retrieved. The *sorting matched patterns* step arranges the list of returned patterns based on their frequency. As a consequence, the matched patterns with higher frequencies are placed first, regardless of the matched relations and target concepts.

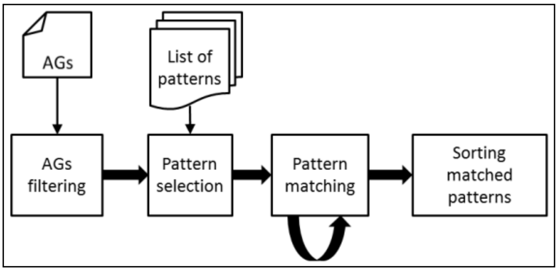


Figure 2. Post-coordination method for building SNOMED CT expression from Annotation Groups (AGs)

2.3. Evaluation Method

In order to assess whether the resulting list of patterns can be used to find meaningful post-coordinated expressions a gold standard was produced. It was manually created by a SNOMED CT expert. The post-coordination expressions of the gold standard were created based on each AG and by looking at the clinical narrative in order to assert the correct relations consistent with the SNOMED CT Concept Model. The evaluation consists in manually comparing the list of resulting matched patterns ordered by frequency with the post-coordinated expression from the gold standard. In cases with overlapped patterns, the one with highest frequency is taken. The following five categories are defined for the evaluation (with the respective counts in brackets): HIT, PARTIAL HIT, NO HIT, EMPTY, NO EXP and NO EXP AND EMPTY.

HIT means that all elements of an AG are matched against a meaningful pattern. For example, the AG {Subgaleal area, Bleeding, Blood coagulation panel} corresponds to the gold standard post-coordinated expressions:

'*Bleeding* (*finding*)' and **RoleGroup** some
('**Finding site**' *some* '*Subgaleal area* (*body structure*)')

plus the concept '*Blood coagulation panel* (*procedure*) ' which is not part of any post-coordinated expression.

NO HIT means that although some matched patterns are retrieved, none of them agrees with the gold standard. For example, the AG {*Screening mammography*, *Interval*, *Month*}, has the gold standard expression:

'*Screening mammography* (*procedure*)' and
'**Time aspect**' some *Interval* (*qualifier value*)

but if the SNOMED CT relation '**Time aspect**' does not appear in any of the patterns, therefore, no post-coordinated expression can be created.

PARTIAL HIT means that the content of the gold standard expressions are partially matched. EMPTY occurs when no patterns are matched but the gold standard provides post-coordination expressions for the AG under scrutiny.

NO EXP means that it is not possible to post-coordinate the AG elements according to the gold standard. For example, the AG {*Bisoprolol* (*substance*), *milligram (qualifier value), Twice a day (qualifier value)*}, cannot be post-coordinated because SNOMED CT Concept Model does not cover it.

Finally, we calculate the precision and recall based on the resulting matched patterns. The precision corresponds to the number of meaningful matched patterns divided by the total number of retrieved matched patterns. The recall is obtained dividing the number of meaningful patterns by the total number of expressions in the gold standard. We assume NO EXP and EMPTY are true negatives cases.

3. Results

Based on the SNOMED CT distribution files from January 2016, we obtained 956 structural patterns out of 357,165 pre-coordinated concepts with frequencies from 27,413 to 1. In particular, 284 patterns had frequency of 1. The most frequent pattern is the one depicted in Fig. 1, defining a clinical finding or disorder in terms of morphology and site.

The post-coordination method processed the list of 169 AGs in order to produce their corresponding list of matched patterns (not including AGs with only one element). It retrieved matched patterns to only 56 AGs. The AG with the most matched patterns contains nine concepts and it was associated with 377 matched patterns.

The evaluation provided the following counts (in brackets) for each category: HIT (39), PARTIAL HIT (0), NO HIT (1), EMPTY (2) and NO EXP (16), NO EXP AND EMPTY (111). As a result, we obtained a precision 9.31% and a recall of 95.9%. Moreover, the average position of a meaningful matched pattern is 7.1 in the returned list of matched pattern sorted by frequent with an average size of 30.3 patterns of list of matched patterns.

4. Discussion and Conclusion

The result from the post-coordination method shows a high rate of meaningful post-coordination expressions based on the list of patterns extracted from SNOMED CT. Consequently this list provides the most frequent patterns used for representing clinical discourse with SNOMED CT post-coordinated expressions. However, precision was very low, mainly, because the first version of the post-coordination method presented here retrieves all possible matching patterns. We have observed that the average position of a meaningful matched pattern in the resulting list of matched patterns sorted by pattern frequency is lower than the middle of the average size of the resulting list of matched patterns. The use of this frequency to limit the list of matching patterns and improve the precision will be explored in subsequent work. Yet, our method misses some meaningful post-coordination expressions, mainly for two reasons:

Firstly, there are no SNOMED CT pre-coordinated concepts that include the required pattern like in:

'*Procedure* (*procedure*)' and **'Time aspect'** some '*Time frame* (*qualifier value*)'

Secondly, the post-coordinated expression has as focus concept from the Situation with explicit context hierarchy, which has not been used within any AG, due to coding rules that guided the manual annotation task in ASSESS CT. Besides, there are several SNOMED CT relations with equivalent domain and range restrictions such as **Procedure site - Direct** and **Procedure site - Indirect**. Here, the selection of the meaningful pattern cannot be decided based on a set of concepts only, it requires analysing the semantic content of the clinical text.

Future work should allow matching patterns without focus concepts, using the top-level as default, e.g. Procedure. The focus node might then be found in previous AGs, because anaphora or ellipsis are frequent phenomena in medical texts. A necessary extension is also the addition of new structural patterns that are frequent in medical texts although they are not used within SNOMED CT definitions. This would however require a sufficiently large training set of clinical text "translated" into compositional SNOMED CT expressions, which requires time and in-depth knowledge of the SNOMED CT Concept Model.

Acknowledgements

The text annotations were performed in the EU project ASSESS CT (www.assess-ct.eu).

References

[1] O. Bodenreider. Biomedical Ontologies in Action: Role in Knowledge Management, Data Integration and Decision Support. *Yearb Med Inform*. (2008), 67–79.
[2] IHTSDO, SNOMED CT. The Global Language of Healthcare. Available: http://www.ihtsdo.org/snomed-ct [Accessed: 31-Oct-2016].
[3] J.A. Miñarro-Giménez, C. Martínez-Costa, S. Schulz. Qualitative assessment of annotations using SNOMED CT. *Proc. of the 7th Workshop on Ontologies and Data in Life Sciences* **1692** (2016), K.1–2.
[4] P. López-García, S. Schulz. Structural Patterns under X-Rays: Is SNOMED CT Growing Straight? PLOS One 2016 Nov 3;**11**(11):e0165619