Combining lexical and structure-based methods to align clinical archetypes to SNOMED CT

J.L. Allones[1], M. Meizoso[1], M. Taboada[1], D. Martinez[2], and S. Tellado[2]

[1] Department of Electronics and Computer Science, University of Santiago de Compostela, Spain (joseluis.allones@usc.es maria.meizoso@usc.es maria.taboada@usc.es)

[2] Department of Applied Physics, University of Santiago de Compostela (diego.martinez@usc.es serafin.tellado@usc.es)

Abstract. Semantic interoperability of health systems will be only possible if clinical data models, such as OpenEHR Archetypes, are agreed by experts and aligned to standard terminology systems. In this paper we present an automated approach combining mapping algorithms to align clinical archetype terms to SNOMED CT concepts.

Keywords: Terminology Mapping, Knowledge Representation, Clinical Archetypes, SNOMED-CT, Semantic Interoperability

1 Motivation of the work

Providing semantic interoperability between health information systems is one of main current challenges of eHealth [1]. On the one hand, biomedical terminologies have been proposed as standard to code patient data and thus avoid misinterpretation of patient data and the errors usual in traditional paper records. One of the most important examples is SNOMED CT [2], which is determined to provide comprehensive terminology for encoding all aspects of Electronic Health Records (EHRs). On the other hand, nowadays important organizations [3] are moving up the use of entry forms to capture clinical data in a structured way in EHRs. Clinical data models, such as the European openEHR Archetypes [4], have been proposed for defining the structure of the information to be captured by these data entry forms. Mapping clinical terms in free text of these clinical data models to concepts of standard medical terminologies is a crucial step to provide interoperability between health information systems. However, at the present time, openEHR archetypes of the most important repositories contain mostly free text, and links to terminologies are infrequent. Because some medical terminologies, such as SNOMED CT [2], contain about 300,000 medical concepts, the manual mapping becomes very time-consuming. Therefore, we propose an automated approach to bind archetype clinical terms in free text to their equivalent SNOMED CT concepts.

2 Materials

2.1 OpenEHR archetypes

OpenEHR archetypes provide a detailed structure to model and define the clinical information required to record a particular clinical statement (e.g., tobacco

use). Archetypes are modeled using the Archetype Definition Language (ADL), which is comprised of three sections (see Fig. 1):

- **Header section** including metadata about the archetype.
- **Definition section** which is hierarchically organized and includes the structure, values and occurrences related with clinical data.
- **Ontology section** including *term definitions* and *term bindings* subsections. *Term definition* consists of the names and descriptions of the clinical terms present in the body section. *Term binding* stores the bindings between clinical terms (referenced by local archetype identifiers) and biomedical terminology concepts.

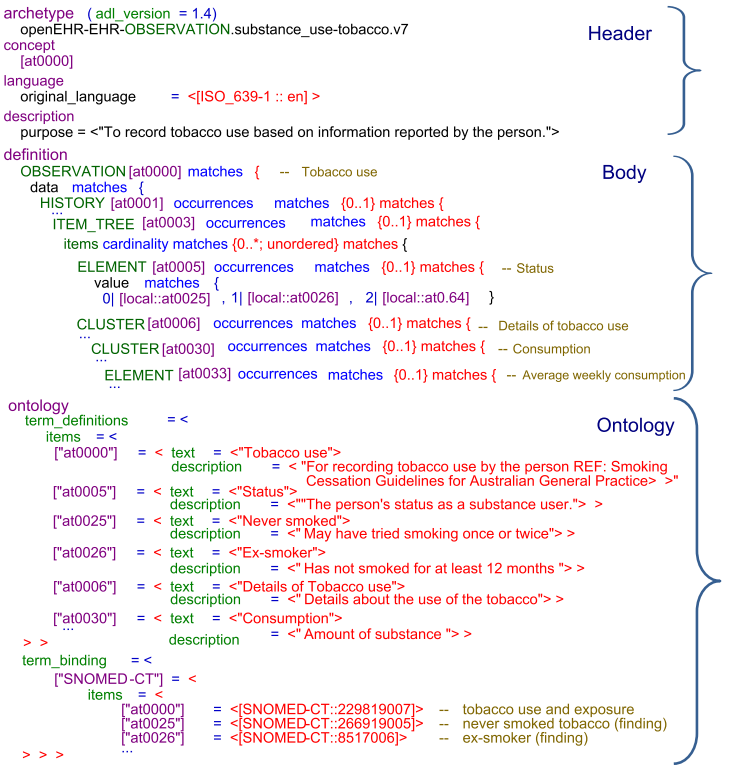

Fig. 1. Tobacco-use archetype (adl file).

We have chosen a set of 25 openEHR observation archetypes from the repository created by NHS [5]. An ADL parser from the openEHR implementation was used to remove the details of archetype syntax and preserve only the clinical terms with clinical meaning. The result of parsing is a dependency tree of terms representing the hierarchical organization of the archetype (see Fig. 2).

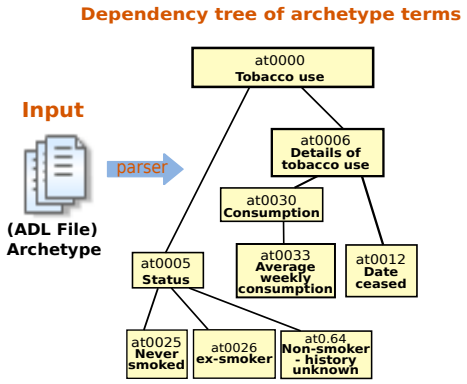

Fig. 2. Part of the tree resulting of parsing the ADL file shown in Fig 1.

2.2 SNOMED CT

SNOMED CT is a comprehensive clinical terminology that provides a standard for clinical information [2]. It contains about 300,000 active concepts, which have an unique id and several associated descriptions, each one representing a human-readable term that describes the same clinical idea (i.e., a synonym). Each concept has one or more semantic relationship to other concepts. These relationships can be classified in the following two categories. The hierarchical relationships *IS A* represent a concept by linking it with other concept in a subsumption hierarchy. The attribute relationships associate two concepts specifying a characteristic of one of these concepts. For example, the *interprets* relationship designates the judgement aspect being evaluated or interpreted.

2.3 UMLS

The Unified Medical Language System (UMLS) [6] is a repository of biomedical ontologies and vocabularies developed by the US National Library of Medicine (NLM). UMLS provides several terminological resources [7][8] , which use a knowledge-intensive approach based on symbolic, natural-language processing and computational-linguistic techniques, in order to discover biomedical concepts referred in a phrase or text.

3 Methods

Several mapping techniques, classically used in ontology matching [9], were developed to bind archetype clinical terms to SNOMED CT concepts:

3.1 Name-based techniques

Name-based techniques search SNOMED CT concepts with some description (preferred term or synonym) lexically similar to the archetype terms. These techniques involve the following steps (see Fig. 3) :

– Preprocessing archetype terms as follows:
  - First, terms are tokenized into their constituent words (tokens). Prepositions and broad and unspecific words are discarded.
  - Next, tokens are concatenated in different orders, so several variations of the original term are generated. Furthermore, because sometimes archetype nodes may not have been modeled explicitly in an archetype (i.e., their meanings are completed or deduced from other information of the archetype), they are not specific enough, and therefore tokens are also concatenated with upper level nodes from the archetypes hierarchy in order to obtain more specific statements. For example, in Fig. 3, the term *1 minute* is combined with the tokens of the root term *apgar score*, generating the new term *apgar 1 minute*.
  - Finally, terms are normalized, including plurals, singulars, case-insensitive, replacement of abbreviations making use of the SNOMED Word Equivalent Table, which provides the most used abbreviations.
– Normalization of SNOMED CT concept descriptions exactly like the archetype terms.
– Searching for lexical matching between normalized archetype terms and concept descriptions. Two lexical matching are considered: *full match* (or exact match) occurs when an archetype term is exactly the same as some SNOMED CT description and *partial match* (or approximate match) occurs when the archetype term is contained inside some SNOMED CT description.

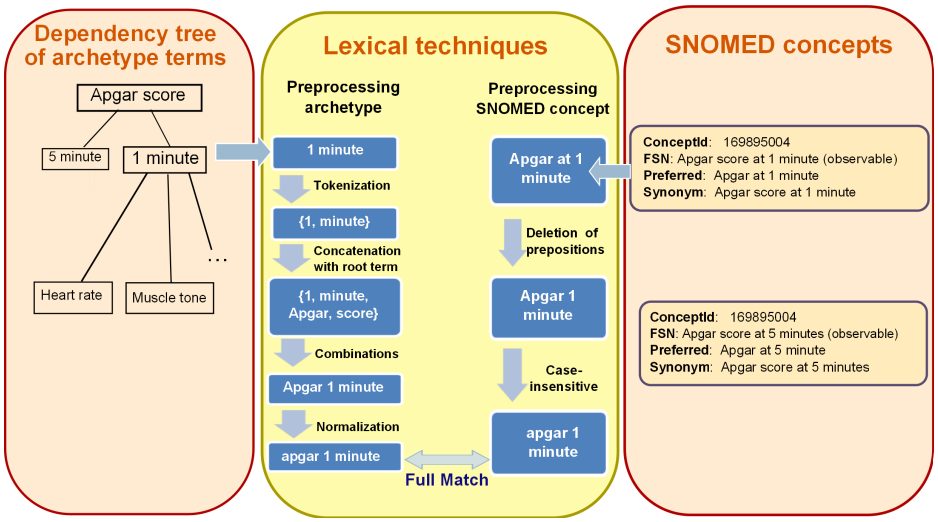

Fig. 3. Steps to lexically map archetype terms to SNOMED CT concepts. An example of application is shown for the *apgar score* archetype

3.2 External terminological resources

Several UMLS operations (*Exact search*, *Truncate Search*, *Normalized word* and *Metamap*) were applied in different stages during automatic mapping process. *Exact Match* retrieves only concepts that include a synonym that exactly matches the archetype term. *Normalized word* removes lexical variations such as plural and upper case text and compares archetype terms to the Metathesaurus normalized word index. *Truncate Search* retrieves concepts with synonyms that begin or end with the letters or words of the archetype term. *MetaMap* is a highly configurable program to map biomedical text to the UMLS Metathesaurus concepts, which has several high-level components: tokenization, part-of-speech tagging, lexical lookup, acronym/abbreviation identification, syntactic analysis, variant generation, etc.

3.3 Structure-based techniques

Our approach, unlike other works in this field [10] and [11] , exploits the structural similarity between subtrees of the archetypes and SNOMED CT subgraphs obtained through *interprets* and *IS A* relationships. As a result, candidate bindings are detected identifying archetypes terms and SNOMED concepts with similar neighbors. For example, in Fig. 4 considering that previously a lexical mapping between term *Depth* of the archetype *Respiration* and the SNOMED CT concept *Depth of respiration* was detected; the method traverses the relationship *interprets* of the concept *depth of respiration*, and extracts a set of candidates concepts (*Shallow breathing*, *Normal breathing*, *Deep breathing*, *Depth of breathing uneven*, etc.) to map to the values of the term*Depth* (*Shallow*, *Normal*, *Deep*). Then, the method applies *partial match* to bind *Shallow* to *Shallow breathing*, *Normal* to *Normal breathing* and *Deep* to *Deep breathing*.

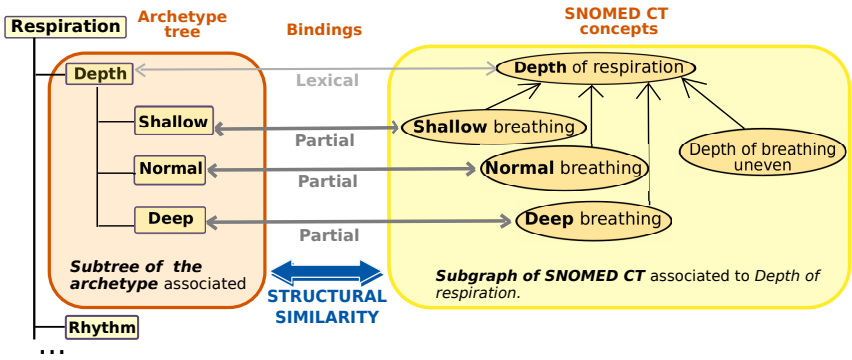

Fig. 4. Structural similarity between a subtree of respiration archetype and a SNOMED CT subgraph

4 Results and Discussion

A set of 25 OBSERVATION archetypes, with a total of 477 manually linked terms, was used to test the automated approach. The terms were linked with

96.1% precision, 71.7% recall and 1.2 concepts on average for each mapping. Our study also revealed that a high ratio of archetype terms which are semantically related. Unlike other works in automated mapping of archetypes [11][10][12], our approach took advantage of the SNOMED relationship structure, restricting the use of approximate lexical techniques to relevant SNOMED portions. Thus, it was possible to map undetected SNOMED concepts by more exact lexical techniques. This contribution allowed to rise the recall of our method by 10%.

5 Conclusions

The aim of OpenEHR archetypes is sharing clinical data in a unambiguous and accurate way. Standard terminologies, such as SNOMED CT, provide an appropriate method of expressing unambiguous and interoperable clinical data terms. Therefore, we propose an automated method to map archetype terms to SNOMED CT concepts. The method exploits the SNOMED CT relationships to limit the searches to relevant portions of the terminology. This research shows that it is possible to automatically map archetype terms to a standard terminology with a high precision and recall, with the help of appropriate contextual and semantic information of both models.

Acknowledgements This work has been funded by the Ministerio de Educación y Ciencia, through the national research project *Gestion de Terminologias Medicas para Arquetipos* TIN2009-14159-C05-05

References

1. SemanticHEALTH project: Semantic interoperability for better health and safer healthcare. http://ec.europa.eu/information_society/activities/health/docs/publications/2009/2009semantic-health-report.pdf
2. SNOMED-CT: Systematized nomenclature of medicine-clinical terms. http://www.ihtsdo.org/snomed-ct/
3. NHS: Connecting for health project. http://www.connectingforhealth.nhs.uk/ (2010)
4. OpenEHR: archetypes. http://www.openehr.org/
5. NHS: NHS connecting for health archetype repositories. https://svn.connectingforhealth.nhs.uk/svn/public/nhscontentmodels/TRUNK/cm/archetypes/
6. UMLS. http://www.nlm.nih.gov/research/umls/
7. Metathesaurus: Web service operations. https://uts.nlm.nih.gov//doc/devGuide/webservices.html#meta
8. MetaMap. http://mmtx.nlm.nih.gov/
9. Euzenat, J., Shvaiko: Ontology Matching. Springer-Verlag (2007)
10. Yu, S., Berry, D., Bisbal, J.: An investigation of semantic links to archetypes in an external clinical terminology through the construction of terminological "shadows". In: IADIS, Freiburg, Germany (2010)
11. Qamar, R.: Semantic Mapping Of Clinical Model Data To Biomedical Terminologies To Facilitate Interoperability. PhD thesis, University of Manchester (2008)
12. Lezcano, L., Sanchez-Alonso, S., Sicilia, M.: Associating clinical archetypes trough umls metathesaurus term clusters. Journal of Medical Systems (2010)