Using SNOMED CT Concept Definitions for Natural Language Processing

*Presenter: Presenter: April Russell, Hua Cheng, James Chan, Thomas Polzin and Gilan Saadawi - M*Modal*

Audience

Healthcare providers interested in mapping clinical narrative to SNOMED CT and SNOMED extensions.

Objectives

Using the SNOMED CT hierarchy and concept definitions to automatically develop a Natural Language Processing resource for the accurate detection of concepts in clinical narratives.

Abstract

The coverage and detail of SNOMED CT is potentially a very valuable resource for Natural Language Processing (NLP). However, SNOMED CT concept definitions are clinically motivated and not linguistically. As a consequence, the correlation between a concept's descriptions (fully specified name, preferred name, or synonyms) and its manifestation in clinical narratives is rather loose. Finding a complex SNOMED concept in a text document using simple NLP techniques such as string matching based on the concept's descriptions tends to perform very poorly. In this work we attempt to use the SNOMED concept hierarchy and definitions to derive a meaningful NLP resource for robust concept recognition within clinical narratives. Within SNOMED CT each concept is defined by a set of relationships to other concepts. We show that we can use these relationships, for instance "finding site" or "morphologic abnormality", to automatically derive a resource that allows the NLP to find theses relevant subcomponents and, given the proper syntactic context, to aggregate these subcomponents back into the overall concept. Thus, instead of locating concept descriptions in the narrative, we match relevant parts of a concept definition and define an aggregation rule that assembles theses parts to a coordinated concept given that the proper syntactic constraints are met. We discuss here two different approaches:

Method 1: We automatically derive aggregation rules based on fully defined concepts in certain subsets within SNOMED CT with minimal manual modifications. We match the relevant parts independently with robust string matching techniques and then apply these rules to aggregate the matched parts to the overall concept given that the proper syntactic constraints are met. For example, consider the concept *'125605004 Fracture of Bone' (Disorder)*. We use the two defining relationships, associated morphology *of '72704001 Fracture' (morphologic abnormality)* and finding site of *'272673000 Bone structure' (body structure)* to automatically compile aggregation rules for all concepts related by an IS-A relationship to this concept. This automatic approach can also be used to derive aggregation rules that associate laterality information with body structure concepts.

Method 2: Unfortunately the automated compilation does not carry over to primitive SNOMED concepts since these do not contain sufficient inherent structure to derive adequate aggregation rules. Ironically, this is also true when fully defined concepts are too granular or too detailed. The resulting aggregation rules would require parts the clinical narrator simply does not document. In order to deal with these two cases, we manually define guidelines that either explicitly add relationships (for primitive concepts) or change or remove existing relationships (for fully defined concepts). Additionally, we can also automatically stipulate particular syntactic constraints based on a concept definition. For instance, the value of the "device used" relationship is frequently realized as a prepositional phrase in the respective narrative. Method 2 was very useful in post coordinating procedures using the following rule:

BaseProcedure #|ProcedureSite|ProcedureMorphology|ProcedureApproach|UsingDevice

We believe that using these two approaches has empowered our NLP to effectively map narrative to SNOMED CT concepts in clinical documentation.