SNOMED CT Quality Assurance via a Lexical Approach

Ibrahima Niang, Dr. Yan Chen

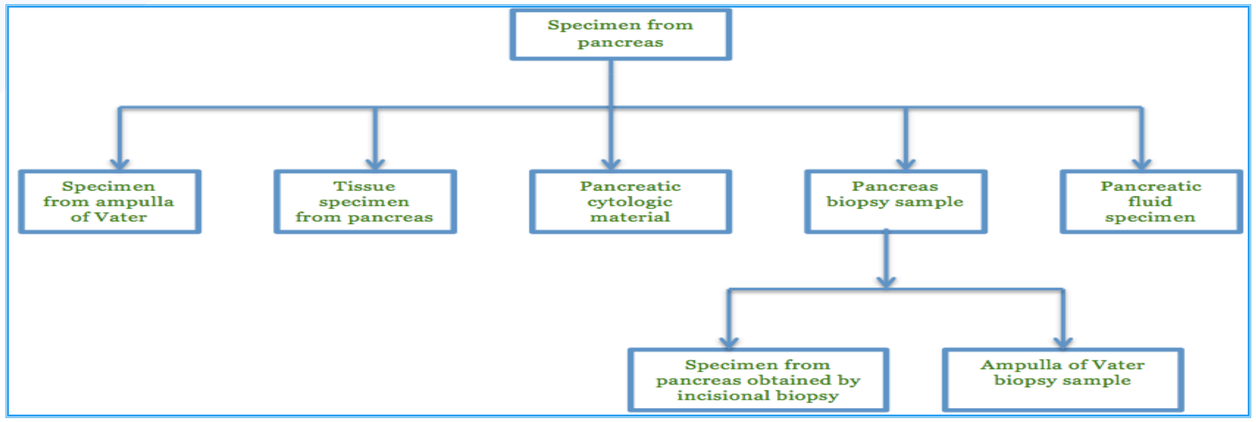
Computer Information Systems Department, Borough of Manhattan Community College, CUNY

Abstract

The SNOMED Clinical Terms (SNOMED CT) is the most comprehensive, multilingual clinical healthcare terminology in the world, which provides the core general terminology for the electronic health record (EHR). The medical terminology contains more than 311,000 active concepts with unique meanings and formal logic-based definitions organized into hierarchies. SNOMED CT is continuously updated to meet the needs of users around the world. Revisions are released twice a year. Due to its broad scope and inherent complexity, it is unavoidable that errors will find their way into SNOMED's knowledge content, particularly as it continues to expand. Thus, quality assurance is essential to ensure the integrity of its contents. In this research, we will study a lexical approach to aid the quality assurance of SNOMED CT. In particular, we will focus on the wrong "is-a" relationships (or hierarchical relationship). Errors discovered during the auditing process will be reported to help ensure the quality and safety of the medical terminology.

Methodology

The first step is concepts reading. Since SNOMED CT is organized into hierarchies, recursion will be used to get each concept and their children. The second step is concept refinement. Stop-words like a", "the", "of" in the FSN will be removed since they are insignificant. The third step is complexity comparison. This comparison identifies potential erroneous relationships by determining if the parent is a substring of child. The final step is the errors report file. Potential wrong relationships will be reported in a file along with the hierarchies and identifiers to help auditors improve the quality of the medical terminology.



Results

To test the efficiency of the algorithm, we have called the method IndentifySuspiciousISA(…) on a random hierarchy: the specimen hierarchy. The results obtained consist of 177 parent-child pairs of potential erroneous wrong is-a relation.

Concept Id	Concept FSN
123038009	Specimen (specimen)
48469005	Cytologic material (specimen)
123038009	Specimen (specimen)
168121005	Miscellaneous samples (specimen)
123038009	Specimen (specimen)
257261003	Swab (specimen)
123038009	Specimen (specimen)
258415003	Biopsy sample (specimen)
258415003	Biopsy sample (specimen)
122550002	Specimen obtained by fine needle aspiration procedure (specimen)
122550002	Specimen obtained by fine needle aspiration procedure (specimen)
309061008	Breast fine needle aspirate sample (specimen)
122550002	Specimen obtained by fine needle aspiration procedure (specimen)
309146009	Thyroid fine needle aspirate sample (specimen)
122550002	Specimen obtained by fine needle aspiration procedure (specimen)
309508006	Soft tissue lesion fine needle aspirate sample (specimen)
258415003	Biopsy sample (specimen)
396483002	Specimen from skin obtained by shave excision (specimen)
123038009	Specimen (specimen)
258433009	Smear sample (specimen)

Introduction

The Systematized Nomenclature of Medicine-Clinical Terms (SNOMED CT) is a multilingual clinical healthcare terminology, which provides the core general terminology for the electronic health record (EHR). Currently used in 52 countries, SNOMED CT contains more than 311,000 active concepts with unique meanings and formal logic-based definitions organized into 19 hierarchies. SNOMED CT can be used to represent clinically relevant information consistently, reliably and comprehensively as an integral part of producing EHRs.

The SNOMED CT concept model specifies the way in which SNOMED CT concepts are defined using a combination of formal logic and editorial rules. SNOMED CT content is represented using three types of component: Concepts, Descriptions and Relationships.

A concept is a clinical idea associated with a unique identifier (up to 18 digits). The meaning of the term is specified by an association with a term known as the Fully Specified Name. Concepts are organized in hierarchies, from the general to the specific. This allows detailed clinical data to be recorded and later accessed or aggregated at a more general level. Currently, There are well over 300,000 active *concepts* in the SNOMED CT.

SNOMED CT descriptions link appropriate human readable terms to concepts. A concept can have several associated descriptions, each representing a synonym that describes the same clinical concept. Each translation of SNOMED CT includes an additional set of descriptions, which link terms in another language to the same SNOMED CT concepts. Every description has a unique numeric description identifier.

SNOMED CT relationships link concepts to other concepts whose meaning is related in some way. These relationships provide formal definitions and other properties of the concept. Each relationship includes: a unique identifier, the identifier of the source concept, the identifier of the relationship type concept ("is a" or "has a") and the identifier of the destination concept.

Due to SNOMED's broad scope and inherent complexity, it is unavoidable that errors will find their way into SNOMED's knowledge content, particularly as it continues to expand. One of the highest priorities for Members of IHTSDO is to ensure the quality and safety of SNOMED CT.

Implementation

The implementation of the algorithm I designed is described as the following in JAVA:

```
/*
 * @author Ibrahima Niang
 */
public class Dataline {

public static void main(String[] args) {
        try
         {
                Scanner in = new Scanner(System.in);
                System.out.println("Enter a conceptId");
                String  s = in.nextLine();
                long conceptId= Integer.parseInt(s);
                 IndentifySuspiciousISA(conceptId);
         }
        catch (Exception e)
      {
         System.out.println(e.getMessage());
      }
}
public static void IndentifySuspiciousISA(long concept) throws IOException
{…}

public static boolean isStopWord(String word) throws IOException
{…}

public static String RemoveStopWord(String concept) throws IOException
{…}

public static boolean isSuspicious(String parent, String child) throws IOException
{…}

 public static void saveRelationship(String parent, long parentId, String child, long childId) throws IOException
 {…}
.
```

Discussion

Our goal to develop an efficient algorithm to identify and report potentially erroneous relationships is almost met. For now we have the algorithm running. Moving forward, I will be working on testing the validity of the algorithm to the 19 hierarchies of SNOMED CT. So far, the results obtained are only potential candidate for wrong is-a relationship. Hence, the opinion of an auditor is needed to ensure the validity of these first results.

Acknowledgement

Louis Stokes Alliance for Minority Participation (LSAMP). NYC

References

SNOMED CT Starter Guide. 3rd ed. US:IHTSDO, 2014. Print.

Agrawal, Ankur, Yehoshua Perl, Yan Chen, Gai Elhanan, and Mei Liu. "Identifying Inconsistencies in SNOMED CT Problem Lists Using Structural Indicators." AMIA Annual Symposium Proceedings. American Medical Informatics Association. Web. 4 Oct. 2014. <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3900119/>.

Dr. Linda Bird. An Introduction to SNOMED CT Implementation [PDF document]. Retrieved from Lecture Notes Online Web site: http://www.himssasiapac.org/14/docs/speakersPresentations/HIMSSAP_14DHW_SpeakersPresentations_SNOMEDCT_LindaBird.pdf

Sticco, Caitlin. Using the SNOMED CT Error Taxonomy to Maximize the SNOMED CT Content Request System. Web. August 2011.