Semantic Tagging of Medical Narratives with Top Level Concepts from SNOMED CT Healthcare Data Standard

Saman Hina
School of Computing
University of Leeds

Eric Atwell
School of Computing
University of Leeds

Owen Johnson
School of Computing
University of Leeds

Abstract

*Medical narratives written by clinicians constitute critical information in healthcare domain and are required to be correct with respect to contextual meaning. SNOMED CT (**S**ystematized **No**menclature of **Med**icine -- **C**linical **T**erms) is a standardized reference terminology that consists of 390023 SNOMED CT concepts with SNOMED CT codes. This paper describes the extraction of SNOMED CT concepts from free text discharge summary reports. For the evaluation of the medical concepts, we used 300 discharge summaries corpus provided by University of Pittsburgh Medical Centre, and compared it with the SNOMED CT concept file which is a preprocessed and cleaned file listing SNOMED CT concepts. In this paper we present the ongoing research on SNOMED CT concept extraction from discharge summaries using natural language processing and introducing SNOMED CT core concepts as a single gazetteer list for concept extraction. Out of 390023 concepts, 21563 concepts were found in the test set of discharge summaries from the SNOMED CT core concepts gazetteer list. A modified approach extracted the 23 top level concept tags which will be useful for linguist analysis research.*

1. Introduction

The secure transfer of medical narratives in the form of discharge summaries, progress notes, verbal autopsies, etc., require data quality. Data quality is the most important factor in healthcare domain. This is because a patient's data could be analyzed by various potential users such as clinicians, healthcare providers, and researchers using disparate data sources which need to be standardized in such a manner that the data could be retrieved in a structured format. To meet the requirements of data quality, domain specific data standards containing authentic and useful information which helps in explaining the domain knowledge semantically, have been developed. Data standards are fundamental to the seamless exchange of data and they help improve the ability of partners (internal and external) to exchange data efficiently and accurately. They also assist secondary data users understand, interpret, and use data appropriately. SNOMED CT is a standardised reference terminology that consists of millions of SNOMED CT concepts for medical domain. It was developed by the College of American Pathologists and the United Kingdom's National Health Service by merging two health care terminologies SNOMED RT and Clinical Terms Version 3 [1]. For the structure of classification of different coding terminologies such as READ Codes, SNOMED RT and SNOMED CT, it is clear that SNOMED CT will be the most suitable choice in the UK NHS (National Health Service) [2].

2. Background

Clinical documents such as discharge summaries, patient's consultation notes, etc., are written by clinicians in natural language free text and contain medical concepts in unstructured format. Some of the sections in medical documents such as discharge summary reports or progress notes consist of natural language free text along with medical concepts. These sections come under defined headers, for example, final diagnosis and discharge instructions. The problem with defined sections is that they are not always in one format and defined in different cases. Adding to the complexity of problem. headers cannot be identified by any punctuation mark such ":" or ";", and the medical concepts under each section cannot be identifed with these section names.

To explain these concepts semantically, annotation guidelines are required to be provided by domain experts. The authenticated and standardized terminology thus obtained enables disparate systems in health care sector to maintain structured and consistent information [3]. A strategy to extract diseases and procedures using very limited amount of natural language processing in finding phrases in the discharge summary has been reported. UMLS was used to examine the type of concept [4]. Another system used lexicons for the semantic matching words and phrases [5]. Health Information Text Extraction (HITEx) tool also used the UMLS Database to extract UMLS concepts for the principal

diagnosis [6]. A number of studies report other NLP systems for the information extraction such as MEDLEE [7], AMBIT: Acquiring Medical and Biological Information from Text [8], and MetaMap [9] for different evaluations in biomedical field. In the present study, we have used SNOMED CT concept file provided by Leeds Institute of Health and Sciences.

3. Methodology

In the present study, an incremental approach to improve the information extraction from medical narratives has been adopted. Dictionary method was adopted as an initial implementation for the SNOMED CT concept extraction, followed by using Hash Gazetteer which is CREOLE Processing resource in GATE [10], [11]. To improve concept extraction task, this approach has been improved and modified by classifying all the concept tags of SNOMED CT. This classification will help the researchers to analyze the corpus with respect to each concept type. For example, any potential user can identify that how many concepts are disorders, substance, etc. This has been done by extracting specific tags from the underlying model in the SNOMED CT original concept file.

3.1. Dictionary Based Concept Extraction

For our preliminary approach we developed subset of SNOMED CT concepts as a python dictionary containing single word as well as multiword concepts as the type and SNOMED CT code as a value. AutoSlog has been reportedly used to assemble automatic dictionary for the extraction of information specific to domain [12], however, in present study the dictionary was manually created using python dictionary method. Natural language written by a clinician involves data that can also give partial wrong matches. It can be read as a concept but not completely extracted as required in the document. Single word concepts can be perfectly matched but the concepts having more than one word sometimes also contain single concepts in them. For example; “cough with fever” has two single concepts “cough” and “fever”. Figure 1 shows the single concept extraction from a discharge summary. This method also used some basic NLP analysis like token splitting, punctuation removal and change case for the best possible retrieval from the dictionary values. However, we realised that the method was only matching single word concept and was not equally efficient in detecting multiword concepts. Another interesting observation made was that “sprain” is single word concept but it is missing in the ouput file. This is because in SNOMED CT concept file, “sprain” exists as multiword concept as “Ulnohumeral sprain”.

Diagnosis :
Left Shoulder / Neck strain / sprain .
Nausea, Constipation

Treatment Rendered :
Dilaudid 2mg PO x 1 Flexeril 10 mg x 1
Discharge Medications :
Flexeril 10mg TID

Disposition , Follow up and Instructions to Patient :
Please follow-up with your Orthopaedic physician on Wednesday 6/9/02 as previously arranged by patient . Please follow-up with PCP . Take codeine prescribed by PCP with food and water to prevent nausea and constipation .

Outputfile:

Extracted Concepts with codes:
Codeine/ 85990009
Nausea/422587007
Constipation/14760008

Missing Concepts:
Left Shoulder
Neck strain
sprain

Figure 1. Example of portion from discharge summary and single concepts extracted from dictionary method

In an attempt to achieve a better output for the extraction of SNOMED CT concepts in the corpus of 300 discharge summaries, we applied language processing resources to capture core concepts from the natural language text. For this reason we extracted 390023 SNOMED CT core concepts as gazetteer for the concept extraction. These core concepts include finding, disorder, specimen, product, body structure, organism, event, physical objects, procedures, special concept, staging scale, observable entity, and substance. This baseline approach aimed to find concept existence in gazetteer list and later on SNOMED CT codes and categories were added as a feature to each annotation. These SNOMED CT core codes were the features of concept types. In addition to identification of SNOMED CT concepts, start and end offset of SNOMED CT concepts were displaced in the annotation lists.

3.2. Natural Language Processing

For preprocessing the data, GATE – General Architecture for Text engineering [13] was used for

applying language processing stages over the corpus. These stages include tokenizer, sentence splitter and part of speech tagger, Noun phrase chunker and name entity recognition. SNOMED CT concept Gazetteer was placed at the end of all processing resources to check the maximum possible overlapping between language chunks and SNOMED CT core concepts. It was found that it gives error in running the application in this sequence. This is due to ANNIE running with defaults. Later on, we placed SNOMED CT concept Gazetteer after sentence splitter as suggested by GATE message pane.

Many NLP systems can perform variety of tasks that can run in a modular approach to obtain the scope of encoding medical text in natural language. For example, Health Information Text Extraction (HITEx) tool developed at National Center for Biomedical Computing and Informatics for integrating Biology and Bedside (i2b2) [14], also selected GATE- General Architecture for Text Engineering [13].

3.3. SNOMED CT Concepts Gazetteer

We introduced SNOMED CT core concepts as the SNOMED CT concept gazetteer list in the processing resources. This was followed by a series of language processing to run over the corpus and find possible relations identified in each document.

Noun Phrase Chunker was used in order to analyze the possible occurrences of SNOMED CT read as noun phrase. Figure 2 shows all the processing resources used over the discharge summary reports.

The sequence was followed by running "ANNIE-A Nearly New Information Extraction" application with defaults. The placement of the gazetteer list was followed according to ANNIE pipeline to avoid error in running the application.

- In the first step, our application reset all the previous annotations to re-initialize and to avoid any duplication of annotation in running every processing resource.
- The Tokeniser split the token annotations, then the sentence splitter split into sentence annotations. This resource constructed rational combinations like "haven't" to "have n't" and produced token and space token annotations according to part of speech tagger requirements.
- Sentence Splitter finds sentences on the basis of tokens and it uses the gazetteer of abbreviations to annotate sentence splits.
- The SNOMED CT concept gazetteer generates the list of 390023 core concepts to find out the

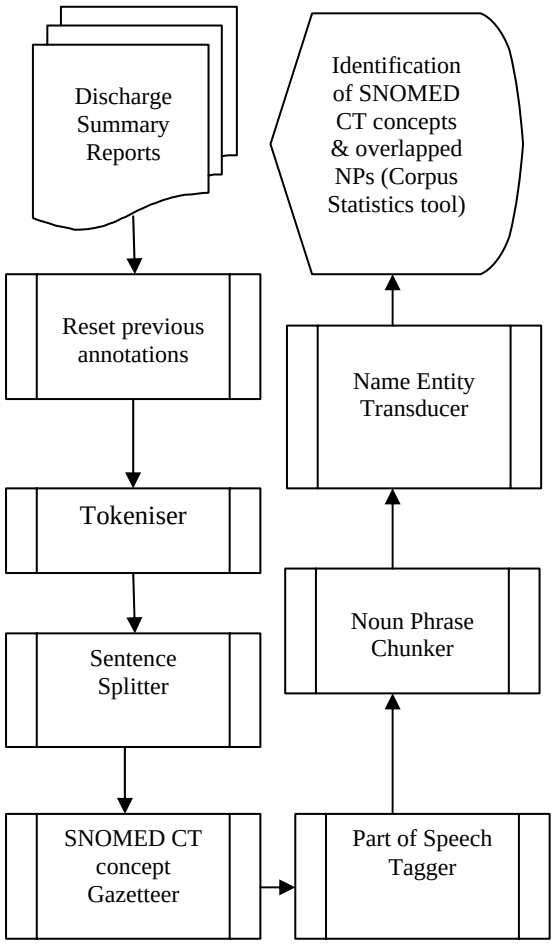

Figure 2. Core Processing resources including SNOMED CT concept gazetteer

to find out the possible occurrences of SNOMED CT in the corpus.

- The Part of Speech tagger adds POS feature values to the tokens. This tagger used Penn Treebank tag set used in the analysis of medical notes [15].
- Noun phrase chunker finds noun phrases present in each document. This phase refined the application into more specific approach that analyzed the relation of each SNOMED CT concept having long phrases with the existing noun phrases.
- In this case, Named Entity Transducer is used to find the ambiguities of the concepts with the other annotations. The SNOMED CT concept gazetteer finds some entities that were suggested by the gazetteer but were ambiguous. For example; "^210m^Bismuth (substance)" can be identified as person name "Bismuth". Named Entity Transducer contains number of JAPE

grammars to define patterns and eliminate ambiguities to some extent.
- Corpus Statistics component was used to calculate the summary of token types compared with the original markup.

3.4. Classifying SNOMED CT Concepts Gazetteer into SNOMED CT Concept Tags

An advantage of SNOMED CT data standard is that besides clinical terms, general concepts such as occupation, social concept, physical object, life style, physical force, etc., can be captured as well. The previous system flow was capturing noun phrases overlapping within the concepts while the extended system flow is expanding purely on the structure of SNOMED CT data standard using the multiple gazetteers with respect to different concept tags. Figure 3 shows the expansion of SNOMED CT data standard into important concept classes.

SNOMED CT Top Level Concept Tags

Administrative concept
Attribute
Body Structure
Disorder
Event
Morphologic Abnormality
Navigational concept
Observable Entity
Occupation
Organism
Person
Physical Force
Physical Object
Procedure
Product
Qualifier value
Record Artifact
Regime/Therapy
Situation with explicit context
Social concept
Substance

Figure 3. List of SNOMED CT top level concept classes

The advantage of using underlying model discussed in preceding text is to help the users in understanding the medical narratives more specifically by extracting every concept by its class name. The researchers will then have the advantage of using this approach to categorize the concept and then identify the potential relationship between the concepts. For example;

- Which product is prescribed for any particular disorder?
- Which Physician has prescribed the product?

The data standard SNOMED CT is well defined with every concept associated with its class name which is helpful for end user to understand the type of concept more specifically. The extraction of underlying model was done from SNOMED CT original concept file. Instead of using single gazetteer to extract the concepts from the corpus, the current system has been modified by developing separate gazetteers for each class defined in SNOMED CT file. Modified application pipeline has replaced SNOMED CT gazetteer into separate gazetteer, removed the noun phrase chunker and Named entity transducer. Results were better in terms of identifying the concepts separately with the respective tags compared to previous which were overlapping the multiword concepts with nested tags.

4. Evaluation

For the evaluation of our method we used the corpus quality assurance tool of GATE to analyze the frequencies of each annotation type. To check the overlapping of noun phrases we selected both annotation types to be compared within the corpus. Table 1 shows the results obtained from the corpus quality assurance tool.

Table 1. Frequencies of Annotation Types found in the corpus from single gazetteer method

Annotation	Frequency
SNOMED CT concepts lookup	35892
Sentence	13480
Noun Phrases Overlapping	54269
Micro Summary	103641

By implementing the method of using separate gazetteers, the comprehensible representation of the top level SNOMED CT concepts tags has been achieved. By extracting concept with respect to top level SNOMED CT concept tags and the frequency of each concept tag is given in Table 2. The evaluation has been made over the same corpus of 300 discharge summaries.

Results have shown the clear difference of concept extraction from previous method by disambiguated nested multiword concept by

assigning separate tag of their own top level concept tags.

Table 2. Frequencies of Top Level Class Tags found in the corpus from multiple gazetteer method

Top Level Class Tags	Frequency
Administrative concept	28
Attribute	280
Body Structure	14
Disorder	560
Regime/Therapy	45
Body Structure	14
Event	8
Morphologic Abnormality	144
Navigational Concept	1
Observable Entity	83
Occupation	18
Organism	31
Person	39
Physical Force	1
Physical Object	7
Procedure	115
Product	502
Qualifier value	3093
Record Artifact	1
Regime/ therapy	45
Situation	121
Social concept	10
Substance	322
Total Concepts	**5482**

5. Discussion

The present study evaluated the possible occurrences of SNOMED CT concepts in large amount of natural language free text present in clinical documents and aimed to identify the relationship between the noun phrases contained the SNOMED CT core concepts. The processing resources employed were challenging in preprocessing the data with language processing and using gazetteer list to identify the SNOMED CT core concepts. Some researchers used Link Grammar Parser for parsing English language text for clinical documents [16]. It was also used in the present study for initial experiments but it was found advanageous to use a single tool for all processing resources. It has maximized the possible findings of SNOMED CT top level concepts in the corpus and also need to define the relationships between the SNOMED CT top level concepts. Many SNOMED CT core concepts were found correctly but not all of the findings were noun phrases. The findings of SNOMED CT core concepts were accurate to some extent but still have discrepancies, for example reading a label "Diagnosis #1:" or "Discharge Diagnosis" found in annotation list having "major type = concept". Similarly, "Bladder Carcinoma" found as a noun phrase but not found by the SNOMED CT concept list. These inaccuracies have to be corrected either manually or by writing JAPE - Java Annotation Patterns Engine rules [17]. These rules will also include inter sentence patterns as depicted in Figure 4.

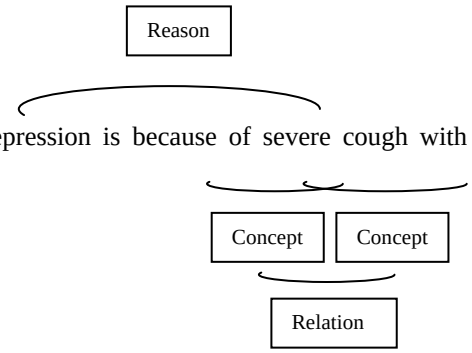

Figure 4. Example of inter sentence pattern

Our approach aimed to provide decision making feature for potential users in biomedical field and NLP researchers. The approach used by HITEx is close to our study but UMLS tables for mapping the concepts were used [6]. To the best of our knowledge SNOMED CT reference terminology is the most consistent and complete vocabulary covering maximum concepts and the relationship between the concepts. In present study, Snoflake Browser by Dataline Software [18] was found useful in studying the classification and relationship between the SNOMED CT concepts and detailed description including the information about CTV3 and SNOMED RT codes for each concept. Another browser has been used to verify the concept classes extracted from the original SNOMED CT concept file [19].

6. Conclusion

Natural Language Processing is very effective in extracting valuable information from large amount of texts in clinical documents. Our application involved a series of linguistic processing to identify the SNOMED CT concepts and successfully recognised

core concepts from the gazetteer list. 35892 concepts were found by developing single gazetteer with ambiguities in multiword concept. By extending this approach to multi gazetteer 5482 concepts were found in total 23 concept classes, which is considerable improvement in current application pipeline. The present research study is continuing and needs to be sophisticated in terms of annotating the missing SNOMED CT concepts with respect to each concept class and resolving the remaining problem of ambiguities in the concepts.

7. Future Work

It is envisaged in future embodiments of present study to insert the whole structure of SNOMED CT which is expected to give appropriate and perfect results extracted from natural language free text present in discharge summaries or any clinical document having natural language free text. This will require the addition of remaining concept classes in the gazetteer list. By structure we aim to define relationship between the SNOMED CT concepts. This will authenticate consistent information exchange between potential users such as clinicians and researchers, etc.

8. Acknowledgements

This work was supported by Leeds Institute of Health Sciences in terms of providing essential trainings about “Data Standards and Terminologies” and University of Pittsburgh Medical Centre for providing the data under Data Use Agreements. Ms Hina's research is supported by NED University of Engineering and Technology, Karachi, Pakistan. Authors would also like to acknowledge Abdul Baqi Sharaf for his support and Niraj Aswani from GATE team members for advising technically using the GATE software.

9. References

[1] M. Q. Stearns, *et al.*, "SNOMED clinical terms: overview of the development process and project status," *Proc AMIA Symp*, pp. 662–666, 2001.

[2] E. Coiera, *Guide to Health Informatics* 2nd ed. London: Hodder Arnold 2003.

[3] H. Wasserman and J. Wang, "An Applied Evaluation of SNOMED CT as a Clinical Vocabulary for the Computerized Diagnosis and Problem List," *AMIA Annu Symp Proc*, pp. 699-703, 2003.

[4] W. Long, "Extracting Diagnosis from Discharge Summaries," *AMIA Annu Symp Proc*, vol. 2005, pp. 470-474, 2005.

[5] J. Patrick, *et al.* (2008). *Developing SNOMED CT Subsets from Clinical Notes for Intensive Care Service*. Available: www.hinz.org.nz

[6] Q. Zeng, *et al.*, "Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system," *BMC Medical Informatics and Decision Making*, vol. 6, p. 30, 2006.

[7] C. Friedman, *et al.*, "Automated encoding of clinical documents based on natural language processing," *J Am Med Inform Assoc*, vol. 11, pp. 392 - 402, 2004.

[8] R. Gaizauskas, *et al.*, "Acquiring Medical and Biological Information from Text (AMBIT)," in *Proceedings of UK e-Science All Hands Meeting 2003*, Nottingham, 2003, pp. 370 - 373.

[9] A. Aronson, "Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program," *Proc AMIA Symp*, pp. 17 - 21, 2001.

[10] H. Cunningham, *et al.*, "GATE - a TIPSTER-based General Architecture for Text Engineering. in he TIPSTER Text Program (Phase III) 6 Month Workshop," 1997.

[11] R. Gaizauskas, *et al.*, "GATE: an environment to support research and development in natural language engineering," in *Tools with Artificial Intelligence, 1996., Proceedings Eighth IEEE International Conference on*, 1996, pp. 58-66.

[12] E. Riloff, "Automatically Constructing a Dictionary for Information Extraction Tasks," presented at the Eleventh National Conference on Artificial Intelligence, Washington D.C, 1993.

[13] H. Cunningham, "GATE, a General Architecture for Text Engineering," *Computers and the Humanities*, vol. 36, pp. 223-254, 2002.

[14] D. Demner-Fushman, *et al.*, "What can natural language processing do for clinical decision support?," *Journal of Biomedical Informatics*, vol. 42, pp. 760-772, 2009.

[15] S. V. Pakhomov, *et al.*, "Developing a corpus of clinical notes manually annotated for part-

of-speech," *International Journal of Medical Informatics*, vol. 75, pp. 418-429, 2006.

[16] D. Sleator and D. Temperley, "Parsing English with a link grammar," 1991.

[17] Hamish Cunningham, *et al.* (2000, JAPE: a JAVA Annotation Patterns Engine

[18] D. S. Ltd. (2010, 14-04-2010). Available: http://snomed.dataline.co.uk/scripts/snomed-browser.dll?c7ee

[19] X-lab. (2010, 12-12-10). *NPEx SNOMED-CT Browser*. Available: http://snomedbrowser.co.uk/

