Extracting the Concepts in Clinical Documents Using SNOMED-CT and GATE

Saman Hina, PhD Student[1], Eric Atwell, PhD[1], Owen Johnson, MSc[1,] Rebecca West, MA[1]

[1]University of Leeds, Leeds, UK

Abstract

*This paper is in response to the fourth i2b2/VA Shared-Task and Workshop Challenges in Natural Language Processing for Clinical Data. The task was completed through the use of SNOMED-CT (**S**ystematized **N**omenclature of **M**edicine -- **C**linical **T**erms) a comprehensive international standard for clinical terminology and software components available within GATE (General Architecture for Text Engineering). We developed a rule based system called NLP-SNOMED which identified the concepts and noun phrases within a training corpus of 300 discharge summaries by using the JAPE rules with GATE. The system was then tested over the test corpus of 477 discharge summaries. We found that our NLP-SNOMED system achieved the results of 39% for the first data set and 49% for the second data set in annotating both the SNOMED-CT concepts and the noun phrases.*

Introduction

The need to have quality electronic medical records (EMR) is crucial in the health care sector. Major investments in EMR are seen as a key driver for improving health care in most developed countries including the USA[1]. The vital information within EMRs is shared between health professionals and takes many forms, e.g. discharge summaries, progress summaries and patient's consultation notes. The EMR supports clinical and administrative users with accessibility, alerts, reminders and clinical decision support. The ongoing challenge is that the information contained in EMR's is often in an unstructured format with large amounts of free text written in natural language by the clinician. Although there has been much debate about introducing more structure into EMR current thinking is that this approach is short sighted and unlikely to be of benefit to either the patient or the physician [2,3]. Primarily because clinical medicine is so diverse and constantly evolving, there is a need to accept natural language and free text within EMR and clinical documentation. By developing new NLP systems, researchers seek to identify and extract the clinical information within EMRs from this narrative text and data standards will provide consistency in representation and coding of the data [4].

The fourth i2b2/VA Shared Task and Workshop Challenges in Natural Language Processing for Clinical Data presented an opportunity to work with a large corpus of discharge summaries anonymised for research using Natural Language Processing (NLP) techniques. Natural language approaches evolved simple parsing of medical text[5].With the development of MEDLEE by Friedman which used clinical ontologies to code information[6,7,8]. Long extracted the diseases and procedures by applying NLP to extract the diagnosis from discharge summaries [9]. HITEx - Health Information Extraction tool extract the concepts from the principal diagnosis using UMLS (Unified Medical Language System), a metathesaurus which includes SNOMED-CT [10] AMBIT: Acquiring Medical and Biological Information from Text [11], and MetaMap [12] for different evaluation in biomedical field.

Initial aim of this research was not specifically to win the contest but, to explore the use of SNOMED CT as concepts, in the analysis of patient data from disparate data sources. For this purpose, we have corpus of clinical documents;

1. Verbal Autopsies from Ghana
2. Data from an EMR describing a patient consultation recorded by a large number (n=400) medical students of University of Leeds.

The NLP-SNOMED approach being developed at Leeds is intended to be broadly applicable to research and learning within the field of medical informatics by using GATE – General Architecture for Text Engineering [14] and SNOMED-CT, an international multi-lingual health terminology developed by The College of American Pathologists and the United Kingdom's National Health Service [13]

Data and Methods

Two Datasets were released by the i2b2 organizers. Both contained discharge summaries which had been anonymised and approved for research and were from several sources.

1. Data set-1 was released at the start of the challenge. This data set contained patient's progress notes and discharge summaries from different healthcare partners. Data Set 1 containing the

discharge summaries have been selected from "unannotated" folder. This data set does not have any ground truth files available (concepts annotated manually for evaluation), so we have selected to develop rule based system.
2. Data set-2 was released after the development of the systems to test and, contained a further 400 discharge summaries.

We have developed a rule based system NLP-SNOMED to annotate the key medical concepts for the two sets of discharge summaries. This paper presents the system and evaluates its performance in identifying both SNOMED-CT concepts and noun phrases. The application annotates both the noun phrases and SNOMED CT terms as "concepts", found in the corpus.

NLP-SNOMED Rule-Based Approach:

This NLP tool is based on a pipeline architecture common in NLP tools and a key feature of the GATE tool. The application pipeline starts with pre-processing of the core concept file supplied and is followed by building a semantic annotation pipeline to undertake language processing through the use of GATE - General Architecture for Text Engineering [14]. The following sections describe the architecture in detail and this is summarized below in (Figure 1).

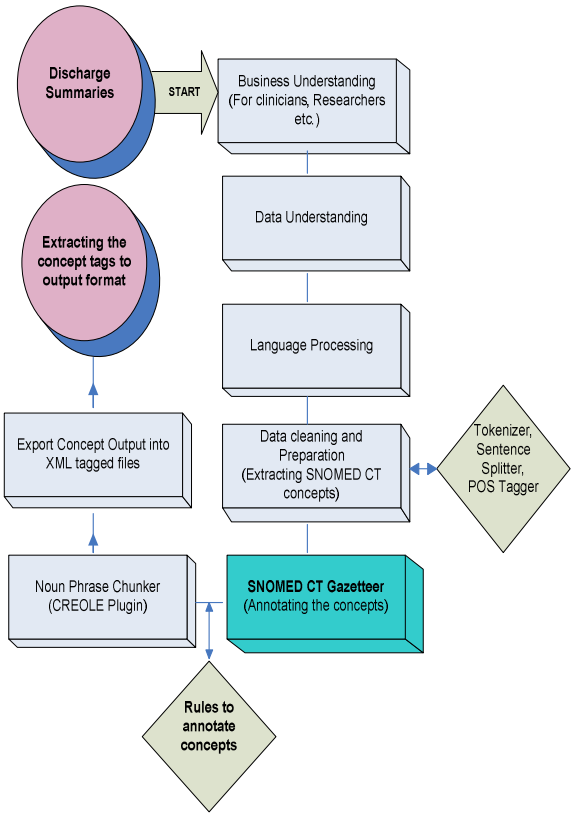

Figure 1. NLP-SNOMED system flow

Business Understanding

This research aims to extract SNOMED-CT concepts within the supplied discharge summaries to identify and document the issues related to natural language when written by clinicians and to explore the associated data standard and terminologies. This application was produced following the guidelines provided by i2b2 organizers for the concept extraction phase. The requirement was to extract the concepts in doing so have the ability to classify the medical problems, test and treatments. As part of the process completed noun phrases and adjective phrases were also annotated.

Data Understanding

From the data set we selected the Data Set 1 containing the discharge summaries and omitted the progress notes. When the data was first observed, it was established that it contained "structural information." The noisy data consisted of different header sections in different formats such as "Admission Date", "Social History", "Past Medical History", "PRINCIPAL DIAGNOSIS" etc. The Data set 1 and Data set 2 corpus had different header format; the headers in training set were in "UPPER CASE" and the test set in "Title Case". The common reference of each document was the punctuation mark ":" after each heading. Therefore a rule was written and applied to identify each heading by the punctuation mark ":", within each document. This formed the basis to be able to classify the medical concepts into problem, test and treatment.

Language Processing

The English Tokeniser was used first. The Token splits allow for the application to apply rules for annotating the concepts. The Sentence splitter was used to analyze the boundaries, in order to calculate the line offsets afterwards. Finally the application used a part of speech tagger which uses Penn Treebank tag set. This tagger has been adopted and has been used in the analysis of medical notes [15].

Data Cleaning and Preparation

The SNOMED-CT core concept file was selected from the UMLs source and consisted of the following attributes; CONCEPT ID, CONCEPT STATUS, FULLY SPECIFIED NAME, CTV3ID, SNOMED ID, IS PRIMITIVE as shown in (Figure 2). Only the "FULLY SPECIFIED NAME" was required. Therefore, a Python program was built to remove all attributes except the "FULL SPECIFIED NAME". By achieving this a file was produced to enable the build of the gazetteer in order to create a lookup for the concept names.

```
-
100334005  10   DERMOLAR SHAMPOO (product) XU05qC-D2417   1
100361005  10   DIFIL SYRUP (product) XU06KC-D2499   1
100362003  10   DIFIL TAB8 (product)  XU06LC-D2501   1
100390004  10   DL-ALPHA TOCOPHEROL ACETATE INJECTION
(product)  XU06pC-D2569   1
10039002   0    ^210m^Bismuth (substance)  XU06qC-125B2   1
100391000  10   D-LIMONENE SHAMPOO (product)    XU06rC-D2571
     1
100420006  10   DUOVAC -Ma5 (product) XU07LC-D2633   1
10042008   0    Structure of intervertebral foramen of fifth
thoracic vertebra (body structure)   XU07MT-1175A   1
100335006  10   DEXAMETHASONE 2.0 MG INJECTION (product)
     XU05rC-D2421   1
100336007  10   DEXAMETHASONE INJECTION (product)    XU05sC-
D24231
100363008  10   DINEOTEX (product)   XU06MC-D2503   1
```

Figure 2. Extract from SNOMED-CT core concept file.

Developing the Gazetteer for SNOMED-CT Concept Annotations

To annotate the SNOMED-CT concepts in the corpus a gazetteer list was developed of all the concept names by using GATE. (Figure 3) shows the gazetteer entries after cleaning the SNOMED-CT concept file.

```
Abdominal rigidity
Abdominal rigidity absent
Abdominal rigidity of epigastrium
Abdominal rigidity of left lower quadrant
Abdominal rigidity of left upper quadrant
Abdominal rigidity of periumbilical region
Abdominal rigidity of right lower quadrant
Abdominal sacrocolpopexy
Abdominal seizure
Abdominal skin crease
Abdominal skin fold thickness
Abdominal skin pouch structure
Abdominal skin ptosis
Abdominal skin scar
Abdominal somatic dysfunction
Abdominal stoma
```

Figure 3. Extracted from SNOMED-CT core concept gazetteer.

When a match was obtained between the gazetteer and the natural language written in the discharge summary JAPE-Java Annotation Pattern Engine [16] was used to write a Concept Rule to annotate the concepts from SNOMED-CT list as "concept".

SNOMED-CT concepts → concept

NP-Noun Phrase Chunker

After filtering the SNOMED-CT concepts in the corpus, a noun phrase chunker was used sourced from the CREOLE plugin within GATE to find noun phrases present in each discharge summary in the corpus. The aim was to annotate the remaining concepts which are complete noun phrases.

Rule#1: To annotate every NOUN PHRASE as concept

After annotating noun phrases, this rule annotated the concepts overlapping with the noun phrases, which were not captured by the SNOMED-CT concept gazetteer. Now the concept annotation appeared as follows;

Concept = SNOMED-CT concepts + Noun Phrases

This rule refined the application into a more specific approach that analyzed the relation of each SNOMED-CT concept containing long phrases with the existing noun phrases. The rule excluded the SNOMED-CT concepts which were noun phrases.

Rule#2: Annotate Headers to filter the concepts

Discharge summaries contains heading sections like; Primary Diagnosis, Secondary Diagnosis, History of Present Illness, Past Medical History, Social History etc. Some of these headings have been annotated by gazetteer automatically. This required correction as these headings should not be annotated as "concepts". This was overcome by applying a rule using JAPE first by annotating these headers as "Heading" and then to exclude these headings from concept annotations. The Headings were identified and then marked by using the ":" symbol followed by the headings. The headers were marked in order to extract the type of concepts.

Rule#3: For extracting the offsets of concepts

The i2b2 Organizers provided the guidelines to mark the concepts with the offset. Offsets were marked as stated in the guidelines.

Start offset = Line Number : word number
End offset = Line Number : word number

The output produced by the application needed post processing to extract the offsets into the required format provided by i2b2 organizers but the application is showing concepts offset in GATE[14].

Exporting the Concepts into Output Format:
GATE provided the visual resource to view and output the annotations along with the annotation lists

providing all the details for each annotation. Another plugin from CREOLE (Flexible Exporter) was used to export the annotation into XML and a python program was then used to convert the XML into the required format by i2b2 organizers;

c= "concept text" offset ¦¦ t= "concept type"

GATE[14] showed the output in visual format and to extract the annotated concepts, we converted the output files into XML tagged format. These output files contains concept tags within the text of file. To extract the concepts from the XML files, post processing have been done using Python script. The final output generated by ESNOMED application, contained only concepts but not the type of concepts as shown in (Figure 4).

```
c="Propionibacterium acnes"
c="Physical"
c="infectious disease"
c="blood culture"
c="extremity"
c="deep "
c="fibrillation"
c="complaint"
c="acute"
c="prednisone"
c="pitting edema"
c="potassium diet"
c="myocardial "
c="frequent"
c="cardiovascular disease"
```

Figure 4. Sample Output extracted from tagged files produced by GATE.

Evaluation:

The evaluation phase was based on finding the frequencies of concepts present in the corpus and the SNOMED-CT concepts as well. Table 1 shows the frequencies of all the annotations analyzed by NLP-SNOMED. On observing the annotations, concept annotations with noun phrases produced better results than extracting SNOMED-CT concepts separately. The problem is natural language written by clinician is similar but not exactly written followed by data standard like SNOMED-CT. These results do not meet the requirements of the contest because SNOMED CT gazetteer aimed to extract only the concept terms and not the type of the concepts.

Annotations	**Data set 1**		**Data set 2**	
Concepts (SNOMED-CT+Noun Phrases)	39%	92782	49%	206355
SNOMED-CT Concepts	24.41 %	58084	27%	129674
Noun Phrases	22.8 %	54269	25.6 %	122667

Table 1. Frequencies of SNOMED-CT concepts and Noun phrases in the Data set 1 and Data set2.

So, if any evaluation script aimed to measure the format given by the organizers cannot work on output extracted from N:LP-SNOMED application.

Discussion

Discharge summaries often contain long but important phrases For example; "Acute duodenal ulcer with hemorrhage AND with perforation but without obstruction". With SNOMED_CT having the benefit of being a nomenclature it has the ability to identify this. SNOMED-CT does appear to have finer concept granularity and a richer expressiveness than other classification systems. SNOMED-CT concepts have some concepts which have roots in some other languages such as French and German ("Familial multiple café-au-lait macules without neurofibromatosis" or "WaldenstrÖm macroglobulinemia"). The characters "é" and "Ö" are not readily accessible on a standard English keyboard used by a clinician to input natural language into an EMR or a discharge summary. There would therefore be no mat One solution is to convert these terms into English names to increase the chance of matching natural language to the SNOMED-CT concepts.

Another important concern is the familiarity of the clinician with the terminologies based on standard like SNOMED-CT. It seems unreasonable so they could write the terms according to the data standards and terminology and then to expect that clinicians could write the terms according to such a large and complex set of data standards and terminology in order to make subsequent automatic extraction more efficient.

Conclusion

For the SNOMED CT concept extraction, the use of the gazetteer provided by GATE worked well enough, especially in capturing the long multiword concepts. Moreover the use of a coding standard provides data quality and security in sharing the patient's information. Natural Language Processing helped effectively in extracting some valuable

information from clinical narratives. Our aim is to continue to improve on the current limitations of our solution by refining the rule- based approach with more sophisticated language processing to improve ton the accurate annotation of the correct concepts.

Acknowledgements

The 2010 i2b2/VA challenge and the workshop are funded in part by the grant number U54-LM008748 on Informatics for Integrating Biology to the Bedside from National Library of Medicine. This challenge and workshop are also supported by resources and facilities of the VA Salt Lake City Health Care System with funding support from the Consortium for Healthcare Informatics Research (CHIR), VA HSR HIR 08-374 and the VA Informatics and Computing Infrastructure (VINCI), VA HSR HIR 08-204. MedQuist, the largest transcription technology and services vendor, generously co-sponsored the 2010 i2b2/VA challenge meeting at AMIA. Authors would also like to acknowledge Niraj Aswani from GATE team and Abdul Baqi Sharaf from University of Leeds for their guidance.

References

1. Biden (2009), Press Release from US Vice President, J Biden, White House http://www.whitehouse.gov/the-press-ffice/vice-president-biden-announces-availability-nearly 12-billion-grants-help-hospitals accessed online 30th August 2010.
2. Wears RL, Berg M, Computer technology and clinical works: still wait for Godot, JAMA 293 (March 9 (10)), 2005, pp1261-1263.
3. Thompson DA, Eitel D, Fernandes CM, Pines JM, Amsterdam J, Davidson SJ, Coded chief complaints – automated analysis of free text complaint. Acad.Emerg.Med. 13 (7), 2006 pp 774-782.
4. Shahpori R, Doig C, "Systemized Nomenclature of Medicine-Clinical Terms direction and its implications on critical care", *Journal of Critical Care,* 2009.
5. Lussier YA, Bodenreider, O. 2007.Clinical Ontologies for discovery applications. In: BAKER CJO, CHEUNG KH, ed. 2007.Semantic Web: Revolutionizing knowledge discovery in the life sciences: Springer pp. 101-119.
6. Bodenreider, O, Mitchell JA, McCray, AT. 2002. Evaluation of the UMLS as a terminology and knowledge resource for biomedical informatics. *Proc AMIA Symp*. pp.61-5.
7. Morrison, FP, Li Li, MS, Lai, A et al. 2009. Repurposing the Clinical Record: Can an Existing Natural Language Processing System De-identify Clinical Notes? *Am Med Inform Assoc.* 16(1) pp. 37–39.
8. Friedman C, Shagina L, Lussier Y, et al. 2004. Automated encoding of clinical documents based on natural language processing. *J Am Med Inform Assoc*. **11**(5), pp.392-402.
9. Long W, Extracting diagnosis from discharge summaries", *AMIA Annu Symp,* 2005, pp. 470–474.
10. Qing T Zeng**,** Sergey Goryachev**,** Scott Weiss**,** Margarita Sordo**,** Shawn N Murphy and Ross Lazarus, "Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system", *BMC Medical Informatics and Decision Making,* 2006, 6:30.
11. Gaizauskas R, Hepple M, Davis N, Guo Y, Harkema H, Roberts A, and Roberts I, "AMBIT: Acquiring Medical and Biological Information from Text", *ISMB/ECCB*, Poster, 2004.
12. Aronson AR, "Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program", *Proc AMIA Symp,* 2001, pp. 17–21.
13. College of American Pathologists. http://www.cap.org
14. Cunningham H, "GATE, A General Architecture for Text Engineering", *Computers and the Humanities*, 2002, Vol. 36, pp. 223-254
15. Pakhomov SV, Coden A, CG. Chute, "Developing a corpus of clinical notes manually annotated for part-of-speech", Int J Med Inform, 2006, 75(6), pp. 418-429.
16. Cunningham H, Maynard D, Tablan V, "JAPE: a Java Annotation Patterns Engine", *Research Memorandum,* dcs.shef.ac.uk, 2000, pp. 10-19.