Automatic Extraction and Mapping of Discharge Summary's Concepts into SNOMED CT

Rabia Batool[1], Asad Masood Khattak[2], Tae-Seong Kim[1] and Sungyoung Lee[2]

Abstract— **Patient data is critical in healthcare domain. Secure, consistent, and coded information increases the efficiency and encourages collaboration within and between organizations. Data entered by physician consists of free text containing medical terminologies. Processing of such data in e-health or clinical decision support systems is a challenging task. This paper presents a mapping and transformation system of discharge summaries of hospital written in natural language to standardized data, which can be easily used by computer based medical applications. We process 889 free text discharge summaries and map medical concepts to SNOMED CT (Systematized Nomenclature Of Medicine Clinical Terms) concepts using natural language processing techniques. Proposed system uses stemming, reordering, synonyms, indirect mapping, and stopword removal to increase accuracy of conversion.**

Index Terms— **SNOMED CT; Natural language processing; Discharge Summaries; Healthcare;**

I. INTRODUCTION

Quality of patient care depends upon the collection and dissemination of information about patient [1]. Health data is critical and needs to be stored and retrieved in a structured form. Healthcare standards exist at multiple levels. High level standards are defined which determine the structure of the medical knowledge and provide information model for medical records [1]. They help in information exchange among different healthcare systems. Low level standards are the vocabulary used to describe the clinical encounters [1].

Moving patient's records from paper or physical filing systems to standardized computer based system creates utility for patients, providers, and decision support systems [2]. It also plays very important role in communicating information within different healthcare providers or stakeholders [3]. Usually the physicians store data in fragmented English free text along with medical terminologies. Almost 50% of the data in patients medical records is physicians free text notes [4]. The attempt to change the habit of documentation of physicians has failed due to increase in time and inconvenience caused by using formatted computer interface [3]. The processing of such free text in discharge summaries is not only hard for searching and retrieval, but is also inconvenient for e-health and clinical decision support systems.

[1]R. Batool and T.-S. Kim are with the Department of Biomedical Engineering, Kyung Hee University, Seocheon-dong, Giheung-gu, Yongin-si, Gyeonggi-do, Republic of Korea rabia@oslab.khu.ac.kr, tskim@khu.ac.kr

[2]A. M. Khattak and S. Y. Lee are with the department of Computer Engineering, Kyung Hee University, Seocheon-dong, Giheung-gu, Yongin-si, Gyeonggi-do, Republic of Korea asad.masood@oslab.khu.ac.kr, sylee@oslab.khu.ac.kr

In this research paper, we propose a system to automatically extract medical terminologies from discharge summaries and convert them to SNOMED CT codes respectively. SNOMED CT is standardized, comprehensive, and multilingual clinical reference terminology that consists of millions of medical concepts with codes. It provides effective access to information required for decision support, consistent reporting, and analysis [5]. This terminology provides disparate systems throughout the healthcare providers to manage information in structured form with consistency [6]. We also explore different ways to increase accuracy of the proposed system that includes stemming, synonyms, reordering, and indirect mapping (from FMA to SNOMED). The Foundational Model of Anatomy Ontology (FMA) is an evolving domain ontology that represents knowledge about human anatomy which is understandable by human and machine-based systems. Its ontological framework can be applied and extended to all other species [7].

The proposed system processes 889 free text discharge summaries to extract medical concepts and convert them into standardized codes. The quality of conversion increases by applying the proposed series of natural language processing techniques.

Rest of the paper is organized as follow. Section II discusses the related work closely aligned with our work. Section III describes the proposed system architecture and its components. Section IV describes data set used for implementation, testing, and results of the proposed system. Section V concludes the work and explores future research directions.

II. RELATED WORK

Several research groups are working on natural language processing in biomedical informatics. They have developed applications to process clinical text of certain type. Haug et al. [8] developed automated admit diagnosis encoding system to decrease time delay involved in waiting for human coder. System is able to convert reasons for admitting patient from free text statement into more than 450 ICD diagnoses. They divided the concepts into two groups. First group can be entered into information system based on probabilistic approach and previous successful encoding without reviewing from HIS personnel. Second group consists of concepts which required user review before storing into information system. They did not consider spelling mistakes in free text. Christensen et al. [9] introduced a system which uses Bayesian network to represent information in medical domain. They used directed acyclic graph to represent different

concepts and their relations. Bottom-up chart parser is used to extract whole concept from text.

Hina et al. [10] built a system to extract information from clinical documents using SNOMED CT concepts. They applied natural language processing techniques using GATE to extract valuable information from discharge summaries. SNOMED CT gazetteer was used to provide standardization to noun phrases, extracted from clinical discharge summaries. Long et al. [11] used natural language processing techniques to find disease and procedures in discharge summaries. They used unified medical language system (UMLS) to convert disease and procedure into standard vocabulary. Patrick et al. [12] processed clinical notes by variety of natural language processing and extracted the complete subset of SNOMED CT that would be necessary for an intensive care unit by using subtree of concept types in SNOMED hierarchy. Different tools are also available to extract medical information from free text. Some of them are Medlee [13], AMBIT [14], and HITEx: Health Information Text Extraction [15]. In continuation to the existing research, we apply natural language processing techniques on discharge summaries to extract medical information and focus on accuracy of mapping by integrating stemming, synonyms, reordering, and indirect mapping. The proposed system is using SNOMED CT for standardization of extracted medical information.

III. Proposed system architecture

In this section, we present our proposed system for extraction and conversion of medical concepts from discharge summaries to SNOMED CT concepts. To extract concepts from discharge summaries, and to map them to SNOMED CT codes, the proposed architecture is divided into two main components. Overall architecture is shown in Fig. 1 and the detail of each component is given below.

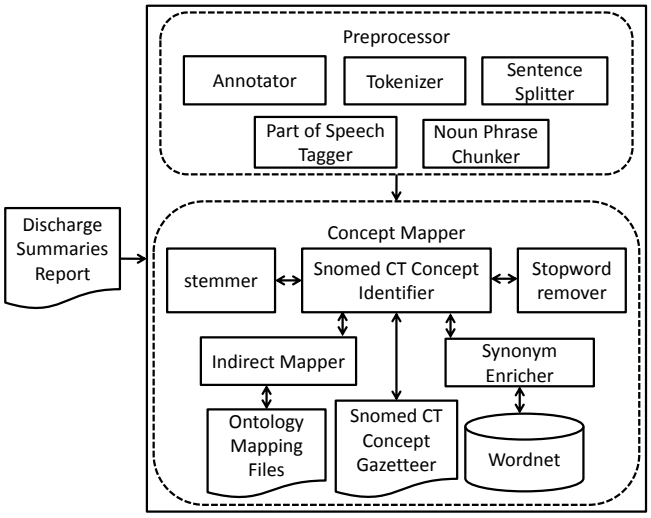

Fig. 1. Detailed Architecture of Proposed system

```
Discharge Status :
Discharged

Condition on Discharge :
Stable

Patient States Complaint :
N V D

Diagnosis :
Vertigo , resolved

Treatment Rendered :
head MRI showed no evidence of stroke or mass , but did show
severe stenosis of the left internal carotid artery
Discharge Medications :
same as admission

Disposition , Follow up and Instructions to Patient :
We are not sure exactly what was causing your vertigo and nausea,
but you have improved significantly , and a head MRI showed no
acute change to suggest a cause of your symptoms .
The MRI did show a blockage in the L carotid artery , which
should be examined further .
Please call Dr office on Monday and arrange to see him within a
week .
```

Fig. 2. Example of portion from discharge summaries

A. Preprocessor

Preprocessor is used to extract medical concepts from discharge summaries. Proposed system uses General Architecture for Text Engineering (GATE) to preprocess discharge summaries. GATE is an open source software capable of solving text processing problems. It is in use for all types of computational task involving human language [16]. The *Annotator* resets previous annotation made by the doctors in discharge summaries. *Tokenizer* converts discharge summaries into tokens. Tokens are words, symbols, and other meaningful elements in the text. *Sentence Splitter* is used to split text into sentences on the basis of tokens, and *Part of Speech Tagger* is used to identify different parts of speech in discharge summaries that are required by *Noun Phrase Chunker*. The *Part of Speech Tagger* is a modified version of Brill's speech tagger [17]. *Noun Phrase Chunker* extracts noun phrases to map with SNOMED CT concepts. This *Noun Phrase Chunker* is a java implementation of the Ramshaw and Marcus Base NPchunker [18]. After the execution of all subcomponents on discharge summaries, we get noun phrases. We have extracted 147211 noun phrases from 889 discharge summaries using GATE.

B. Concept Mapper

After preprocessing, we need to map single and multi word concepts to standardized SNOMED CT concepts. Following the extraction of noun phrases, *Preprocessor* passes the noun phrases to *Concept Mapper*. *SNOMED CT Concepts Identifier* uses different modules to map noun phrases with SNOMED CT concepts. Stemming is the process of reducing inflected or derived words to their stem, base, or root form. If noun phrase with original word is not found in SNOMED CT concepts, *Stemmer* is used to apply stemming to words. *Indirect Mapper* uses FMA ontology to map noun phrase to SNOMED CT indirectly. Concepts that are not present in SNOMED CT and present in FMA can be extracted using FMA to SNOMED CT mapping files. Stopwords are words

that are filtered out prior to processing of natural language data. *Stopword Remover* checks for list of words in noun phrases and removes them if found to map with concepts accurately. WordNet is a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms in Wordnet and it also provides short and general definition [19]. *Synonym Enricher* uses Wordnet to find synonyms and definitions of each word in noun phrase if original noun phrase is not present in SNOMED CT. After applying series of modules on extracted noun phrases, *SNOMED CT Concepts Identifier* is able to convert noun phrases to SNOMED CT codes more accurately.

IV. Evaluation

A. Data set

The data set used for testing of system contains discharge summaries from Informatics for Integrating Biology and the Bedside (i2b2). There are total 889 discharge summaries with 147211 noun phrases. To standardize natural language used by physicians in discharge summaries, SNOMED Clinical Terms (SNOMED CT) is used. We used SNOMED data source from UMLS which has 341659 unique concepts with their semantics. We also used the Foundational Model of Anatomy ontology (FMA) for indirect mapping of discharge summaries noun phrases to SNOMED CT concepts. OAEI - Ontology Alignment Evaluation Initiative is an initiative that has provided benchmark ontologies and their mappings for testing of ontology matching systems [20]. One of the tracks is related to large biomedical ontologies thats consists of SNOMED, FMA and NCI ontologies [21]. We used those mapping files for indirect mapping from FMA to SNOMED CT. Wordnet is used to extract meanings and definitions of each concept and word, so if there is any variation in choices of words, it can be handled using synonyms and word definition from wordnet.

B. Results

We present ongoing research to convert noun phrases from discharge summaries to SNOMED CT concepts. Fig.2 shows example of portion from discharge summary. Extracted noun phrases from discharge summary shown in Fig.2, their equivalent SNOMED CT concept and codes are shown in Table I. We have been testing our system with 889 discharge summaries. The proposed system applies stemming, synonyms, indirect mapping and stopword removal to covert medical concept to SNOMED CT code. Table I also shows more concepts, mapped to SNOMED codes successfully by proposed system after applying series of modules.

- Most of the concepts exactly match and don't require further processing e.g. "Stroke", "Left internal carotid artery", "Vertigo".
- There are some concepts that are not present in SNOMED CT ontology but present in FMA. OAEI 2012 mapping files are used to map concept which are not found in SNOMED CT concepts and exist in FMA. Like "Right axilla" and "Manubrium" are not present in SNOMED CT concept, but they are present in FMA ontology. FMA to SNOMED mapping contains "Right axilla" and "Manubrium" concepts of FMA maps to "Structure of right axillary region" and "Bone structure of manubrium" concept of SNOMED CT respectively as shown is Figure 3.
- Stopword removal also helps in mapping. Proposed system removes some words from concepts like 'of', and 'to'. This contributes to an increased accuracy of the proposed system. We also found that body directions "right", and "left" is missing for some concepts in SNOMED CT which exists in discharge summaries like "Right thalamic hemorrhage" found in discharge summaries; however, SNOMED CT has "Thalamic hemorrhage". Some adjectives were also present in discharge summaries like "severe stenosis" correctly maps to "Stenosis" by removing adjectives, physician used to describe severity of disease. So in such cases the proposed system first checks full phrases for mapping, if not found then system removes stopwords from noun phrases to map to SNOMED CT concept.
- Synonyms and definitions of words from Wordnet improve accuracy of the results e.g., "vulvar cancer" maps to "Cancer of vulva" which cannot be directly matched. Through Wordnet, the proposed system found "vulvar is relating to the vulva".
- Some concepts are present but order of words in concepts is changed like "Small cell lung carcinoma" is present as "Small cell carcinoma of lung" in SNOMED CT. The proposed system maps these concepts by removing stopword first then ignoring word order.

TABLE I
Medical concept from discharge summaries and their standardized codes mapped by proposed system

Medical concept from discharge summaries	SNOMED CT concept	Code
Vertigo	Vertigo	399153001
Head MRI	MRI of head	241601008
GI bleed	GI bleeding	74474003
Severe stenosis	Stenosis	415582006
VULVAR cancer	Cancer of vulva	363367000
Metastatic liver	Metastasis to liver	94381002
Right thalamic hemorrhage	Thalamic hemorrhage	230711001
Carcinoma of the kidney	Carcinoma of kidney	254915003
Right axilla	Structure of right axillary region	19654004
Manubrium	Bone structure of manubrium	37285002
Left internal carotid artery	Left internal carotid artery	58379002

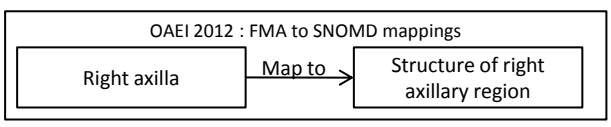

Fig. 3. FMA to SNOMED mapping extracted from OAEI's track related to large biomedical ontologies

- *Stemmer* also improves the proposed system performance. Some words require stemming like "GI bleed", and "Pneumonias" map to "GI bleeding" and "Pneumonia" respectively by stemming concepts from discharge summaries.
- After noun phrase chunking, proposed system also considers combination of more than one noun phrase from same sentence which are connected by preposition to make large concept, e.g., "carcinome of the kidney"

To measure accuracy of our proposed system, we analyzed 250 random concepts using UMLS SNOMED CT browser. Majority of the concepts were found, but still some noun phrases like "Post-infarction unstable angina pectoris" were not mapped to any SNOMED CT concept due to absence of corresponding concepts in SNOMED CT. The proposed system is not able find their equivalent SNOMED CT code.

V. CONCLUSION AND FUTURE WORK

In this research work, we have demonstrated a system to extract and map medical concepts from discharge summaries to SNOMED CT concepts. To avoid information loss our proposed system applies series of natural language processing techniques to convert noun phrases from discharge summaries into SNOMED CT codes. Integration of different techniques has improved the results as shown in Table I. The proposed system helps in processing of natural language to make it usable for healthcare applications like clinical decision support system and making system interoperable. There exist some noun phrases which are not mapped to any SNOMED CT concept. Domain expert analysis is required to find best alternative for those noun phrases which can be mapped to SNOMED CT concepts. Spelling mistakes are very common in free text; however, we did not consider this problem in this paper. Future work includes spelling mistakes recovery, word disambiguate problem, and detailed semantic analysis to extract relations of concepts.

ACKNOWLEDGMENT

This research was supported by the MKE(The Ministry of Knowledge Economy), Korea, under the ITRC(Information Technology Research Center) support program supervised by the NIPA(National IT Industry Promotion Agency) (NIPA-2012-(H0301-12-2001)).

This work was supported a grant from the NIPA(National IT Industry Promotion Agency) in 2012. (Global IT Talents Program).

REFERENCES

[1] W. Chan, C. Centiu, J. Morris, and M. Kurtz, "Uniform data standards for capturing patient medical record information at the point of care," *Journal of Healthcare Information Management*, vol. 13, pp. 85–96, 1999.

[2] T. Torrey, "The benefits of electronic medical records," (Last visited in january 2013). [Online]. Available: http://patients.about.com/od/electronicpatientrecords/a/ EMR-benefits.htm

[3] D. Heinze, M. Morsch, and J. Holbrook, "Mining free-text medical records." in *Proceedings of AMIA Symposium*. American Medical Informatics Association, 2001, p. 254.

[4] A. Turchin, N. Kolatkar, R. Grant, E. Makhni, M. Pendergrass, and J. Einbinder, "Using regular expressions to abstract blood pressure and treatment intensification information from the text of physician notes," *Journal of the American Medical Informatics Association*, vol. 13, no. 6, pp. 691–695, 2006.

[5] "Snomed," (Last visited in March 2012). [Online]. Available: www.ihtsdo.org/snomed-ct

[6] H. Wasserman and J. Wang, "An applied evaluation of snomed ct as a clinical vocabulary for the computerized diagnosis and problem list," in *AMIA Annual Symposium Proceedings*, vol. 2003. American Medical Informatics Association, 2003, p. 699.

[7] "Foundational model of anatomy," (Last visited in March 2013). [Online]. Available: http://sig.biostr.washington.edu/projects/fm/AboutFM.html

[8] P. Haug, L. Christensen, M. Gundersen, B. Clemons, S. Koehler, and K. Bauer, "A natural language parsing system for encoding admitting diagnoses." in *Proceedings of the AMIA Annual Fall Symposium*. American Medical Informatics Association, 1997, p. 814.

[9] L. Christensen, P. Haug, and M. Fiszman, "Mplus: a probabilistic medical language understanding system," in *Proceedings of the ACL-02 workshop on Natural language processing in the biomedical domain-Volume 3*. Association for Computational Linguistics, 2002, pp. 29–36.

[10] S. Hina, E. Atwell, and O. Johnson, "Secure information extraction from clinical documents using snomed ct gazetteer and natural language processing," in *Internet Technology and Secured Transactions (ICITST), 2010 International Conference for*. IEEE, 2010, pp. 1–5.

[11] W. Long, "Extracting diagnoses from discharge summaries," in *AMIA Annual Symposium Proceedings*, vol. 2005. American Medical Informatics Association, 2005, p. 470.

[12] J. Patrick, Y. Wang, P. Budd, A. Rector, S. Brandt, J. Rogers, R. Herkes, A. Ryan, and B. Vazirnezhad, "Developing snomed ct subsets from clinical notes for intensive care service," in *7th Annual HINZ Conference and Exhibitions, Rotorua, New Zealand*, 2008.

[13] C. Friedman, L. Shagina, Y. Lussier, and G. Hripcsak, "Automated encoding of clinical documents based on natural language processing," *Journal of the American Medical Informatics Association*, vol. 11, no. 5, pp. 392–402, 2004.

[14] R. Gaizauskas, M. Hepple, N. Davis, Y. Guo, H. Harkema, A. Roberts, and I. Roberts, "Ambit: Acquiring medical and biological information from text," in *Proc UK e-Science programme All Hands Conference*, 2003, pp. 2–4.

[15] "Hitex manual v2.0," (Last visited in january 2013). [Online]. Available: https://www.i2b2.org/software/projects/hitex/hitex_manual.html

[16] H. Cunningham, "Gate, a general architecture for text engineering," *Computers and the Humanities*, vol. 36, no. 2, pp. 223–254, 2002.

[17] E. Brill, "A simple rule-based part of speech tagger," in *Proceedings of workshop on Speech and Natural Language*. Association for Computational Linguistics, 1992, pp. 112–116.

[18] L. Ramshaw and M. Marcus, "Text chunking using transformation-based learning," in *Proceedings of the Third ACL Workshop on Very Large Corpora*. Cambridge MA, USA, 1995, pp. 82–94.

[19] G. Miller *et al.*, "Wordnet: a lexical database for english," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[20] "Ontology alignment evaluation initiative," (Last visited in january 2013). [Online]. Available: http://oaei.ontologymatching.org/2012/

[21] "Oaei 2012 large biomed track," (Last visited in january 2013). [Online]. Available: http://www.cs.ox.ac.uk/isg/projects/SEALS/oaei/2012/