Original Article

Toward the Automatic Generation of the Entry Level CDA Documents

Sungwon Jung[1], Seunghee Kim[2], Sooyoung Yoo[2], Jinwook Choi[2]

Interdisciplinary Program, College of Engineering, Seoul National Univ.[1],
Dept. of Biomedical Engineering, College of Medicine, Seoul National Univ.[2]

Abstract

Objective: CDA (Clinical Document Architecture) is a markup standard for clinical document exchange. In order to increase the semantic interoperability of documents exchange, the clinical statements in the narrative blocks should be encoded with code values. Natural language processing (NLP) is required in order to transform the narrative blocks into the coded elements in the level 3 CDA documents. In this paper, we evaluate the accuracy of text mapping methods which are based on NLP. **Methods**: We analyzed about one thousand discharge summaries to know their characteristics and focused the syntactic patterns of the diagnostic sections in the discharge summaries. According to the patterns, different rules were applied for matching code values of Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT). **Results**: The accuracy of matching was evaluated using five-hundred discharge summaries. The precision was as follows: 86.5% for diagnosis, 61.8% for chief complaint, 62.7%, for problem list, and 64.8% for discharge medication. **Conclusion**: The text processing method based on the pattern analysis of a clinical statement can be effectively used for generating CDA entries. ***(Journal of Korean Society of Medical Informatics 15-1, 141-151, 2009)***

Key words: Clinical Document Architecture, Natural Language Processing, SNOMED CT

Received for review: July 3, 2008; **Accepted for publication:** February 9, 2009.

Corresponding Author: Jinwook Choi, Dept. of Biomedical Engineering, College of Medicine, Seoul National Univ.28 Yongon-Dong, Jongno-Gu Seoul, Korea, 110-744
Tel: +82-2-2072-1260, **Fax:** +82-2-745-7870, **E-mail:** jinchoi@snu.ac.kr

* This research was supported by the MKE (Ministry of Knowledge Economy), Korea, under the ITRC (Information Technology Research Center) support program supervised by the IITA (Institute of Information Technology Advancement)" (IITA-2008-(C1090-0801-0002)).
* This research was supported by the Seoul National University Hospital, Korea (21-2004-042).

Ⅰ. Introduction

These days, various kinds of Hospital Information Systems (HISs) are being developed in and outside of medical institutions with the development of medical computerization. Order communication System (OCS) and Picture Archiving and Communication System (PACS) were representative former HISs, but now, Electronic Medical Record (EMR) system is being established and used along with previous systems. Besides, unlike EMR which is limited to each hospital, researches about nationwide Electronic Health Record (EHR) are being actively conducted. As more hospitals and medical institutions adopt EMR system, more institutions will generate and store clinical documents electronically. This enables medical institutions to provide patients with a facilitated treatment and also to enhance the quality of service by way of sharing patients' medical treatment information. However, in order for a clinical document generated in a heterogeneous institution to be shared without distortion or loss in its meaning and content, interoperability of information should be achieved. To solve this problem, Health Level 7 (HL7) has established Clinical Document Architecture (CDA) standard, an XML-based document markup standard that specifies the structure and semantics of clinical documents[1)2)]. CDA functions as a useful and powerful standard for exchange of clinical documents among heterogeneous software systems. Currently, CDA Release 2 has been developed. CDA Release 1 became an American National Standard Institute (ANSI) approved standard in November, 2000 and CDA Release 2 became an ANSI-approved standard, in May, 2005. There are many studies actively ongoing about applying CDA domestically and internationally[3-9)].

A CDA document consists of a Header and a Body. The Header contains metadata needed for managing and searching documents and the Body defines the structure to represent actual clinical data. A structured Body may have many sections, and each section can contain a narrative block of a human readable narrative form and any number of encoded entries which is machine processable.

In order to increase semantic interoperability of a CDA document, CDA entry-level which is computer processable is required. However, much of the information is in textual form, which is not suitable for use by automated applications. An effective method that automatically maps text in clinical documents into standardized concepts would be needed to make CDA entries.

Many reports have been published on methods that automatically map clinical text to concepts within a standardized coding system. Some researches used methods based on string matching, statistical processing, or linguistic processing, utilizing part of speech tagging[10)11)]. The other work developed a coding method that uses advanced NLP techniques to generate structured encoded output consisting of findings and corresponding modifiers[12)].

In this paper, we describe three stages of text mapping process, Exact Mapping, Preprocessing Mapping, and Text Processing Mapping, to generate CDA entries automatically.

II. Materials and Methods

We analyzed narrative statements of Seoul National University Hospital(SNUH) discharge summaries to derive text mapping methods and evaluated each section's mapping performance, in order to generate an upgraded CDA entry from the previously developed Entry Mapper of CDA Studio®.

1. Entry Mapper of CDA Studio®

CDA Studio® is a tool to convert clinical documents stored in a legacy database into CDA documents, consisting of the Designer, the Mapper, the Generator, and the Entry Mapper (Fig. 1). In this paper, we will focus the Entry Mapper of CDA Studio®. Detailed description of other components is in[6)].

The Entry Mapper reads a CDA document and creates entry-level automatically. However, it had a very low entry mapping rate because mapping was based on Exact Mapping. For chief complaints, only 44 out of total 118 detail items(37.3%), for diagnosis, only 94 out of 186 detail items(50.5%), for problem list, 94 out of 295 detail items(31.9%), and for discharge medication, only 354 out

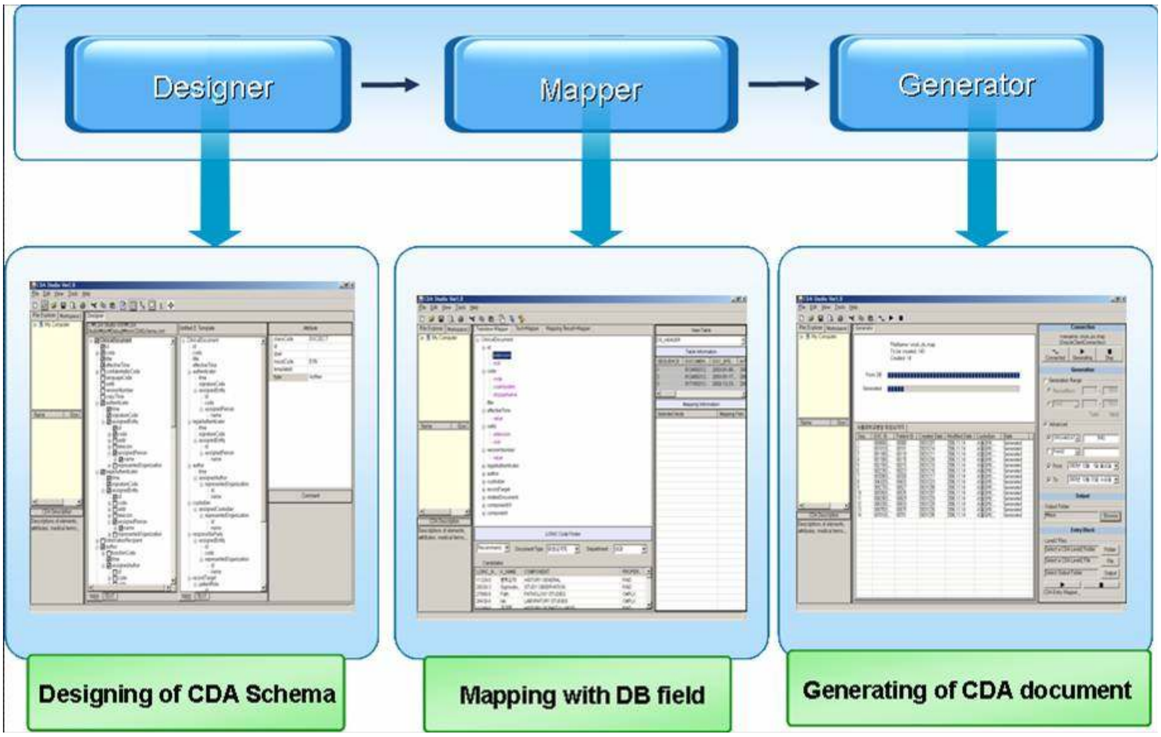


Figure 1. Main components of CDA Studio®

of 580 detail items(61.1%) were mapped. In case of discharge medication section, it showed relatively higher mapping rate than other sections, since mapping was conducted only with one-word ingredient.

2. SNUH discharge summaries

We analyzed 15,618 discharge summaries in 2003. They have about 30 types of data. Among those, data which are expressed as CDA Header in CDA document processing are as follows.

•Patient identification, sex, age, specialty, authenticator, primary physician, legal authenticator, author, composition date, recorded date.

The rest of the data can be expressed in CDA Body sections and are divided into three types, structured, semi-structured, and narrative according to the form of data content. Table 1 shows Section Name, Document Number and Section Type. Document Number is how many discharge summaries have each section.

Structured type has only one extracted detail item, not many detail items, therefore is described in a simpler way

Table 1. Characteristics of SNUH discharge summaries

Section Name	Document Number (%)	Section Format
Admission date	15,618(100.0)	Structured
Discharge date	15,618(100.0)	Structured
Ward	15,587(99.8)	Structured
Onset date	8,590(55.0)	Structured
Progress at admission	11,510(73.7)	Structured
Discharge disposition	10,933(70.0)	Structured
Keeping status	8,809(56.4)	Structured
Signature status	6,809(43.6)	Structured
Chief complaint	14,743(94.4)	Semi-structured
Diagnosis	13,541(86.7)	Semi-structured
Problem list	11,495(73.6)	Semi-structured
Discharge medication	12,135(77.7)	Semi-structured
Observation impression	11,026(70.6)	Narrative
Reservation	7,387(47.3)	Narrative
History of present illness	15,415(98.7)	Narrative
Physical examination	10,074(64.5)	Narrative
Progress after admission	14,447(92.5)	Narrative
Discharge instructions	2,140(13.7)	Narrative
Operation impression	1,577(10.1)	Narrative
Follow-up	2,015(12.9)	Narrative

than other sections. For instance, admission date, discharge date, and onset date are presented just as date like “2003-12-01,” ward is as “ward 114-internal,” and progress at admission and discharge disposition sections are just as “death, maintaining, improving, full recovery, transfer.” Keeping status and signature status sections are presented as “Y (yes)” or “N (no)” so that it easily generates entries.

Semi-Structured type presents data which can be subdivided into several detail items, so it needs the process of item detection. Most of the data, however, are in English and composed of relatively simple noun phrases, so it does not need additional natural language processing such as noun phrase extraction or part of speech tagging. For example, data in a diagnosis section, described in a document as “Typhoid Fever Due To Salmonella Typhi” needs to go through the process of dividing into “Typhoid Fever” and “Salmonella Typhi” but it does not need additional natural language processing like tagging. In case of the discharge medication section, as shown in the example of Figure 2, it needs the process of extracting the ingredient item from the sentence using Regular Expression (RE), since mapping to SNOMED CT is possible only with the ingredient of medication or product name.

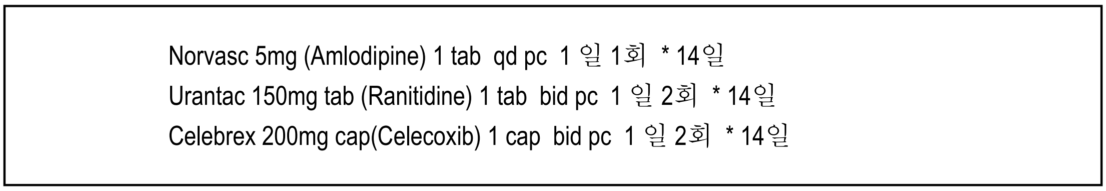
In narrative type of sections, one data has several freely described detail items, thus requiring extraction of candidate phrases through sentence identifying processing and natural language processing. Figure 3 is an example of the observation impression section, indicating that this section needs a method to figure out how to encode and which candidate phrase to encode, along with sentence identifying processing. Moreover, it requires a method of relating meanings of each sentence in case of subordinate sentence construction.

3. Analysis of diagnosis section

In order to know effective text mapping processes, we selected diagnosis section as representative section in discharge summary and mapped diagnosis to SNOMED CT manually. As the result of the manual mapping, 496 out of 1,000 diagnoses (49.6%) were mapped to SNOMED CT.

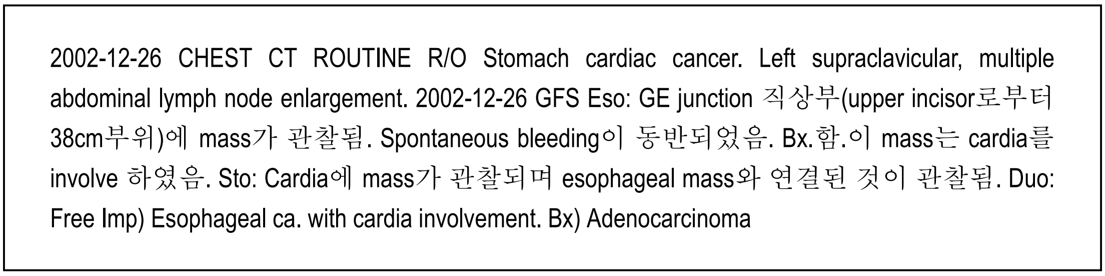
The remaining 504 diagnoses were mapped through following methods. First, some diagnoses were mapped by removing or separating modifiers. Modifiers generally appeared after a comma or a preposition. For example, for “Hepatocellular Carcinoma, s/p PET,” we selected the candidate phrase “Hepatocellular Carcinoma” by eliminating the comma and the following words. After the modifier elimination, the concept was successfully mapped to SNOMED CT.

Second, we separated qualifiers located at the beginning of diagnosis. “Chronic Renal Failure,” for example, was successfully mapped to SNOMED CT after splitting the diagnosis into a qualifier “Chronic” and a candidate concept



Norvasc 5mg (Amlodipine) 1 tab qd pc 1 일 1회 * 14일
Urantac 150mg tab (Ranitidine) 1 tab bid pc 1 일 2회 * 14일
Celebrex 200mg cap(Celecoxib) 1 cap bid pc 1 일 2회 * 14일

Figure 2. A sample example of a discharge medication section



2002-12-26 CHEST CT ROUTINE R/O Stomach cardiac cancer. Left supraclavicular, multiple abdominal lymph node enlargement. 2002-12-26 GFS Eso: GE junction 직상부(upper incisor로부터 38cm부위)에 mass가 관찰됨. Spontaneous bleeding이 동반되었음. Bx.함.이 mass는 cardia를 involve 하였음. Sto: Cardia에 mass가 관찰되며 esophageal mass와 연결된 것이 관찰됨. Duo: Free Imp) Esophageal ca. with cardia involvement. Bx) Adenocarcinoma

Figure 3. A sample example of an observation impression section

"Renal Failure."

Third, we improved the mapping rates by realigning diagnosis adding a preposition. "Liver Cirrhosis" was mapped to SNOMED CT when it was realigned as "Cirrhosis of Liver". Another diagnosis "Drug Eruption" was also mapped to SNOMED CT when it was realigned as "Eruption due to Drug."

Fourth, we tried to improve mapping result by expanding abbreviations into full names. "DM" was mapped to SNOMED CT when it was expanded to the full term as "Diabetes Mellitus."

Fifth, we removed some special characters. For example, "Hemolytic-Uremic Syndrome" was mapped to SNOMED CT successfully when it was changed to "Hemolytic Uremic Syndrome" by removing the special symbol "-".

Several additional methods were adopted for improving mapping performance. They were mainly done by using synonyms or combining above methods. For example, "Breast Cancer" was mapped to SNOMED CT when "Cancer" was replaced with its synonym "Carcinoma" and the diagnosis was realigned into a candidate concept "Carcinoma of Breast." Similarly, "Advanced Gastric Cancer" was mapped to SNOMED CT when the qualifier "Advanced" was separated, "Cancer" was replaced with its synonym "Carcinoma", "Gastric" was modified into its noun form "Stomach," and the diagnosis was realigned using a preposition into a candidate concept "Carcinoma of Stomach." Table 2 illustrates mapping rates to SNOMED CT of various mapping methods explained above.

Through these manual mapping processes, we derived three mapping steps, Exact Mapping, Preprocessing Mapping, and Text Processing Mapping, to improve mapping performance of previously developed Entry Mapper.

• Database structure for SNOMED CT

SNOMED CT is a comprehensive terminology to express the content of clinical documents and mainly used for encoding, searching and analyzing clinical data. SNOMED CT is composed of Concept ID (CID), Concept Status (CS), and Fully Specified Name (FSN).

In this study, we converted raw text files of SNOMED CT into Microsoft Access database to enhance the availability and efficiency of CID search (Fig 4). FSN, however, which is used for mapping of actual concepts, has semantic type of each medical term such as "disorder" and "substance" in "Chronic proctocolitis (disorder)" and "Bithionol (substance)." Therefore, we separated semantic types, created and added a new row labeled "TYPE." In order to increase efficiency of search for FSN, we also added a row, "Fully Specified Name Stemming (FSNS)" by stemming and removing special symbols and stop words. For instance, "Chronic proctocolitis" was added to FSNS as "chronic proctocol" by stemming and "Diabetes mellitus without complication" was added to FSNS as "diabet mellitus complic" after removing the stop word *without* and stemming.

Table 2. Results of manual mapping to SNOMED CT for diagnosis

Mapping method	Diagnosis number (%)
Exact mapping	496(49.6)
Remove words following a comma or a preposition	73(7.3)
Split the starting qualifier from the phrase	33(3.3)
Realign using a preposition	57(5.7)
Expand an abbreviation	39(3.9)
Remove a special symbol	2(0.2)
Etc. (modify POS; use a synonym; combine different methods)	300(30.0)
Total	1,000(100)

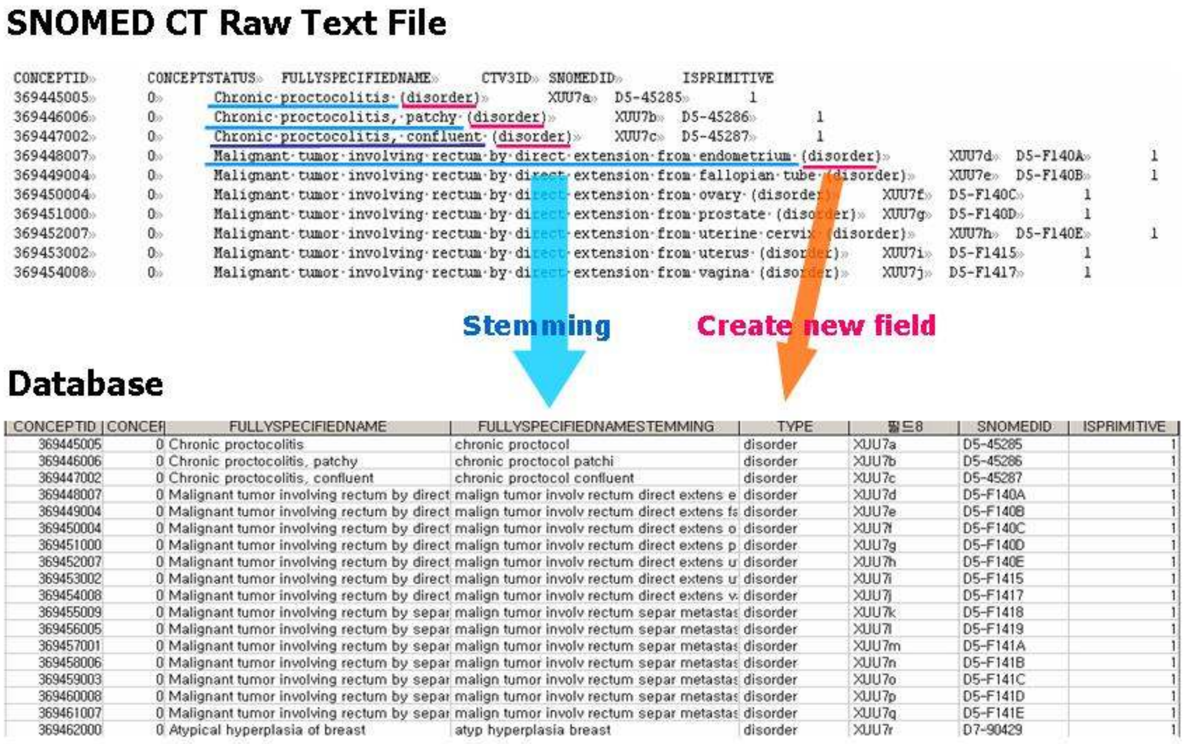
The percentage shows a mapping rate for 1,000 diagnoses which were excluded mapping failure.

• Text mapping methods

Each diagnosis is mapped to SNOMED CT through three stages of mapping process, Exact Mapping, Preprocessing Mapping, and Text Processing Mapping (Fig 5).

The first process is Exact Mapping stage wherein a query is made to SNOMED CT database with the extracted medical term without any processing. For example, a query is made with the term "Hepatocellular Carcinoma, s/p PET," without removing words following coma. If mapping is made in this stage, a CDA entry is generated without proceeding to next stages of mapping process. If not, it proceeds to the next mapping stage.

The second process is Preprocessing Mapping stage. In this step, a query is made to the row of FSNS of SNOMED CT database with preprocessed terms. Preprocessing steps



SNOMED CT Raw Text File

```
CONCEPTID»      CONCEPTSTATUS»  FULLYSPECIFIEDNAME»     CTV3ID» SNOMEDID»       ISPRIMITIVE
369445005»      0»      Chronic·proctocolitis·(disorder)»       XUU7a»  D5-45285»       1
369446006»      0»      Chronic·proctocolitis,·patchy·(disorder)»       XUU7b»  D5-45286»       1
369447002»      0»      Chronic·proctocolitis,·confluent·(disorder)»    XUU7c»  D5-45287»       1
369448007»      0»      Malignant·tumor·involving·rectum·by·direct·extension·from·endometrium·(disorder)»       XUU7d»  D5-F140A»       1
369449004»      0»      Malignant·tumor·involving·rectum·by·direct·extension·from·fallopian·tube·(disorder)»    XUU7e»  D5-F140B»       1
369450004»      0»      Malignant·tumor·involving·rectum·by·direct·extension·from·ovary·(disorder)»     XUU7f»  D5-F140C»       1
369451000»      0»      Malignant·tumor·involving·rectum·by·direct·extension·from·prostate·(disorder)»  XUU7g»  D5-F140D»       1
369452007»      0»      Malignant·tumor·involving·rectum·by·direct·extension·from·uterine·cervix·(disorder)»    XUU7h»  D5-F140E»       1
369453002»      0»      Malignant·tumor·involving·rectum·by·direct·extension·from·uterus·(disorder)»    XUU7i»  D5-F1415»       1
369454008»      0»      Malignant·tumor·involving·rectum·by·direct·extension·from·vagina·(disorder)»    XUU7j»  D5-F1417»       1
```

Database

CONCEPTID	CONCEP	FULLYSPECIFIEDNAME	FULLYSPECIFIEDNAMESTEMMING	TYPE	필드8	SNOMEDID	ISPRIMITIVE
369445005	0	Chronic proctocolitis	chronic proctocol	disorder	XUU7a	D5-45285	1
369446006	0	Chronic proctocolitis, patchy	chronic proctocol patchi	disorder	XUU7b	D5-45286	1
369447002	0	Chronic proctocolitis, confluent	chronic proctocol confluent	disorder	XUU7c	D5-45287	1
369448007	0	Malignant tumor involving rectum by direct	malign tumor involv rectum direct extens e	disorder	XUU7d	D5-F140A	1
369449004	0	Malignant tumor involving rectum by direct	malign tumor involv rectum direct extens fa	disorder	XUU7e	D5-F140B	1
369450004	0	Malignant tumor involving rectum by direct	malign tumor involv rectum direct extens o	disorder	XUU7f	D5-F140C	1
369451000	0	Malignant tumor involving rectum by direct	malign tumor involv rectum direct extens p	disorder	XUU7g	D5-F140D	1
369452007	0	Malignant tumor involving rectum by direct	malign tumor involv rectum direct extens u	disorder	XUU7h	D5-F140E	1
369453002	0	Malignant tumor involving rectum by direct	malign tumor involv rectum direct extens u	disorder	XUU7i	D5-F1415	1
369454008	0	Malignant tumor involving rectum by direct	malign tumor involv rectum direct extens va	disorder	XUU7j	D5-F1417	1
369455009	0	Malignant tumor involving rectum by separ	malign tumor involv rectum separ metastas	disorder	XUU7k	D5-F1418	1
369456005	0	Malignant tumor involving rectum by separ	malign tumor involv rectum separ metastas	disorder	XUU7l	D5-F1419	1
369457001	0	Malignant tumor involving rectum by separ	malign tumor involv rectum separ metastas	disorder	XUU7m	D5-F141A	1
369458006	0	Malignant tumor involving rectum by separ	malign tumor involv rectum separ metastas	disorder	XUU7n	D5-F141B	1
369459003	0	Malignant tumor involving rectum by separ	malign tumor involv rectum separ metastas	disorder	XUU7o	D5-F141C	1
369460008	0	Malignant tumor involving rectum by separ	malign tumor involv rectum separ metastas	disorder	XUU7p	D5-F141D	1
369461007	0	Malignant tumor involving rectum by separ	malign tumor involv rectum separ metastas	disorder	XUU7q	D5-F141E	1
369462000	0	Atypical hyperplasia of breast	atyp hyperplasia breast	disorder	XUU7r	D7-90429	1

Figure 4. Database for SNOMED CT

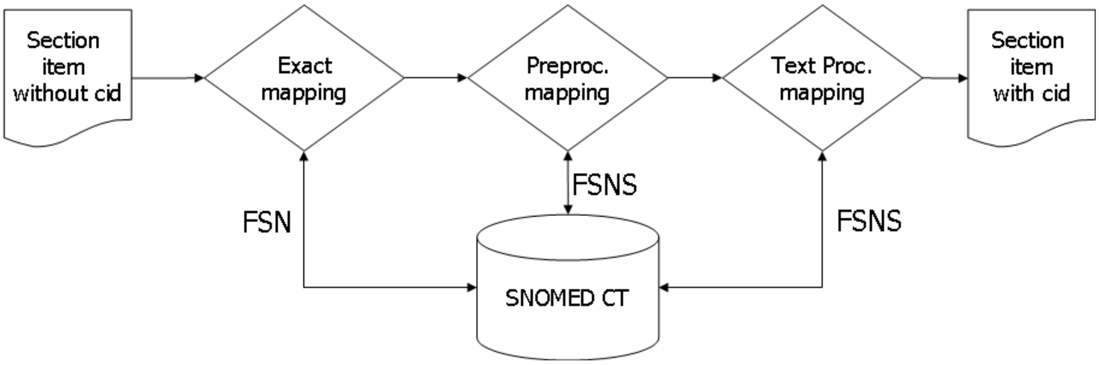


Figure 5. Three stages of mapping process

consist of Abbreviation Extension, Trimming, Removing punctuation mark, Removing stop word, and Stemming (Fig 6). A term is changed after the process of Abbreviation Extension like from “DM” to “Diabetes Mellitus.” Next, it is cleaned by Trimming and removed tems following comma or preposition mark such as from “Hepatocellular Carcinoma, s/p PET” to “Hepatocellular Carcinoma.” And then, it is removed a special symbol and stop word. Finally, it is changed root form by stemming like from “Localization-related epilepsy, NOS” to “Localization related epilepsy.” If mapping to SNOMED CT is made in this stage, a CDA entry is generated without proceeding to the third stage. If not, it proceeds to the last stage of Text Processing mapping.

The last process is Text Processing Mapping stage wherein a query is made after the process of Splitting Qualifier like from “Chronic renal failure” to two phrases of “Chronic” and “renal failure,” or Replacement like from “liver cirrhosis” to “cirrhosis liver,” or Synonym Extension like from “Gastric Cancer” to “carcinoma stomach,” or

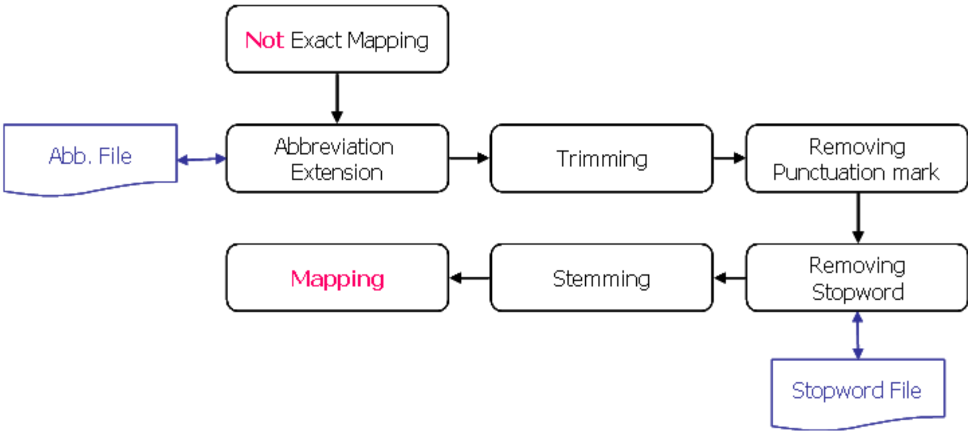


Figure 6. Steps of preprocessing mapping

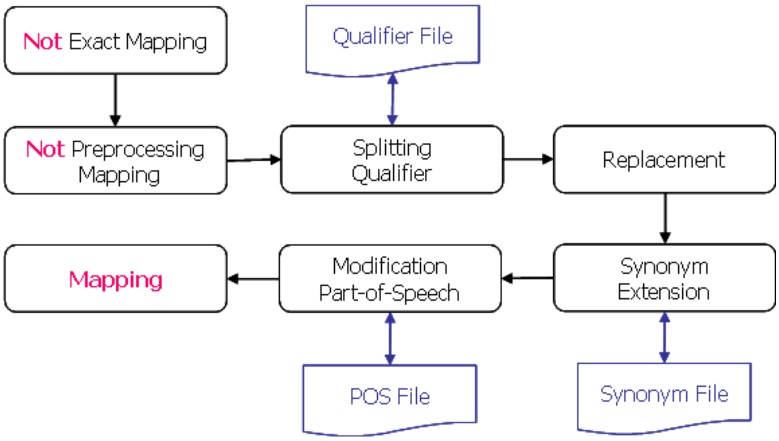


Figure 7. Steps of text processing mapping

Modification Part-of-Speech like from "Esophageal Cancer" to "carcinoma esophagus." In case of "Esophageal Cancer," it is converted to "carcinoma of esophagus" first. Then, stop word "of" is removed from "carcinoma of esophagus" and converted to a candidate phrase "carcinoma esophagu" by stemming (Fig 7).

• Implementation of Modified Entry Mapper

We implemented Modified Entry Mapper based on above three text mapping methods to map text in clinical documents into SNOMED CT automatically. The Modified Entry Mapper is composed of three detailed components, Encoder, Text Analyzer, and Code Mapper (Fig. 8).

Encoder searches with parser elements of <Section> which corresponds to the section of a CDA document, and reads the sentence corresponding to the value of <text> which is the child node of <Section> elements, then delivers it to Text Analyzer.

Text Analyzer is composed of Item Detector and Term Extractor (TE). Item detector is a module which extracts detail items, using Regular Expression, from the sentence delivered by Encoder. TE extracts the remaining part as a candidate phrase using Regular Expression after excluding numbers and special symbols unnecessary for searching SNOMED CT. For this process of extracting a candidate phrase, TE utilizes abbreviation word file and matching full term file, qualifier file, synonym file, part-of-speech for each word file, and stopword file.

Code Mapper makes queries to SNOMED CT database for extracted phrases and searches SNOMED CT after consecutive stages of Exact Mapping, Preprocessing Mapping, and Text Processing Mapping.

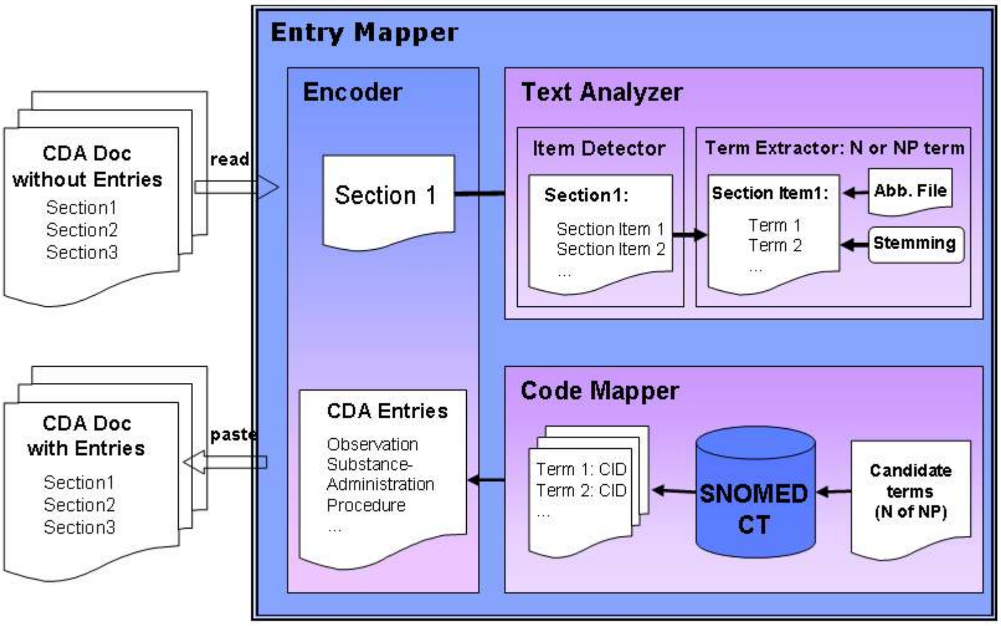


Figure 8. Architecture of Entry Mapper

For candidate phrases having matching SNOMED CT, proper entries are generated by Encoder according to the type of section they belong to. CDA documents containing an entry are generated using <SubstanceAdministration> element for the discharge medication, using <Procedure> element for the procedure, using <FutureEncounter> element for the future encounter, and using <Observation> element for other items which requires encoding.

III. Result

For the test, we randomly selected 500 discharge summaries among 15,618 SNUH discharge summaries and prepared 5 sets of test for each 100 discharges. Mapping performance of each test was evaluated by the result of mapping rate. In case of Exact Mapping, the result was from trying Exact Mapping alone. In case of Preprocessing Mapping, the result was from summing up Exact Mapping rate and Preprocessing Mapping rate. The result of Text Processing Mapping was summing up all three stages of mapping rate, Exact Mapping rate, Preprocessing Mapping rate and Text Processing Mapping rate.

From 500 documents, 913 detail items for the diagnosis, 581 for the chief complaint, 1,481 for the problem list and 2,422 for the discharge medication, were extracted using Regular Expression. The result of automatic mapping is illustrated in Table 3. Each mapping stage indicates total number of mappings and mapping rate. Using mapping

Table 3. Results of automatic mapping to SNOMED CT

	Diagnosis	Chief complaint	Problem list	Discharge medication
Mapping stages				
Exact mapping	468 (51.3%)	217 (37.4%)	377 (25.5%)	1,570 (64.8%)
Preprocessing mapping	518 (56.7%)	266 (45.8%)	561 (37.9%)	1,570 (64.8%)
Text processing mapping	790 (86.5%)	359 (61.8%)	929 (62.7%)	1,570 (64.8%)
Total	913	581	1,481	2,422

methods of this study, we obtained following results of mapping rate: 86.5% for the diagnosis, 61.8% for the chief complaint, 62.7% for the problem list, and 64.8% for the discharge medication.

In all sections except discharge medication, mapping rate gets higher as the mapping proceeds to next stage. Discharge medication section had the same mapping rate in all three stages of Exact Mapping, Preprocessing Mapping, and Text Processing Mapping, since most items of discharge medication section were composed of one word which can be mapped to SNOMED CT by Exact Mapping alone and those that failed in Exact Mapping were not qualified for next mapping stages, thus indicating all the same mapping rate at every stage.

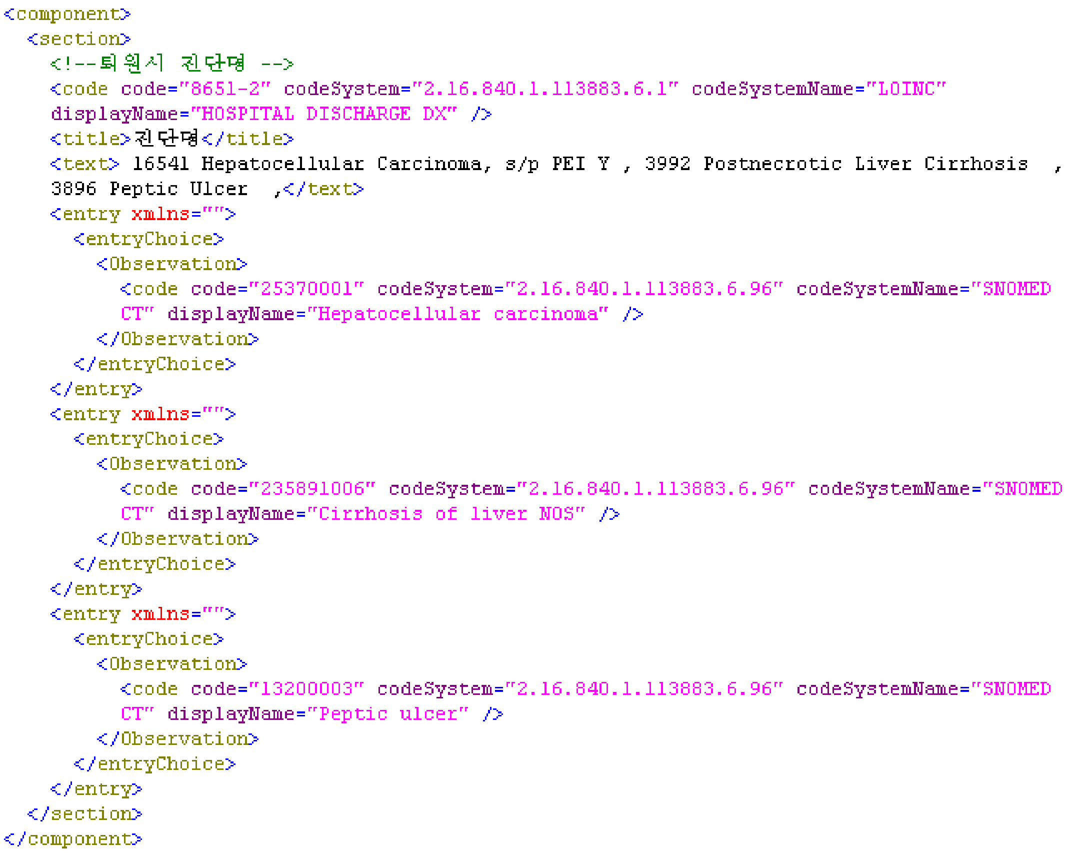
Based on the test method and result described above, we converted SNUH discharge summaries into CDA documents containing entry-level, using the Modified Entry Mapper. Figure 9 is example of CDA documents which were automatically generated by the Modified Entry Mapper.

Ⅳ. Discussion

In order to enable interoperability in exchanging of clinical documents between different institutions whether heterogeneous kind or not, it is necessary that CDA documents should be expressed as containing entry-level.

In this study, we automatically generated CDA documents containing entry-level by analyzing narrative patterns of diagnosis sections in discharge summaries and applying rules to the sections of diagnosis, chief complaint, problem list and discharge medication.

We generated entries through three mapping stages of Exact Mapping, Preprocessing Mapping and Text Processing Mapping to SNOMED CT, and compared the Modified Entry Mapper with previously developed Entry Mapper of CDA Studio® which has only Exact Mapping stage. Table 4 shows mapping results of previously developed Entry Mapper of CDA Studio® and Modified Entry Mapper. The mapping performance improved 37% for the diagnosis,



```
<component>
  <section>
    <!--퇴원시 진단명 -->
    <code code="8651-2" codeSystem="2.16.840.1.113883.6.1" codeSystemName="LOINC"
    displayName="HOSPITAL DISCHARGE DX" />
    <title>진단명</title>
    <text> 16541 Hepatocellular Carcinoma, s/p PEI Y , 3992 Postnecrotic Liver Cirrhosis  ,
    3896 Peptic Ulcer  ,</text>
    <entry xmlns="">
      <entryChoice>
        <Observation>
          <code code="25370001" codeSystem="2.16.840.1.113883.6.96" codeSystemName="SNOMED
          CT" displayName="Hepatocellular carcinoma" />
        </Observation>
      </entryChoice>
    </entry>
    <entry xmlns="">
      <entryChoice>
        <Observation>
          <code code="235891006" codeSystem="2.16.840.1.113883.6.96" codeSystemName="SNOMED
          CT" displayName="Cirrhosis of liver NOS" />
        </Observation>
      </entryChoice>
    </entry>
    <entry xmlns="">
      <entryChoice>
        <Observation>
          <code code="13200003" codeSystem="2.16.840.1.113883.6.96" codeSystemName="SNOMED
          CT" displayName="Peptic ulcer" />
        </Observation>
      </entryChoice>
    </entry>
  </section>
</component>
```

Figure 9. An example of CDA document with entries for semi-structured section (diagnosis section)

30% for the chief complaint, and 32% for the problem list each. These differences of mapping rates between previously developed Entry Mapper of CDA Studio® and the Modified Entry Mapper are caused by two other stages, Preprocessing Mapping and Text Processing Mapping. However, discharge medication section showed no improvement of mapping rate, since mapping was tried with one ingredient word not a phrase in discharge medication and other Text Processing methods except using synonym was inapplicable.

Table 4. Mapping Rates of two Mappers

	Previously developed Entry Mapper of CDA Studio®	Modified Entry Mapper
Diagnosis	50.5%	86.5%
Chief complaint	37.3%	61.8%
Problem list	31.9%	62.7%
Discharge medication	61.1%	64.8%

Despite overall improvement, however, mapping of 14-38% medical terms to SNOMED CT still remained unsuccessful. In many domestic clinical statements, English is being used mixed with Korean as well as many numbers and symbols such as "-" or "+" to indicate negative or positive test result, resulting in much difficulty in mapping to SNOMED CT which is mainly composed of English. Especially, chief complaint section and problem list section had lower mapping rate than in diagnosis section, because both sections contained more Korean words, numbers and symbols than diagnosis section. Similarly in discharge medication section, when mapping to SNOMED CT was tried with the name of medicine after failing with the name of ingredient, still mapping was not possible because many medicine names such as "Steptomycin lg inj 종근당" include Korean words. In diagnosis section, since it usually doesn't have diagnosis in Korean and rarely use numbers and symbols, its mapping rate is much higher than in other sections.

However, there still existed many limitations in mapping of diagnosis to SNOMED CT in diagnosis section. For instance, terms which doctors frequently use such as "Acute Myelocytic Leukemia (AML)" or "Acute Lymphocytic leukemia (ALL)" did not exist in SNOMED CT. Only a broader term like "acute leukemia" or a detailed term like "Acute promyelocytic" was in the SNOMED CT. Other frequently used terms such as "Hemiated Disc Disease" or "Lupus Nephritis" were not in SNOMED CT, therefore no mapping could be made, either. In other cases, mapping was not successful because of errata such as "Hypertention" which should have been "Hypertension" or wrong word spacing.

In conclusion, this study analyzed the narrative patterns of clinical statements and has showed the possibility of automatic generation of CDA entries by Text Processing method using those patterns. Further study should use an integrated method of diverse approaches considering limitations of automatic generation of CDA entries based on post-processing.

REFERENCE

1. Available at: http://www.hl7.org. Accessed December 13, 2007.
2. Available at: http://www.hl7.org/v3ballot/html/welcome/environment/index.htm. Accessed December 13, 2007.
3. Paterson GI, Shepherd M, Wang X, Watters C, Zitner D. Using the XML-based clinical document architecture for exchange of structured discharge summaries. Proceedings of the 35th Hawaii International Conference on System Sciences; 2002 Jan 7-10. pp.1200-1209.
4. Dalmiani S, Marcheschi P, Mazzarisi A. HL7 clinical document architecture to share structured in wide hospital information systems. 22nd International Conference, EurePACS-MIR 2004;2004.
5. Heitmann KU, Schweiger R, Dudeck J. Discharge and referral data exchange using global standards-the SCIPHOX project in Germany. Study Health Technol Inform. 2002;90:679-684.
6. Jung SW, Choe MS, Yoo SY, Park HK, Choi JW. et al. Development of CDA authoring Tool: CDA Studio. Proceedings of 7th International Workshop on HEALTHCOM 2005;2005 June 23-25. pp.307-310.
7. Choe MS, Jung SW, Choi JW. CDA studio: development of an integrated tool for generating CDA documents. Journal of Korean Society of Medical Informatics 2005;11(Supplement2):130-134.
8. Seo HJ, Hong SK, Park JY, Lee JA, Park YR, Kim JH.

Presentation of structural constraints for discharge note according to clinical document architecture standard. Journal of Korean Society of Medical Informatics 2005;11(2):189-197.

9. Kim IK, Lee JY, Kim IK, Cho H, Kwak YS. Clinical document repository system for electronic health record. Journal of Korean Society of Medical Informatics 2005;11(2):199-211.

10. Aronson AR. Effective mapping of biomedical text to the UMLS metathesaurus: the MetaMap program. 2001. AMIA Symposium Hanley & Belfus, 2001: 17-21.

11. Huang Y, Lowe H, Hersh W. A pilot study of contextual UMLS indexing to improve the precision of concept-based representation in XML-structured clinical radiology reports. J Am Med Inform Assoc. 2003;10:580-7.

12. Friedman C, Shagina L, Lussier Y, Hripcsak G: Automated encoding of clinical documents based on natural language processing. J Am Med Inform Assoc 2004, 11(5):392-402.