Information Extraction from Korean Radiology Reports Mingled Two Language

Miyoung Kwak, Seungbin Han, Jinwook Choi
Department of Biomedical Engineering, College of Medicine, Seoul National University

Abstract

This study presents overall of Information Extraction (IE) for SNUH (Seoul National University Hospital) radiology reports coexisted Korean and English using Concept Node (CN) which is a case frame as extraction rule. The following steps are performed: design conceptual model by terminology exploration based on lexical analysis, create a CN definition based on syntactic relationship pattern and implement automatic IE system using CN. Main purposes is to investigate whether syntactic and semantic analysis technique using extraction rule (CN) is effective for typical Korean medical text in mixed two different languages.

Introduction

Electronic Medical Records contain the majority of clinical data in unstructured text. The information in the textual document can be stored in a conceptual format and used to support clinical care by text summarization based on this conceptual format. [1] However, it is necessary to approach different way to extract keywords from texts coexisting mingled various language compare to typical IE system based on only English. From this problem, we investigate whether the IE core technique using CN definition which is an extraction rule could solve to extract this phrase expression limitation for typical Korean medical record.

Material and Methods

Data sources: we used 1500 texts of brain radiology reports available in SNUH between July 1999 and January 2003 for terminology exploration. The original reports consisted of two mixed languages, Korean (45.3%) and English (54.7%). Sentence sample:

1) T2W axial image 상에서 양쪽 frontal lobe 에 lateral ventricle 의 frontal horn 로 high signal intensity 가 보이지 않음.

2) 양측 frontal lobe 의 PVWM 에 patchy 하고 heterogeneous 하게 high signal intensity 를 증가하고 있다.

Conceptual Model

We used fifth step approach to identify the information elements contained in the radiology report for building conceptual model. First, we analyzed the compartmental and radiology-based document by identifying the main concepts on which they rely. Second, to analyze contents in the two different language mingled corpus, we used lexical tool (KLT 2.0) for observing the distribution of lexical information. Third, extract CT(*Candidate Term)* as valuable keyword based on term occurrence. We considered 2211 distinct terms as a CT. Usually, in nominal phrase of CT, adjectives are written with a Korean noun, tagged by 'N' as a modifier, and nouns are written in English, tagged by 'A' as a central finding (eg. 양쪽 frontal lobe). In this observation, as selected valuable CTs, distinct clinical terms written in English are 1737, Korean nouns are 53, Korean verb are 98 and numerical values like date are 323. Fourth, classified CTs into similar meaningful group called semantic entity (SE) and built an object-oriented model of radiology information. Fifth, three major semantic entities based high frequency are selected for annotating each sentence to find extraction rule. From this process, we determined three major attribute based on three major entities as a final template which is structure in RDB for extraction.

Analysis of statistical syntactic sentence using SE network

A sentence is composed of several words belonging to semantic entity (SE) network. We analyzed statistically the syntactic selective patterns of extracted sentences from 1000 documents. To do this, we first manually annotated two major semantic entities: $ <Target_Organ_Name>_에_<Rad_Change>_보이지 않다$

Extraction Rule using CN definition

We created the CN definition as a set of an extraction rule in case frame. Each CN definition specified a set of syntactic and semantic constraints that must be satisfied for the definition to be applied. The previous found verb and postposition is used as trigger word which is deterministic factor weather matched extraction rule occur. In this study, CN definition has 4 slots such as CN name, Trigger, Position, and Semantic Entity

Table 1. CN definition Examples

SLOT NAME	DESCRIPTION OF SLOTS
CN name	Concept node_verb_subject_Radchange
Trigger word	'보이지 않다'
Position	Subject
Semantic Entity	<Rad_Change>

Implementation of IE system

Automatic IE system consisted with four steps: document segmentation, lexical analysis, IE parser (Syntactic and semantic analysis) which is parsing an input sentence using appropriate CN definition on sentence when trigger word occur, and generate templates instances in RDB (Relational Database) format.

<Target_Organ_Name> → 양쪽 frontal lobe

<Rad_Change> → frontal horn high signal intensity

<Change_description> → 보이지 않음.

Results

We tested 10 partitions consisting of 100 texts each. We calculated precision, recall. Average precision is 85.18 and average recall is 93.71. SNUH brain radiology reports, physicians tend to write in two different languages in a sentence. We show this method is worth deliberation for complicated Korean medical documents written in two languages to provide detailed semantic information for user need.

References

[1] Carol, F. "A general Natural language text processor for clinical radiology", J Am Med Informatics Assoc, (1994), pp.161-174.