MEDINFO 2007
HISA Ltd Australia
© 2007 The authors. All rights reserved.

MedinfoParser: A Semantic tool for Information Extraction

Hee Kyong Park[a], Jinwook Choi[a], Elizabeth Liddy[b]

[a] *Department of Biomedical Engineering, Seoul National University, South Korea*
[b] *Center for Natural Language Processing, Syracuse University, U.S.A.*

Abstract and objective

We have developed a semantic tool (MedinfoParser) for the information extraction of medical documents. The main role of the parser is to parse medical sentences and extract key information. It puts Part-of-Speech (POS) tags and conceptual tags for the extraction. The parser consists of a Korean parser, an English parser, a numeric detector, and a noun phrase detector. For the semantic extraction, it gets conceptual id from the UMLS knowledge server.

Keywords:

MedinfoParser, information extraction, medical document

Introduction

In the era of the electronic medical record (EMR), huge amount of medical documents are produced electronically. EMR provides many kinds of templates for the coded data entry, but most clinical notes are still written in free text. To catch the patient's clinical status in a short period of time, EMR should provide a neat tool for summarizing medical records. We have developed MedinfoParser which is a semantic tool for information extraction.

Methods

MedinfoParser consists of a Korean parser, an English parser, a numeric detector, and a noun phrase detector.

Korean parser: Korean parser breaks a document into sentence fragments and tokenizes sentences. It provides POS tags of words and those of morphemes. As a Korean parser we have used KLT parser[1]. We modified the numeric module of KLT parser in order to provide more sophisticated analysis. Numeric expressions are more distinctly denoted by separating the numeral tokens from the alphanumeric combined words. And numeral words are converted into Arabian numbers. A new tag of 'NB' and 'M' are introduced to annotate numeric expressions more precisely.

English parser: In general, Korean medical records are written in Korean and English mixed. However, most medical expressions are written in English. In order to parse those key medical expressions, English parser detects medical terms written in English. By sending queries to UMLS knowledge server, it gets UMLS cuis. Then it puts a 'MED' tag and a UMLS cui at the end of each medical term. Medical documents contain lots of abbreviations and locally defined terms which are used only in a specific institution. As these terms are not searched through UMLS, we designed user dictionaries.

Numeric detector: As the numeric expression contains very important meaning in medical records, we have improved the numeral identification function. Especially in order to extract key information in the chronological order, the identification of the temporal expression is the most significant. For the Numeric detector we designed a rule pattern based on data frames for the semantic extraction. Each data frame works as a unit concept. It has various methods to evoke functions. The numeric detection rule contains data frames which are necessary to extract numeric expressions from the medical records.

Noun phrase detector: Noun phrase (NP) detector identifies medical concepts that are consisted of more than two words. NP boundary can be identified in many ways according to the reviewer's view. Therefore we made principles to clearly identify noun phrases. About one hundred NP lists were extracted from the discharge summaries in Seoul National University Hospital. We analyzed the context of each NP occurrence and have been developing rule patterns for NP detection.

Discussion and conclusion

We have developed a semantic tool for the information extraction of medical documents. The implemented Korean parser and English parser generate the highly readable XML files that can be used for other medical applications. For the future work, we will display the output of MedinfoParser onto the timeline to make a chronological summary. We hope that MedinfoParser will be a promising tool for the information extraction in medical field.

1 From the NLP laboratory of Kookmin university in South Korea