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.***

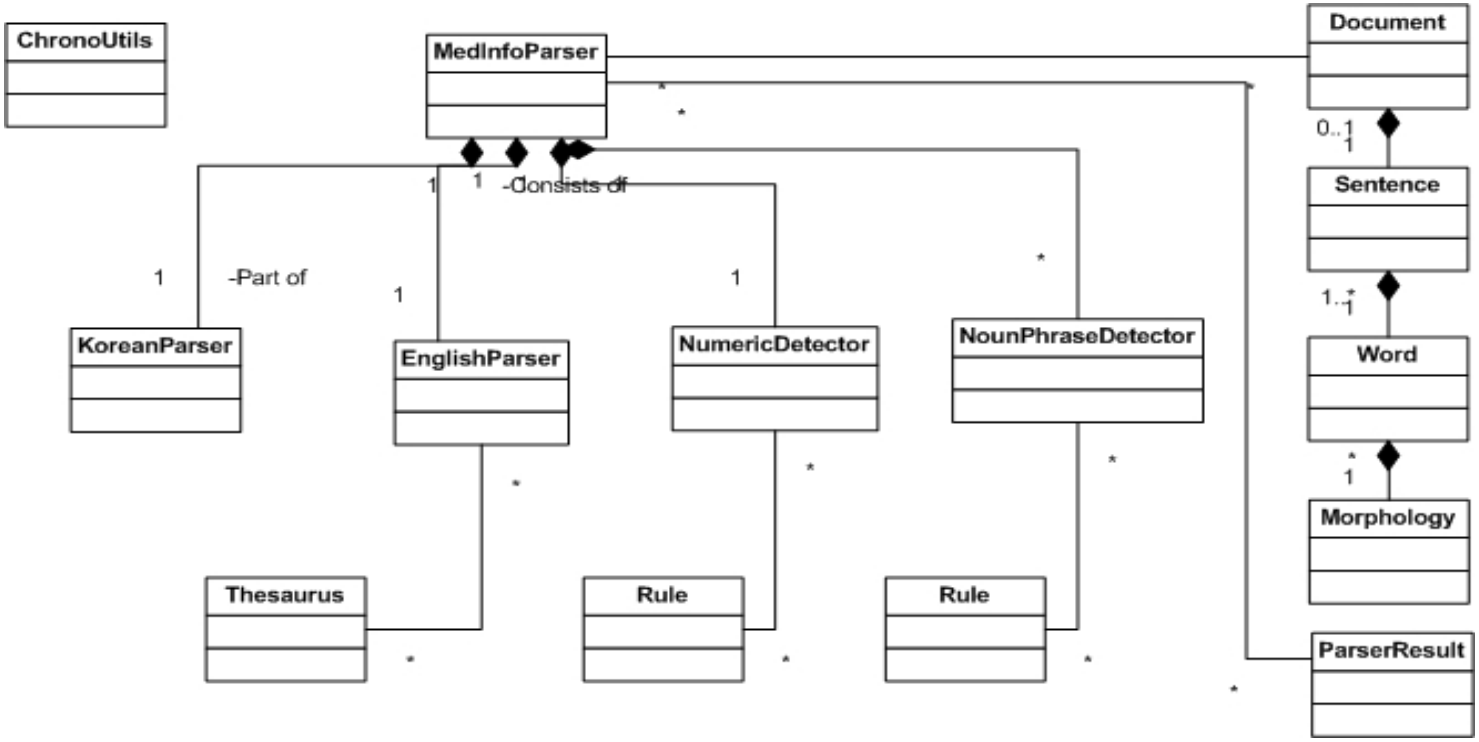
Introduction

- Background
  - Era of the electronic medical record (EMR)
  - Huge amount of medical documents are produced electronically
  - Most clinical notes are still written in free text
  - Difficulties in catching a patient's clinical status in a short period of time
  - Needs for a neat tool for summarizing medical records
- Purpose
  - Develop a semantic tool for summarizing medical records
- We have developed MedinfoParser

MedinfoParser

- A semantic tool for information extraction
- Role
  - Tokenization
  - POS tagging
  - Concept recognition
    - medical concept
    - numeric concept
    - noun phrase identification
- Components
  - Korean parser
  - English parser
  - Numeric detector
  - Noun phrase detector

A Classdiagram of the MedinfoParser



Korean Parser

- Tokenizes and parses Korean medical documents
  - Breaks a document into sentence fragments
  - Tokenizes sentences
  - Puts POS tags after the each parsed word and morpheme

```
<s>
<ins>Child C LC    Chronic PSE    Supportive care 위해 입원함</ins>
<anl>
Child|N|@|(Child|N)
C|N|@|(C|N)
LC|N|@|(LC|N)
Chronic|N|@|(Chronic|N)
PSE|N|@|(PSE|N)
Supportive|N|@|(Supportive|N)
care|N|@|(care|N)
위해|VM|?|(위하|V + 어|e)
입원함|NSM|?|(입원|N + 하|t + ㅁ|e)
</anl>
</s>
<s>
<ins>Ativan therapy 하며 밤에 재움    점차 Liver failure 진행되어 사망</ins>
<anl>
Ativan|N|@|(Ativan|N)
therapy|N|@|(therapy|N)
하며|VM|?|(하|V + 며|e)
밤에|NJ|N|(밤|N + 에|j)
```

An example of Korean parser output

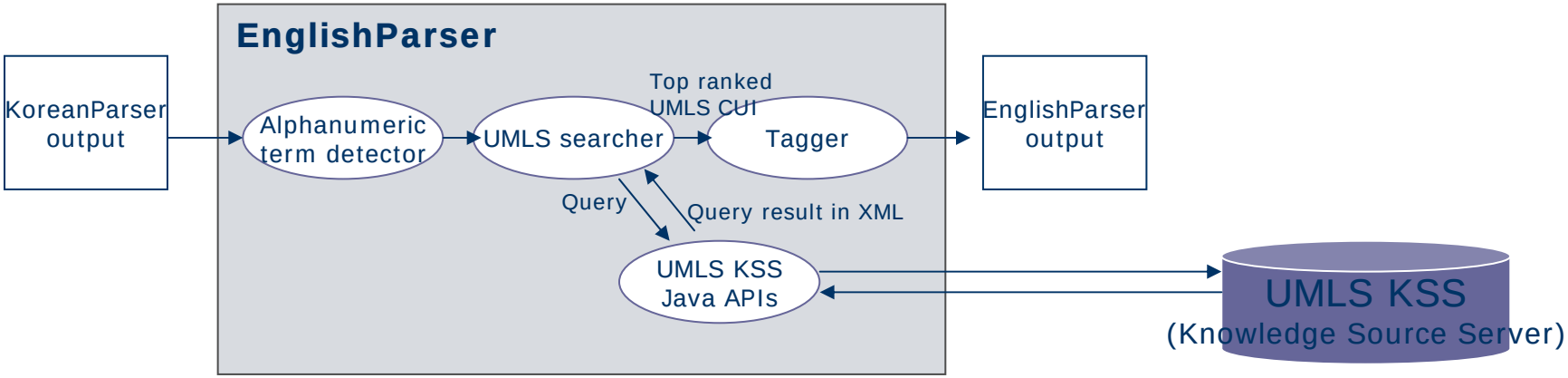
English Parser

- Characteristics of Korean medical records
  - Written in Korean and English mixed
  - Most medical expressions are written in English
  - Contain a lot of abbreviations and locally defined terms
- Parses key medical expressions
  - Detects alphabetic words
  - Sends queries to UMLS knowledge source server
  - If any UMLS CUIs are returned, puts ‘MED’ tag and a top ranked UMLS cui
- To detect medical terms that are not searched from UMLS, we designed user dictionaries

```
<s>
<ins>Child C LC    Chronic PSE    Supportive care 위해 입원함</ins>
<anl>
Child|N|MED|(Child|N|C0008059)
C|N|MED|(C|N|C0227087)
LC|N|MED|(LC|N|C0008565)
Chronic|N|MED|(Chronic|N|C0205191)
```

An example of ‘MED’ tagged result generated by English parser

MED tagging Process




```
  <?xml version="1.0" encoding="UTF-8" ?>
- <ConceptIdCollection version="1.0">
  - <query>
    - <findCUI version="1.0">
        <release>2006AB</release>
        <term>AGE</term>
        <language>ENG</language>
        <approx />
      </findCUI>
    </query>
    <release>2006AB</release>
  - <conceptId>
      <cui>C0001779</cui>
      <cn>Age</cn>
    </conceptId>
  - <conceptId>
      <cui>C0001771</cui>
      <cn>Agar</cn>
    </conceptId>
  - <conceptId>
      <cui>C0242449</cui>
      <cn>Ageism</cn>
    </conceptId>
  - <conceptId>
      <cui>C0030543</cui>
      <cn>Age, Parental</cn>
    </conceptId>
  - <conceptId>
      <cui>C0024915</cui>
      <cn>Age, Maternal</cn>
    </conceptId>
```


<an example of query result in XML>

: the CUIs are ranked by the UMLS KSS

Numeric Detector

- Numeric expressions are key information for summarizing events in chronological order
- Extracts numeric terms and finds clinical events which are relevant to the terms based on numeric detection rules
- The numeric detection rules contain data frames which are necessary to extract numeric expressions from the medical records
- Defined seventeen numeric concepts that are critical in medical field, and other concept classes needed for the numeric concept extraction

```
@drugName
listed in thesaurus;
has umlsID;
;;

@drugUnit
listed in (drug unit) thesaurus;
following number;
;;

@medication
expressed in integer;
following drugName;
followed by drugUnit;
;;
```

A sample numeric detection rule for '*medication*'

Rule patterns

- Designed based on data frames (David W. Embley, 2004)
  - Representational frame that encapsulates the essential properties of everyday data items
- Each data frame contains functions, other concept classes, relations, expressions to extract single concept
- Each data frame works as a unit concept. It has various methods to evoke functions

Noun Phrase Detector

- Identifies boundaries of the medical noun phrases (NPs) which are consisted of more than two words
- Made principles to clearly identify boundaries of ambiguous cases
  - The main idea of the principles : the longest noun phrase which has an atomic medical concept should be identified as a NP
  - E.g., *'… Child C liver cirrhosis …'*
    - Possible NPs : *'liver cirrhosis', 'Child C liver cirrhosis'*
    - Noun phrase detector identifies *'Child C liver cirrhosis' as a NP*
- Have been developing rules for NP detection

```
<s>
<ins>Child C LC    Chronic PSE    Supportive care 위해 입원함</ins>
<anl>
<NP>
Child|N|MED|(Child|N|C0008059)
C|N|MED|(C|N|C0227087)
LC|N|MED|(LC|N|C0008565)
</NP>
<NP>
Chronic|N|MED|(Chronic|N|C0205191)
PSE|N|MED|(PSE|N|C1518860)
</NP>
<NP>
Supportive|N|@|(Supportive|N)
care|N|MED|(care|N|C0580931)
</NP>
위해|VM|?|(위하|V + 어|e)
입원함|NSM|?|(입원|N + 하|t + ㅁ|e)
</anl>
</s>
<s>
<ins>Ativan therapy 하며 밤에 재움    점차 Liver failure 진행되어 사망</ins>
<anl>
<NP>
Ativan|N|MED|(Ativan|N|C0699194)
...
```

An example of noun phrase tagged output

Discussion and Conclusion

- Have developed a semantic tool for the information extraction of medical documents
- Each module produces a highly readable XML file as an output
- Easily adoptable to other medical applications
- Future work
  - Display the output of MedinfoParser onto the timeline to make a chronological summary

References

- NLP laboratory of Kookmin University. http://nlp.kookmin.ac.kr/
- UMLS knowledge source server. http://umlsks.nlm.nih.gov/
- David W. Embley. Toward semantic understanding: an approach based on information extraction ontologies, Proceeding of the 15th Australasian Database Conference, 2004; p.3-12.