AUTOMATIC CONTENT EXTRACTION FOR DESIGNING A FRENCH CLINICAL CORPUS

Louise DELEGER — Cyril GROUIN — Aurelie NEVEOL

How to improve further information extraction process?
How to separate the core medical content from other zones within clinical notes?

General structure of clinical notes

- Contact information for the health care unit in which the note was created. Generic Header is the same for all clinical notes from the same unit.
- Specific Patient Information (Name, Date of Birth, Date of Admission). Part of the Specific Header may be repeated in several notes concerning the same patient.
- Medical content of a note is unique for each document.
- Physician’s signature, greetings, etc.

Automatic identification of zones within clinical records

Identification process

Manual annotation of zones
(2 curators, $\chi=.98$)

CRF model creation
Machine-learning tool
Clinical records
(200 docs)
Automatic zone identification
Manual control (3 curators)
Clinical records
(500 docs)

Automatic zone identification
Clinical records
(500 docs)

- Random selection of clinical notes in French
  - Sample 1 (200 docs): development corpus, to design the system
  - Sample 2 (500 docs): test corpus, to evaluate the final system

- Similarly to [Hirohata et al., 2008, Tepper et al., 2012], CRF model creation (WAPITI tool [Lavergne et al., 2010]) based on:
  - first token of current line, previous line, next line
  - second token of current line
  - bigram of 1st and 1nd token of current line
  - is first token in upper case?
  - relative position of current line in the document
  - line length (number of tokens)
  - presence of blank lines before current line
  - presence of figures on current line
  - presence of emails on current line

Results and Discussion

- Precision, Recall, F-measure

Impact of medical content identification on corpus size

<table>
<thead>
<tr>
<th></th>
<th>Raw corpus</th>
<th>Content zone</th>
<th>Diff.</th>
</tr>
</thead>
<tbody>
<tr>
<td># words</td>
<td>171,722</td>
<td>100,730</td>
<td>-41%</td>
</tr>
<tr>
<td># sentence</td>
<td>18,815</td>
<td>9,013</td>
<td>-52%</td>
</tr>
<tr>
<td>Avg. word nb/doc</td>
<td>343</td>
<td>202</td>
<td>-41%</td>
</tr>
<tr>
<td>Avg. sent. nb/doc</td>
<td>38</td>
<td>18</td>
<td>-53%</td>
</tr>
<tr>
<td>Avg. word nb/sent.</td>
<td>9</td>
<td>11</td>
<td>-22%</td>
</tr>
</tbody>
</table>

References


Acknowledgements

- National Research Agency, project CABeRneT, grant ANR-13-JS02-0009-01
- The authors thank the Biomedical Informatics Department at the Rouen University Hospital for providing access to the LERUDI corpus for this work

All Personal Health Information was replaced with plausible substitutes to protect patient privacy.

CNRS, LIMSI (UPR 3251), Orsay, France

AMIA Annual Symposium – Washington (DC), November 2014