De-identification of French medical narratives

Vasiliki Foufi, PhD, Division of Medical Information Sciences, UNIGE & HUG

@VasilikiFoufi
THEORETICAL FRAMEWORK

• Medical narratives contain:
  – Important medical information for secondary usage
  – Protected Health Information (PHI), Personal Identifiable Information (PII)

• De-identification: process of masking or removing sensitive data
  – Ensure data security and privacy
  – Preserve data integrity

• US Health Insurance Portability and Accountability Act (HIPAA) regulation
METHOD

• Named Entity Recognition (NER) task:
  – Names (patients, doctors, nurses, health insurance companies)
  – Locations
  – Elements of dates
  – Addresses
  – Telephone and fax numbers
  – Social security numbers

• Symbolic (rule-based) method via finite state automata

• Replacement of de-identified PHI by credible surrogate information
METHOD
METHOD

Advantages

• High precision and high recall text recognition
• Correctable, reproducible and sharable rules
• Explainable results
• Hospital production environment

Disadvantages

• Language dependent
• Context dependent
• A lot of working hours and human resources
TOOL

• Unitex corpus processor:
  – Multilingual open-source tool
  – Developed by S. Paumier at the Paris-Est Marne-la-Vallée university
  – Lexicon- and grammar-based
  – Downloadable at http://unitexgramlab.org/
FINITE STATE AUTOMATA

De-identification of dates

De-identification of places
PATIENT & PHYSICIAN NAMES

• Use of internal and external triggers for the identification of person’s names:
  – Titles: Monsieur (Mr), Madame (Mrs), Dr, Prof, Dresse
  – Specializations: Mr X, general practitioner or Mr X specialized in breast cancer

• Proper names not de-identified: maladie de Parkinson [Parkinson’s disease]

• Replacement of person names with false names selected randomly
EXAMPLE

Initial text

- M. Gaudet-Blavignac a été transféré à la clinique de Joli-Mont le 06 janvier 2018.
- Il est convenu de la mise en place d'une aide infirmière quotidienne par la FSASD.

De-identified text

- M. Foufi a été transféré à la Clinique le 30 février 2018.
- Il est convenu de la mise en place d'une aide infirmière quotidienne à domicile.
### DE-IDENTIFICATION RESULTS

<table>
<thead>
<tr>
<th></th>
<th>Dates</th>
<th>Patient names</th>
<th>Physician names</th>
<th>Locations</th>
<th>Total performance</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Precision</strong></td>
<td>98.9%</td>
<td>99.7%</td>
<td>100%</td>
<td>96.3%</td>
<td>99.1%</td>
</tr>
<tr>
<td><strong>Recall</strong></td>
<td>92.3%</td>
<td>99.2%</td>
<td>98.8%</td>
<td>78.7%</td>
<td>93.4%</td>
</tr>
</tbody>
</table>
CURRENT STATUS

- 30 finite state automata for French documents
- Construction of a manually annotated dataset
- De-identification rules for German
NEXT STEPS

• Processing of time intervals
• Other types of medical documents
• Rules for de-identifying medical documents in Italian
REFERENCES

ACKNOWLEDGEMENTS

This project has been financed by the Swiss Personalized Health Network