Das Forum für ICT im Gesundheitswesen
Le forum pour les TIC dans le système de santé

@eHealthSummit
www.ehealthsummit.ch

SWISS eHEALTH SUMMIT
2018 HIMSS Europe

SGMI SSIM SSMI
Schweizerische Gesellschaft für Medizinische Informatik
Société Suisse d'Informatique Médicale
Società Svizzera d'Informatica Medica
Swiss Society for Medical Informatics

STADE DE SUISSE
BERN
11.-12. SEPT. 2018

De-identification of French medical narratives

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

@VasilikiFoufi

In cooperation with

ehealthsuisse

IHE SUISSE Integrating the Healthcare Enterprise

pharmaSuisse

VGIch

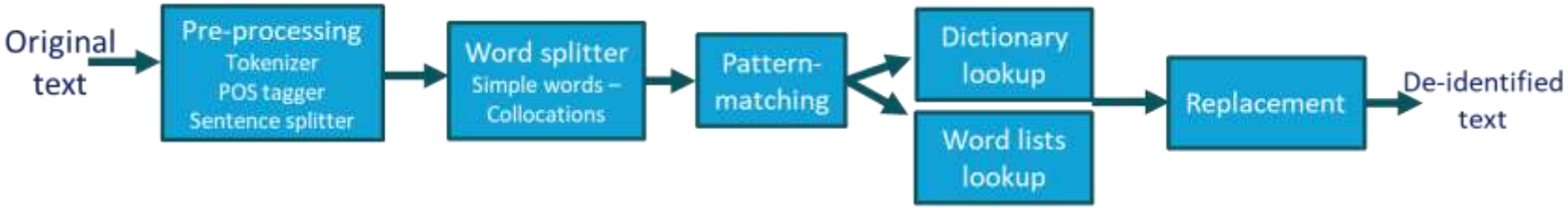
THEORETICAL FRAMEWORK

- 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

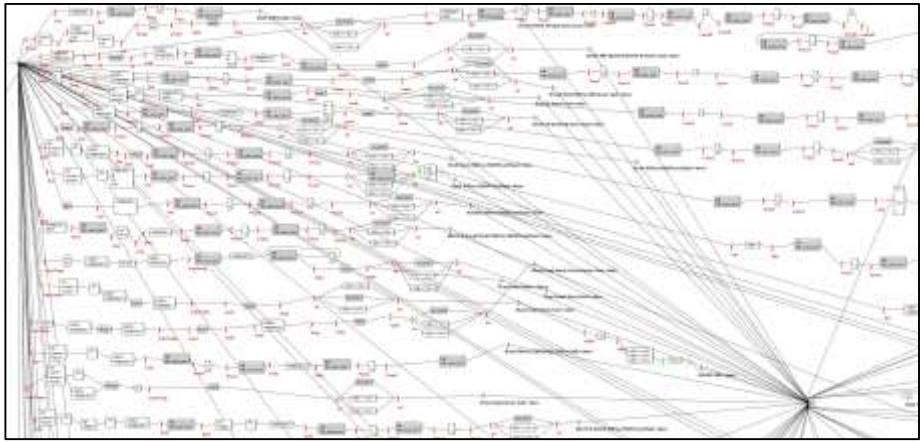
Disadvantages

- Language dependent
- Context dependent
- A lot of working hours and human resources

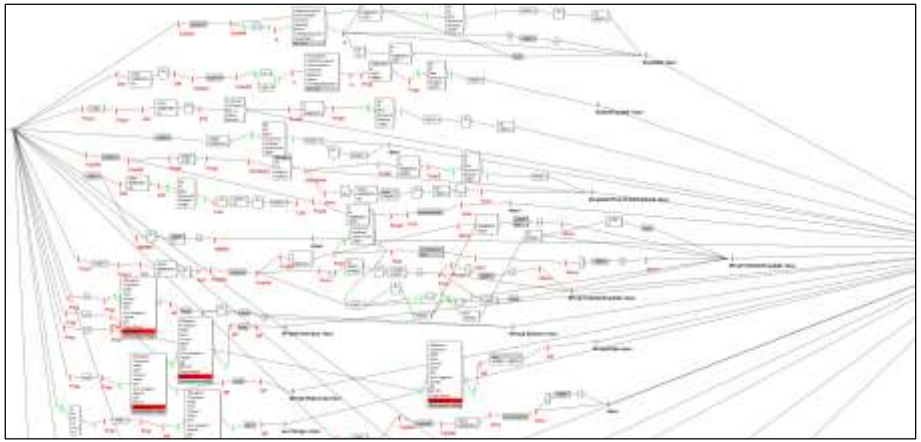
TOOL

- *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

	Dates	Patient names	Physician names	Locations	Total performance
Precision	98.9%	99.7%	100%	96.3%	99.1%
Recall	92.3%	99.2%	98.8%	78.7%	93.4%

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

[1] Douglass M, Clifford GD, Reisner A, Moody GB, Mark RG. Computer-Assisted Deidentification of Free Text in the MIMIC II Database. Computers In Cardiology. 2004;31:341-344.
[2] Meystre S, Savova G, Kipper-Schuler K, Hurdle J. Extracting information from textual documents in the electronic health record: a review of recent research. Yearb Med Inform. 2008:128-44.
[3] Stubbs A, Uzuner Ö. Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/UTHealth corpus. J Biomed Inform. 2015 Dec;58 Suppl:S20-9.
[4] Ferrández O, South BR, Shen S, Friedlin FJ, Samore MH, Meystre SM. BoB, a best-of-breed automated text de-identification system for VHA clinical documents. J Am Med Inform Assoc. 2013 Jan 1;20(1):77-83.
[5] Deleger L, Molnar K, Savova G, Xia F, Lingren T, Li Q, et al. Large-scale evaluation of automated clinical note de-identification and its impact on information extraction. J Am Med Inform Assoc. 2013 Jan-Feb; 20(1): 84-94.
[6] Sweeney L. Replacing personally-identifying information in medical records, the Scrub system. Proc AMIA Annu Fall Symp. 1996:333-7.
[7] Sweeney L. Guaranteeing anonymity when sharing medical data, the Datafly system. Proc AMIA Annu Fall Symp. 1997:51-5.
[8] Levine JM. De-identification of ICU Patient Records. Massachusetts Institute of Technology; 2003.
[9] Neamatullah I, Douglass M, Lehman LH, Reisner A, Villarroel M, Long WJ, et al. Automated De-Identification of Free-Text Medical Records. BMC Medical Informatics and Decision Making. 2008;8:32.
[10] Gardner J, Xiong L, Kanwei L, Lu JJ. HIDE: heterogeneous information Deidentification. Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology (EDBT '09); 2009 Mar 24-26; Saint Petersburg, Russia. 2009.
[11] Goldberger AL, Amaral LA, Glass L, Hausdorff JM, Ivanov PC, Mark RG, et al. PhysioBank, PhysioToolkit, and Physionet: Components of a New Research Resource for Complex Physiologic Signals. Circulation. 2000 Jun 13;101(23):E215-20.
[12] Jaćimović J, Krstev C, Jelovac D. A Rule-Based System for Automatic De-identification of Medical Narrative Texts. Informatica. 2015;39:45-53.
[13] Ruch P, Baud R, Rassinoux AM, Bouillon P, Robert G. Medical Document Anonymization with a Semantic Lexicon. Proc AMIA Symp. 2000:729-733.
[14] Shin SY, Park YR, Shin Y, Choi HJ, Park J, Lyu Y, et al. A De-identification Method for Bilingual Clinical Texts of Various Note Types. J Korean Med Sci. 2015 Jan;30(1):7-15.
[15] Dias FMC. Multilingual Automated Text Anonymization. Instituto Superior Técnico of Lisboa; 2016.
[16] Thomson P, McNaught J, Ananiadou S. Customised OCR Correction for Historical Medical Text. Digital Heritage. 2015:35–41.

ACKNOWLEDGEMENTS

This project has been financed by the Swiss Personalized Health Network