HOW TO DE-IDENTIFY A LARGE CLINICAL CORPUS IN 10 DAYS

Cyril GROUIN Louise DELÉGER Jean-Baptiste ESCUDIÉ Grégory GROISY Anne-Sophie JANNOT Bastien RANCE Xavier TANNIER Aurélie NÉVÉOL

Schedule

Set-up phase

DAY 1
- Install MEDINA suite at the hospital, including:
  - Rule-based module
  - CRF model construction pipeline (Wapiti, TreeTagger)
  - Annotation tool (BRAT)

DAY 2
- User training, including:
  - Detailed presentation of the de-identification protocol
  - PHI annotation
  - MEDINA suite

DAY 3
- Write custom-launch scripts
- Test pipeline throughout

Warm-up phase

DAY 4
- Pre-annotate a set of 20 documents with MEDINA rules
- 2 annotators revise the automatic de-identification
- Compute IAA
- Discuss anotation disagreements
- Create consensus
- Validated de-id corpus contains 20 documents

DAY 5
- Create a custom CRF model on validated corpus
- Pre-annotate a set of 20 documents with CRF model
- 2 annotators revise the automatic de-identification
- Compute IAA
- Discuss anotation disagreement
- Create consensus
- Validated de-id corpus contains 40 documents

Production phase

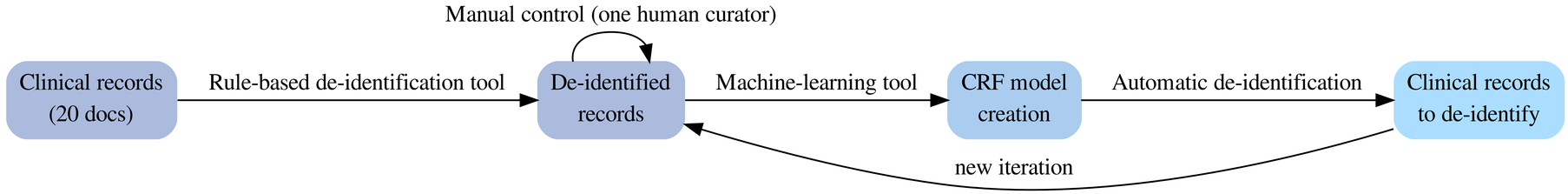
DAY 5
- Create a custom CRF model on validated corpus
- Pre-annotate two sets of 20 documents with CRF model
- 2 annotators each revise one pre-annotated set
- Validated de-id corpus contains 80 documents

DAY 6-10
- Repeat production routine
- Final validated de-id corpus

How to apply an existing clinical records de-identification protocol in a hospital?
How can outside collaborators rapidly use an existing de-identification tool?

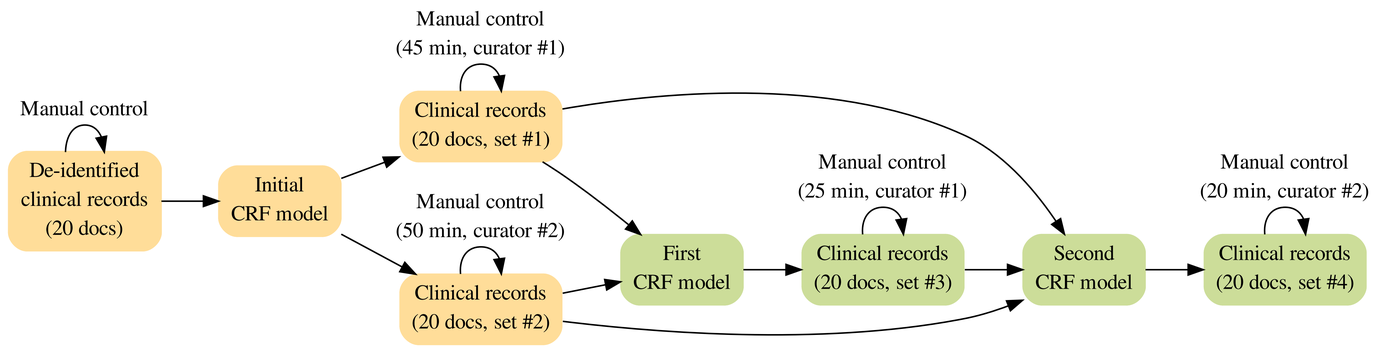
Application of a de-identification protocol by outside collaborators

De-identification protocol



- Selection of 20 documents (no additional annotated documents needed to train the CRF model) [Grouin and Névéol, 2014]
- Automatic de-identification using an existing rule-based de-identification tool: MEDINA [Grouin and Zweigenbaum, 2013]
- Manual control of de-identified documents by one human (no significative statistical difference if performed by two humans) [Grouin et al., 2014]
- CRF model creation based on the 20 de-identified and controlled documents
- Application of the CRF model on new documents to de-identify

De-identification process within the hospital



Features used to build the CRF models:
- Surface features: case, punctuation, digit, token length
- Deep features: part-of-speech, lexical look-up
- External features: position of the token in the record, cluster id

Detailed F-measure for each set

Category [min;max nb]	Corpus set			
	#1	#2	#3	#4
first name [55;227]	.648	.206	.941	.924
last name [213;301]	.824	.691	.914	.947
email [4;12]	.960	1.00	1.00	1.00
hospital [6;15]	.167	.050	.684	.698
address [9;22]	.593	.400	.800	.800
postcode [21;32]	.794	.655	.873	.970
city [8;35]	.436	.348	.831	.959
date [75;120]	.636	.619	.937	.853
phone [104;183]	.883	.875	.997	.967
id [2;27]	.091	.065	.857	.788
Overall [643;843]	.727	.570	.927	.915

References

Grouin, C., Lavergne, T., and Névéol, A. (2014).
Optimizing annotation efforts to build reliable annotated corpora for training statistical models.
In *Proc of Linguistic Annotation Workshop (LAW-VIII)*, pages 54–58, Dublin, Ireland. Coling, Association for Computational Linguistics.

Grouin, C. and Névéol, A. (2014).
De-identification of clinical notes in french: towards a protocol for reference corpus development.
J Biomed Inform, 50:151–61.

Grouin, C. and Zweigenbaum, P. (2013).
Automatic de-identification of french clinical records: Comparison of rule-based and machine-learning approches.
In *Stud Health Technol Inform*, volume 192, pages 476–80.

Acknowledgements

French National Research Agency ANR
- project Accordys, grant ANR-12-CORD-0007-03
- project CABeRneT, grant ANR-13-JS02-0009-01