Automatic De-Identification of French Clinical Records:
Comparison of Rule-Based and Machine-Learning Approaches

Cyril Grouin\textsuperscript{1,2} \hspace{1cm} Pierre Zweigenbaum\textsuperscript{1}

\textsuperscript{1}LIMSI–CNRS, Orsay, France
\textsuperscript{2}INSERM U872 Eq20 & UPMC, Paris, France

August 22\textsuperscript{nd}, 2013

MEDINFO 2013 — Copenhagen, Denmark
1 Introduction

2 Materials and methods
   • Corpora
   • Approaches

3 Results

4 Discussion & conclusion

5 References
<table>
<thead>
<tr>
<th>1</th>
<th>Introduction</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>Materials and methods</td>
</tr>
<tr>
<td></td>
<td>Corpora</td>
</tr>
<tr>
<td></td>
<td>Approaches</td>
</tr>
<tr>
<td>3</td>
<td>Results</td>
</tr>
<tr>
<td>4</td>
<td>Discussion &amp; conclusion</td>
</tr>
<tr>
<td>5</td>
<td>References</td>
</tr>
</tbody>
</table>
Introduction

- **Medical records:**
  - Contain clinical information w/ nominative data:
    - patient and family:
      - I saw your patient Mr John Lasalle...
      - Confident person: Ms Jane Lasalle (daughter)
    - hospital team: Dr John DOE
  - Written in natural language → NLP systems needed
  - Medical records corpora useful to develop/train NLP systems:
    - to create linguistic patterns and statistical models
    - to evaluate the performances of the systems
  - De-identification mandatory: HIPAA (USA), CNIL (France)
- **i2b2 2006 de-identification challenge** [Uzuner et al., 2007]
  - Rule-based methods: regular expressions, lists
  - Machine-learning methods: training corpus
Objectives

To compare two approaches to de-identify French clinical records
- rule-based system: Medina [Grouin, 2013]
- CRF machine-learning via Wapiti [Lavergne et al., 2010]

To evaluate the robustness of each approach
- known corpus in cardiology
- unknown corpus in foetopathology
1. Introduction

2. Materials and methods
   - Corpora
   - Approaches

3. Results

4. Discussion & conclusion

5. References
Two medical domains (from two hospitals):

- **Cardiology**: 312 partially de-identified clinical records
  - no patient names → reintroduction of surrogate PHI
  - remaining clinician names
  - known characteristics: corpus used to build both systems

- **Fœtopathology**: 10 documents obtained from OCR process
  - examination records, scan notes, corresponding letters
  - unknown characteristics: to evaluate the robustness of our systems
Annotation guidelines

**Nine categories** (HIPAA + corpus properties):

- **person names**: *first names, last names* (patient + clinical staff)
- **location names**: *address, town, zip code, hospital*
- **numerical data**: *date, phone #, social security #, serial #*
- **medical device**: *trade mark, model*
Annotation process

First set (100 files)

Annotator A

Annotator B

Consensus (κ=0.81)

κ=0.87
κ=0.93

62 files

Test corpus (62 files)

Second set (212 files)

38 files

single annotation 212 files

Training corpus (250 files)

κ=0.81

First set (100 files)

Annotator A

Annotator B

Consensus (κ=0.81)

κ=0.87 κ=0.93

62 files

Test corpus (62 files)

Second set (212 files)

38 files

single annotation 212 files

Training corpus (250 files)
Annotation process

The chart shows the percentage distribution of annotations for different datasets:
- **Cardio-Train**
- **Cardio-Test**
- **Foeto-Test**

The chart includes categories such as:
- Dates
- First
- Last
- Address
- Phone
- Town
- Hospital
- Zip

Each dataset is color-coded for easy differentiation.
Approaches

Rule-based approach

- Patterns + lists (*towns, first names, last names*) and dictionary (*250,000 inflected forms*)
- Three-step process:
  1. exact match with lists
  2. patterns: character properties, trigger words
  3. patterns: neighborhood of already de-identified entities
Approaches

Machine-learning based approach

- Linear chains CRF
- Features:
  - **surface features:** token, capitalization, digit, punctuation, length
  - **morpho-syntactic:** POS via Tree Tagger [Schmid, 1994]
  - **semantic types:** lexicon, CUI via UMLS [Lindberg et al., 1993]
  - **distributional analysis:** clustering via [Brown et al., 1992]
- No cross-validation
- Automatic feature selection: $\ell_1$ regularization
Evaluation

- **Recall** *(true positive rate)*

\[ R = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}} \]  \hspace{1cm} (1)

- **Precision** *(positive predictive value)*

\[ P = \frac{\text{true positives}}{\text{true positives} + \text{false positives}} \] \hspace{1cm} (2)

- **F-measure**

\[ F = \frac{(1 + \beta^2) \times P \times R}{\beta^2 \times P + R} \] \hspace{1cm} (3)

- **Confidence interval**

→ **Monte Carlo simulation** [Metropolis and Ulam, 1949]
1 Introduction

2 Materials and methods
   • Corpora
   • Approaches

3 Results

4 Discussion & conclusion

5 References
## Results

1. **Cardiology test**

<table>
<thead>
<tr>
<th></th>
<th>P</th>
<th>R</th>
<th>F</th>
<th>Confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rule-based</td>
<td>0.855</td>
<td>0.830</td>
<td>0.843</td>
<td>0.821/0.864</td>
</tr>
<tr>
<td>CRF</td>
<td>0.909</td>
<td>0.858</td>
<td>0.883</td>
<td>0.864/0.901</td>
</tr>
</tbody>
</table>

2. **Fœtopathology test**

<table>
<thead>
<tr>
<th></th>
<th>P</th>
<th>R</th>
<th>F</th>
<th>Confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rule-based</td>
<td>0.678</td>
<td>0.684</td>
<td>0.681</td>
<td>0.633/0.729</td>
</tr>
<tr>
<td>CRF</td>
<td>0.732</td>
<td>0.565</td>
<td>0.638</td>
<td>0.585/0.692</td>
</tr>
</tbody>
</table>
# Results

## Cardiology test (details)

<table>
<thead>
<tr>
<th></th>
<th>Rule-based</th>
<th>CRF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dates (238)</td>
<td>0.920</td>
<td>0.987</td>
</tr>
<tr>
<td>Last names (205)</td>
<td>0.903</td>
<td>0.892</td>
</tr>
<tr>
<td>First names (109)</td>
<td>0.777</td>
<td>0.822</td>
</tr>
<tr>
<td>Hospital (43)</td>
<td>0.500</td>
<td>0.931</td>
</tr>
<tr>
<td>Town (22)</td>
<td>0.688</td>
<td>0.632</td>
</tr>
<tr>
<td>Zip codes (8)</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>Phone (8)</td>
<td>1.000</td>
<td>0.857</td>
</tr>
<tr>
<td></td>
<td>Introduction</td>
<td>Materials and methods</td>
</tr>
<tr>
<td>---</td>
<td>--------------</td>
<td>-----------------------</td>
</tr>
<tr>
<td>1</td>
<td>Introduction</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Materials and methods</td>
<td>Corpora</td>
</tr>
<tr>
<td>3</td>
<td>Results</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Discussion &amp; conclusion</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>References</td>
<td></td>
</tr>
</tbody>
</table>
Discussion

Global results

- CRF performed better ($F=0.883$) than rule-based ($F=0.843$)
- **Rule-based**: performs better for easily formalized entities
  
  *numerical data: zip/phone $F=1.000$*

- **CRF**: given enough training examples (312 docs; 2,500 entities), variety of entities better learned *hospital $F=0.750$*
Discussion

Corpora comparison

- **Cardiology-test**: both systems designed to process this corpus
  - Confidence interval: ±2 points over F-measure
  - **Boundary errors**: rule-based (26 cases) vs. CRF (8 cases)
    → for date ranges and hospital names
  - **Categorization errors**: 20 cases for both systems
    → confusion between first names and last names

- **Fœtopathology**: promising results
  - Rule-based: F=0.681; CRF: F=0.638
  - Confidence interval: ±5 points over F-measure
  - **Boundary errors**: distinct time range and cultural references
    (number of digits in French phone numbers has changed)
Discussion

Comparison across corpora

- From cardiology to fœtopathology:
  - Rule-based: loss of 16 F-measure points
  - CRF: loss of 24 F-measure points

- [Ferrández et al., 2012] From VHA to i2b2 corpus: loss of 26 F-measure points
Discussion

Further work

- Better consideration of categories:
  - first name vs. last name: too much confusing
  - address: human definition

- Use of syntactic features for CRF: POS, chunk

- Combination of both rule-based and CRF systems
Tak for din opmærksomhed!
Husk panelet på anonymisering morgen 8:30
Class-based n-gram models of natural language.  
Generalizability and comparison of automatic clinical text de-identification methods and resources.  
In *AMIA Annu Symp Proc*, Chicago, IL. |
*Anonymisation de documents cliniques : performances et limites des méthodes symboliques et par apprentissage statistique.*  
| References | Lavergne, T., Cappé, O., and Yvon, F. (2010).  
Practical very large scale CRFs. |


