Building a Spanish/Catalan Health Records Corpus with Very Sparse Protected Information Labelled

LREC 2018

Salvador Medina and Jordi Turmo

Universitat Politècnica de Catalunya
Talp Research Center

Carrer de Jordi Girona, 1-3, 08034 Barcelona
{smedina, turmo}@cs.upc.edu

May 08, 2018
Overview

About this project

- **Build Health Record Corpora with labeled Protected Health Information**
  - Unstructured health notes
  - High sparsity of Protected Health Information
  - Multilingual: Spanish and Catalan

- Fetch and select examples by using manual rules
  - That can be defined and understood by non-programmers
  - Implemented using Augmented Transition Networks

- Iterative and interactive process
  - Inspired by active learning
  - New relevant examples are selected in each iteration
  - Rules are added or updated based on these new examples
Overview

About this project

- Build Health Record Corpora with labeled Protected Health Information
  - Unstructured health notes
  - High sparsity of Protected Health Information
  - Multilingual: Spanish and Catalan
- Fetch and select examples by using manual rules
  - That can be defined and understood by non-programmers
  - Implemented using Augmented Transition Networks
- Iterative and interactive process
  - Inspired by active learning
  - New relevant examples are selected in each iteration
  - Rules are added or updated based on these new examples
Overview

About this project

- Build Health Record Corpora with labeled Protected Health Information
  - Unstructured health notes
  - High sparsity of Protected Health Information
  - Multilingual: Spanish and Catalan
- Fetch and select examples by using manual rules
  - That can be defined and understood by non-programmers
  - Implemented using Augmented Transition Networks
- Iterative and interactive process
  - Inspired by active learning
  - New relevant examples are selected in each iteration
  - Rules are added or updated based on these new examples
Contents

1 Introduction
   ■ Overview
   ■ Motivation

2 Methodology
   ■ Introduction
   ■ Definition of Rules
   ■ The iterative method
   ■ Ranking and Selection of EHR

3 Evaluation
   ■ Evaluation Corpora
   ■ Evaluation Framework
   ■ Evaluation Results

4 Conclusions
Motivation

Available Corpora

Several Electronic Health Record (EHR) corpora for Protected Health Information (PHI) can be retrieved from multiple sources:

- **Shared Tasks**
  - 2006 and 2014 i2b2 Challenges [Uzuner et al., 2007]
    [Stubbs and Uzuner, 2015]
  - 2016 CEGS N-GRID Shared Tasks [Stubbs et al., 2017]
- **Re-purposed EHR corpora**
  - Intelligent Monitoring for Intensive Care (MIMIC-II)
    [Neamatzullah et al., 2008]

⇒ Most corpora is in **English**, multilingual corpora is needed
Motivation

Available Corpora

Several Electronic Health Record (EHR) corpora for Protected Health Information (PHI) can be retrieved from multiple sources:

- **Shared Tasks**
  - 2006 and 2014 *i2b2* Challenges [Uzuner et al., 2007] [Stubbs and Uzuner, 2015]
  - 2016 CEGS N-GRID Shared Tasks [Stubbs et al., 2017]

- **Re-purposed EHR corpora**
  - Intelligent Monitoring for Intensive Care (MIMIC-II) [Neamatullah et al., 2008]

⇒ **Most corpora is in English**, multilingual corpora is needed
Motivation

Regulations and directives

- Different countries have different regulations:
  - **Spain**: *Ley Orgánica de Protección de Datos*
  - **Colombia**: Constitution and laws 1273 and 1581
  - **Uruguay**: *Ley de Acceso a la Información Pública*

- Legislation imposes restrictions to
  - Who can access non-anonymized EHR
  - The kinds of entities that must be anonymized
  - The level of protection of different kinds of EHR

⇒ *Existing corpora may need to be adapted or extended*
Motivation

Regulations and directives

- Different countries have different regulations:
  - **Spain**: Ley Orgánica de Protección de Datos
  - **Colombia**: Constitution and laws 1273 and 1581
  - **Uruguay**: Ley de Acceso a la Información Pública

- Legislation imposes restrictions to
  - Who can access non-anonymized EHR
  - The kinds of entities that must be anonymized
  - The level of protection of different kinds of EHR

⇒ Existing corpora may need to be adapted or extended
Motivation

Manual labelling costs

- Health notes usually have a low density of PHI
  - In our corpus, $\sim 0.4\%$ of tokens are people’s names
- PHI classes are very unbalanced
  - In our corpus, $< 0.01\%$ of telephone numbers vs $\sim 1\%$ of locations
- Manual labelling should be consensuated among multiple experts
Contents

1 Introduction
   - Overview
   - Motivation

2 Methodology
   - Introduction
   - Definition of Rules
   - The iterative method
   - Ranking and Selection of EHR

3 Evaluation
   - Evaluation Corpora
   - Evaluation Framework
   - Evaluation Results

4 Conclusions
The Iterative Method

Basic ideas about the method

- Potential PHI in EHR are identified by using a set of rules
- Rules are implemented using Augmented Transition Networks (ATN)
- The rule set is iteratively updated
  - New rules are added
  - Existing ones are updated and grow in complexity
- New EHR are added to the training set in each iteration
Definition of Rules

Characteristics of the manual rules

- Rules are implemented using Augmented Transition Networks
- Phrases are parsed at token level using FreeLing 4.0
  [Padró and Stanilovsky, 2012] including:
  - Language detection
  - Tokenization
  - Lemmatization
  - POS Tagging
  - NER and multi-word detection are disabled
- Gazetteers and regular expressions can be checked
- Partial consumption of tokens is allowed (lAnna → l + Anna)
Definition of Rules
Example of a manual rule (I)

\[
\begin{align*}
&l \in G_{cv} & f \in G_{c vp} & u(f) \\
&POS = \text{Det} & p(f) \in G_d \\
\end{align*}
\]

Figure: Example of an ATN rule. \(l, f\) and POS stand for lemma, form and Part of Speech. \(p(f)\) means to partially consume form \(f\) and \(u(f)\) stands for uppercase. \(G_{cv}, G_{c vp}\) and \(G_d\) are specific gazetteers.
Definition of Rules

Example of a manual rule (II)

"Los derivo a bienestar social para **hablar** con Oliach." (I derive them to social wellness so as to talk to Oliach.)
"Los derivo a bienestar social para hablar con Oliach." (I derive them to social wellness so as to talk to Oliach.)
"Los derivo a bienestar social para hablar con Oliach." (I derive them to social wellness so as to talk to Oliach.)
"Parlo amb l’Anna de la pauta a seguir." (I talk to Anna about the guideline to follow.)
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Subsections</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Introduction</td>
<td>Overview, Motivation</td>
</tr>
<tr>
<td>2</td>
<td>Methodology</td>
<td>Introduction, Definition of Rules, The iterative method, Ranking and Selection of EHR</td>
</tr>
<tr>
<td>3</td>
<td>Evaluation</td>
<td>Evaluation Corpora, Evaluation Framework, Evaluation Results</td>
</tr>
<tr>
<td>4</td>
<td>Conclusions</td>
<td></td>
</tr>
</tbody>
</table>
The Iterative Method

1. Evaluate \( \{F_1, r, p\}(R_i, C_{tr,k}) \)

2. Repeat while \( \exists m \mid R_{i+1}^t = R_i + \{m\} \)
   \( \Rightarrow F_1(R_{i+1}^t, C_{tr,k}) > F_1(R_i, C_{tr,k}) \)
   - Evaluate \( \{F_1, r, p\}(R_{i+1}^t, C_{val}) \)
   - If \( F_1(R_{i+1}^t, C_{val}) > F_1(R_i, C_{val}) \) \( \Rightarrow R_{i+1} = R_{i+1}^t \)
   - If \( r(R_{i+1}^t, C_{val}) < r(R_i, C_{val}) \) \( \Rightarrow \text{discard}(m) \)
   - If \( p(R_{i+1}^t, C_{val}) < p(R_i, C_{val}) \) \( \Rightarrow \text{refine}(m) \)

3. \( \lambda_k = \text{elbow}(\{\text{score}(R_{i+n}[d]) \quad \forall d \in C_{unl,k}\}) \)

4. \( C_{tr,k+1} = C_{tr,k} + \{\text{label}(d)\} \)
   \( \forall d \in C_{unl,k} \mid \text{score}(R_{i+n}[d]) > \lambda_k \)
The Iterative Method

1. Evaluate \( \{F_1, r, p\}(R_i, C_{tr,k}) \)
2. Repeat while \( \exists m | R_{i+1}^t = R_i + \{m\} \)
   \[ \Rightarrow F_1(R_{i+1}^t, C_{tr,k}) > F_1(R_i, C_{tr,k}) \]
   - Evaluate \( \{F_1, r, p\}(R_{i+1}^t, C_{val}) \)
   - If \( F_1(R_{i+1}^t, C_{val}) > F_1(R_i, C_{val}) \) \( \Rightarrow R_{i+1} = R_i \)
   - If \( r(R_{i+1}^t, C_{val}) < r(R_i, C_{val}) \) \( \Rightarrow \text{discard}(m) \)
   - If \( p(R_{i+1}^t, C_{val}) < p(R_i, C_{val}) \) \( \Rightarrow \text{refine}(m) \)
3. \( \lambda_k = \text{elbow}(\{\text{score}(R_{i+n}[d]) \mid \forall d \in C_{unl,k}\}) \)
4. \( C_{tr,k+1} = C_{tr,k} + \{\text{label}(d)\} \)
   \( \forall d \in C_{unl,k} \mid \text{score}(R_{i+n}[d]) > \lambda_k \)
The Iterative Method

1. Evaluate \( \{F_1, r, p\}(R_i, C_{tr,k}) \)
2. Repeat while \( \exists m | R_{i+1}^t = R_i + \{m\} \)
   \[ \Rightarrow F_1(R_{i+1}^t, C_{tr,k}) > F_1(R_i, C_{tr,k}) \]
   - Evaluate \( \{F_1, r, p\}(R_{i+1}^t, C_{val}) \)
   - If \( F_1(R_{i+1}^t, C_{val}) > F_1(R_i, C_{val}) \) \( \Rightarrow R_{i+1} = R_{i+1}^t \)
   - If \( r(R_{i+1}^t, C_{val}) < r(R_i, C_{val}) \) \( \Rightarrow \text{discard}(m) \)
   - If \( p(R_{i+1}^t, C_{val}) < p(R_i, C_{val}) \) \( \Rightarrow \text{refine}(m) \)
3. \( \lambda_k = \text{elbow}(\{\text{score}(R_i+n[d]) \mid \forall d \in C_{unl,k}\}) \)
4. \( C_{tr,k+1} = C_{tr,k} + \{\text{label}(d)\} \)
   \[ \forall d \in C_{unl,k} \mid \text{score}(R_i+n[d]) > \lambda_k \]
Building a Spanish/Catalan Health Records Corpus with Very Sparse Protected Information Labelled

Methodology

The Iterative Method

1. Evaluate \( \{F_1, r, p\}(R_i, C_{tr,k}) \)

2. Repeat while \( \exists m | R_{i+1} = R_i + \{m\} \)

   \[ \Rightarrow F_1(R_{i+1}, C_{tr,k}) > F_1(R_i, C_{tr,k}) \]

   - Evaluate \( \{F_1, r, p\}(R_{i+1}, C_{val}) \)
   - If \( F_1(R_{i+1}, C_{val}) > F_1(R_i, C_{val}) \) \( \Rightarrow R_{i+1} = R_{i+1} \)
   - If \( r(R_{i+1}, C_{val}) < r(R_i, C_{val}) \) \( \Rightarrow \text{discard}(m) \)
   - If \( p(R_{i+1}, C_{val}) < p(R_i, C_{val}) \) \( \Rightarrow \text{refine}(m) \)

3. \( \lambda_k = \text{elbow}(\{\text{score}(R_{i+n}[d]) \; \forall d \in C_{unl,k}\}) \)

4. \( C_{tr,k+1} = C_{tr,k} + \{\text{label}(d)\} \)

   \[ \forall d \in C_{unl,k} \mid \text{score}(R_{i+n}[d]) > \lambda_k \} \)
**The Iterative Method**

1. Evaluate $\{F_1, r, p\}(R_i, C_{tr,k})$
2. Repeat while $\exists m \mid R_{i+1}^t = R_i + \{m\}$
   - $F_1(R_{i+1}^t, C_{tr,k}) > F_1(R_i, C_{tr,k})$
   - Evaluate $\{F_1, r, p\}(R_{i+1}^t, C_{val})$
   - If $F_1(R_{i+1}^t, C_{val}) > F_1(R_i, C_{val}) \Rightarrow R_{i+1} = R_i$
   - If $r(R_{i+1}^t, C_{val}) < r(R_i, C_{val}) \Rightarrow discard(m)$
   - If $p(R_{i+1}^t, C_{val}) < p(R_i, C_{val}) \Rightarrow refine(m)$
3. $\lambda_k = elbow(\{score(R_{i+n}[d]) \mid \forall d \in C_{unl,k}\})$
4. $C_{tr,k+1} = C_{tr,k} + \{label(d)\}$
   - $\forall d \in C_{unl,k} \mid score(R_{i+n}[d]) > \lambda_k$
The Iterative Method

Ranking and selection of EHR: Scoring Function

Documents are scored and ranked using the following scoring function:

\[
\text{score}(d) = \sum_{i \in K} N_i(d) \times (1 - F_1(i)) \times (1 - p_i)
\]

where

\[
p_i = \frac{\sum_{t \in T} N_i(t)}{\sum_{i \in K} \sum_{t \in T} N_i(t)}
\]
The Iterative Method

Ranking and selection of EHR: Threshold Score

# of Documents: Elbow Criterion

Threshold score is the one that corresponds to the *elbow* point of the curve defined by the document’s scores sorted in decreasing order.

**Figure**: Schematic representation of the *elbow* point of an exponential function.
The Iterative Method

Observations

- Prioritizes rules that increase *recall* while $F_1$ is not decreased
- $F_1$ increases monotonically
- Can be applied indefinitely
- Entities of uncommon classes are prioritized
- Documents with no entities are not selected
1 Introduction
   ■ Overview
   ■ Motivation

2 Methodology
   ■ Introduction
   ■ Definition of Rules
   ■ The iterative method
   ■ Ranking and Selection of EHR

3 Evaluation
   ■ Evaluation Corpora
   ■ Evaluation Framework
   ■ Evaluation Results

4 Conclusions
Evaluation Corpora
Characteristics of the evaluation corpora

- We use the Institut Català de la Salut (ICS) Primary Care Service’s corpus of 2011
- Written in Spanish and Catalan, often mixed
- Includes admission, progress, operative and discharge notes
- Cover multiple clinical fields: common illnesses, psychology, dependency, drug use...
Incoherent use of capitalization

“realitzarem inmovilitzаци́o, recomanen e insisteim anar aH DE CALELLA PER CONFIRMAR FISURA I FRACTURA, DIU QUE NO HI ANIRÀ QUE NO VOL ESPERAR-SE 4 H.P: Realitzem inmovilitzаци́o i control en una setmana.” combines fully lowercased phases with fully uppercased ones.

Use of contractions

“Pac que finaliza tto”, where the words Pac and tto are used instead of Paciente (patient) and tratamiento (treatment).
Use of punctuation marks instead of spaces or lack of them

“Algun subcrepitante en bases...Normas.Pulmicort-100 2-1(15 dias).”, the words bases, Normas and Pulmicort-100 are not spaced. What is more, in sentence “Controlada HVhebron anualment.”, HVhebron should be H. V. Hebron, as it refers to Hospital Vall Hebron.

Enumerations of measures and readings from medical analysis

“Usa L/C OD 85°-0.50 +1.00 0.8 /+4.00. OI 115°-1.00 +0.25 0.9 /+3.50.AO 4DP BT en VL.Rx ¿OD NG. OI NG Ad/3.00.”
Inconsistent use of languages, since notes often combine Spanish and Catalan words, phrases or idioms

“M: febre de 39°C tot el dia **a pesar que** la mare li ha donat Dalsy, vomits i mucositat nasal.” is written in Catalan but includes the Spanish expression **a pesar que** (despite of), while sentence “E:herida mordida palma de mano D.P: **neteja**, steri-strip...” is written in Spanish but uses the Catalan verb **neteja** (to clean).
## Evaluation Corpora

Number of entities per PHI category

<table>
<thead>
<tr>
<th>Category</th>
<th>Validation</th>
<th>Test</th>
<th>Resulting Corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>person</td>
<td>372</td>
<td>282</td>
<td>699</td>
</tr>
<tr>
<td>location</td>
<td>99</td>
<td>680</td>
<td>825</td>
</tr>
<tr>
<td>telephone</td>
<td>7</td>
<td>6</td>
<td>17</td>
</tr>
<tr>
<td>Notes</td>
<td>311</td>
<td>5000</td>
<td>1051</td>
</tr>
<tr>
<td>Notes with PHI</td>
<td>299</td>
<td>667</td>
<td>793</td>
</tr>
</tbody>
</table>

Table: Count of instances of PHI corresponding to categories **PERSON**, **LOCATION** and **TELEPHONE** in corpora. Categories **TELEPHONE**, **EMAIL**, **DNI**, **SOCIAL_SECURITY_ID** and **SANITARY_CARD_ID** are excluded. Validation corpus only includes EHR with PHI.
Contents

1 Introduction
   ■ Overview
   ■ Motivation

2 Methodology
   ■ Introduction
   ■ Definition of Rules
   ■ The iterative method
   ■ Ranking and Selection of EHR

3 Evaluation
   ■ Evaluation Corpora
   ■ Evaluation Framework
   ■ Evaluation Results

4 Conclusions
Evaluation Framework

Direct and Indirect Evaluation

Direct Evaluation

**Goal:** Make the manual labeling process *cheaper*
- Evaluate using $F_1$ score achieved by the rule set
- Partial evaluation for boundary identification

Indirect Evaluation

**Goal:** Improve the resulting corpus
- Evaluate using $F_1$ score achieved by a tagger trained using the resulting corpus
- Strict evaluation for boundary identification
Evaluation Results

Direct evaluation over each Iteration: Training

Figure: Evolution of precision, recall and $F_1$ score in the final training corpus over each iteration.
Evaluation Results
Direct evaluation over each Iteration: Validation

Figure: Evolution of precision, recall and $F_1$ score in the validation corpus over each iteration.
Evaluation Results
Final direct Evaluation

<table>
<thead>
<tr>
<th></th>
<th>Eval.</th>
<th>FreeLing</th>
<th>Ruleset</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ALL</strong></td>
<td>Recall</td>
<td>0.494</td>
<td>0.702</td>
</tr>
<tr>
<td></td>
<td>Prec.</td>
<td>0.052</td>
<td>0.489</td>
</tr>
<tr>
<td></td>
<td>$F_1$</td>
<td>0.094</td>
<td>0.576</td>
</tr>
<tr>
<td><strong>PERSON</strong></td>
<td>Recall</td>
<td>0.436</td>
<td>0.772</td>
</tr>
<tr>
<td></td>
<td>Prec.</td>
<td>0.023</td>
<td>0.445</td>
</tr>
<tr>
<td></td>
<td>$F_1$</td>
<td>0.044</td>
<td>0.564</td>
</tr>
<tr>
<td><strong>LOCATION</strong></td>
<td>Recall</td>
<td>0.517</td>
<td>0.371</td>
</tr>
<tr>
<td></td>
<td>Prec.</td>
<td>0.064</td>
<td>0.809</td>
</tr>
<tr>
<td></td>
<td>$F_1$</td>
<td>0.114</td>
<td>0.509</td>
</tr>
</tbody>
</table>

**Table**: Evaluation results in the test set for the general-purpose *FreeLing* NERC module, and for the final set of handcrafted rules.
## Evaluation Results

Final indirect evaluation

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ALL</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Recall</td>
<td>0.721 ± 0.027</td>
<td>0.699 ± 0.042</td>
<td></td>
</tr>
<tr>
<td>Prec.</td>
<td>0.839 ± 0.026</td>
<td>0.769 ± 0.047</td>
<td></td>
</tr>
<tr>
<td>$F_1$</td>
<td>0.774 ± 0.017</td>
<td>0.732 ± 0.039</td>
<td></td>
</tr>
<tr>
<td><strong>PERSON</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Recall</td>
<td>0.784 ± 0.064</td>
<td>0.759 ± 0.093</td>
<td></td>
</tr>
<tr>
<td>Prec.</td>
<td>0.909 ± 0.041</td>
<td>0.730 ± 0.061</td>
<td></td>
</tr>
<tr>
<td>$F_1$</td>
<td>0.840 ± 0.025</td>
<td>0.744 ± 0.057</td>
<td></td>
</tr>
<tr>
<td><strong>LOCATION</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Recall</td>
<td>0.695 ± 0.040</td>
<td>0.676 ± 0.056</td>
<td></td>
</tr>
<tr>
<td>Prec.</td>
<td>0.812 ± 0.022</td>
<td>0.783 ± 0.061</td>
<td></td>
</tr>
<tr>
<td>$F_1$</td>
<td>0.748 ± 0.037</td>
<td>0.726 ± 0.052</td>
<td></td>
</tr>
</tbody>
</table>

**Table:** Mean recall, precision and $F_1$ score obtained by a CRF model trained using the labelled corpus obtained after 3 iterations of the method (1051 health records) compared to the 8-fold cross validation of the test corpus (4350 health records) for the 8 testing partitions. Standard deviation is shown between brackets.
Summary

- We describe a method to build a manually labelled corpus
  - Optimized sparsely populated datasets
  - Retrieval of new examples is based on manual rules
  - Selected examples are manually labeled
  - Rules are iteratively defined or refined

- We created a bilingual Spanish/Catalan EHR corpus for PHI detection

- We evaluated the resulting corpus
  - Direct evaluation: quality of the manual rule-set
  - Indirect evaluation: quality of the resulting corpus
Conclusions

When compared to traditional manually built corpora

- The iteratively built corpus can provide similar results for PHI tasks
- Lower manual labelling effort is required for sparse datasets
- Medical staff can more easily understand and define the fetching rules
Thank you for your attention!
Questions?
References I


References II
