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

Contents

1. Introduction
   - Overview
   - Motivation

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

3. Evaluation
   - Evaluation Corpora
   - Evaluation Framework
   - Evaluation Results

4. Conclusions

Contents

1 Introduction
- Overview
- Motivation

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

3 Evaluation
- Evaluation Corpora
- Evaluation Framework
- Evaluation Results

4 Conclusions

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 Methodoloty

- 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) [Neamatullah 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
  - **Urugay:** *Ley de Acceso a la Información Pública*
- Legislation imposes restrictions to
  - Who can access non-anonyzed EHR
  - The kinds of entities that must be anonymized
  - The level of protection of different kinds of EHR

⇒ Existing corpora may need to be addapted 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
  - **Urugay:** *Ley de Acceso a la Información Pública*
- Legislation imposes restrictions to
  - Who can access non-anonyzed EHR
  - The kinds of entities that must be anonymized
  - The level of protection of different kinds of EHR

⇒ Existing corpora may need to be addapted 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 Methodoloty
- 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

Contents

1. Introduction
    - Overview
    - Motivation
2. Methodoloty
    - Introduction
    - Definition of Rules
    - The iterative method
    - Ranking and Selection of EHR
3. Evaluation
    - Evaluation Corpora
    - Evaluation Framework
    - Evaluation Results
4. Conclusions

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 ($\mathrm{lAnna} \rightarrow \mathrm{l} + \mathrm{Anna}$)

Definition of Rules

Example of a manual rule (I)

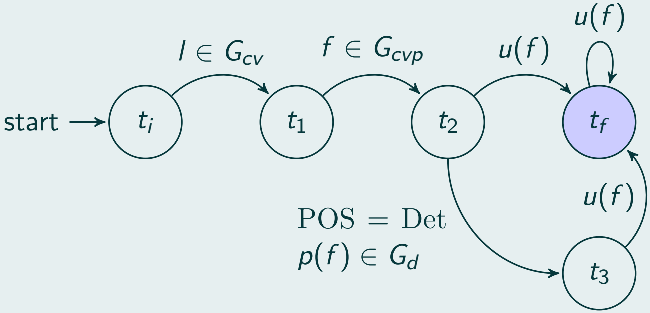
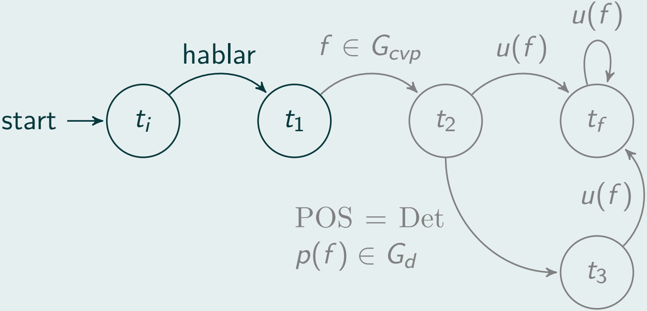


Figure: Example of an ATN rule. l, f and POS stand for lemma, form and Part of Speech. $p(f)$ means to partially consume form f and $u(f)$ stands for uppercase. G_{cv}, G_{cvp} and G_d are specific gazetteers.

Definition of Rules

Example of a manual rule (II)

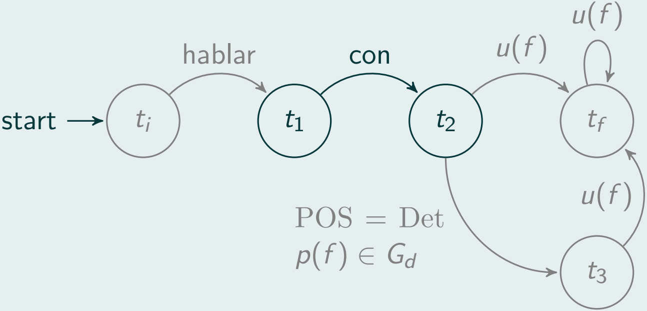
"Los derivo a bienestar social para **hablar** con Oliach." (I derive them to social wellness so as to talk to Oliach.)



Definition of Rules

Example of a manual rule (II)

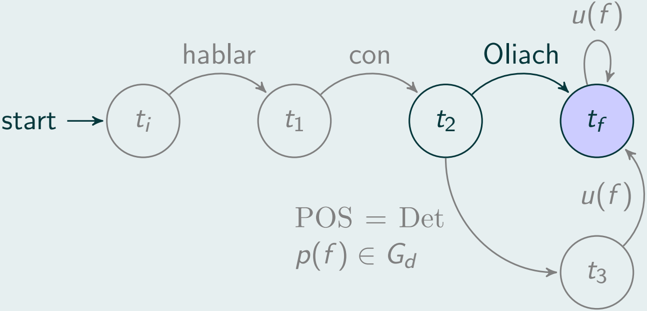
"Los derivo a bienestar social para hablar **con** Oliach." (I derive them to social wellness so as to talk to Oliach.)



Definition of Rules

Example of a manual rule (II)

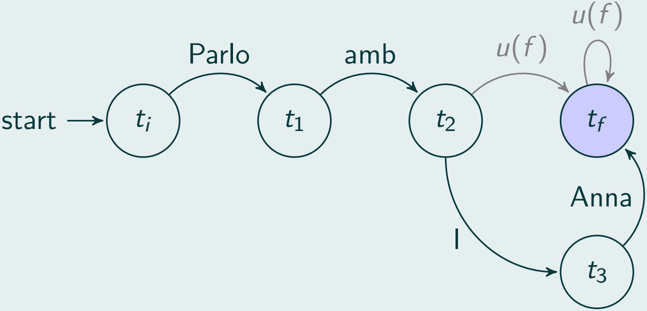
"Los derivo a bienestar social para hablar con **Oliach**." (I derive them to social wellness so as to talk to Oliach.)



Definition of Rules

Example of a manual rule (III)

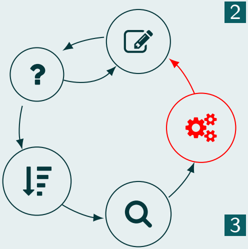
"**Parlo amb lAnna** de la pauta a seguir." (I talk to Anna about the guideline to follow.)



Contents

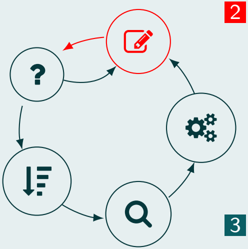
1. Introduction
   - Overview
   - Motivation
2. Methodoloty
   - Introduction
   - Definition of Rules
   - The iterative method
   - Ranking and Selection of EHR
3. Evaluation
   - Evaluation Corpora
   - Evaluation Framework
   - Evaluation Results
4. Conclusions

The Iterative Method



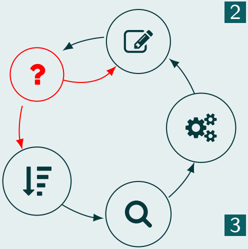
1. Evaluate $\{F_1, r, p\}(R_i, C_{tr,k})$
2. Repeat while $\exists m | R^t_{i+1} = R_i + \{m\}$
   $\Rightarrow F_1(R^t_{i+1}, C_{tr,k}) > F_1(R_i, C_{tr,k})$
   - Evaluate $\{F_1, r, p\}(R^t_{i+1}, C_{val})$
   - If $F_1(R^t_{i+1}, C_{val}) > F_1(R_i, C_{val}) \Rightarrow R_{i+1} = R^t_{i+1}$
   -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

The Iterative Method



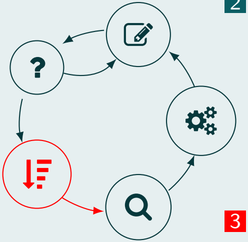
1. Evaluate $\{F_1, r, p\}(R_i, C_{tr,k})$
2. Repeat while $\exists m | R^t_{i+1} = R_i + \{m\}$ $\Rightarrow F_1(R^t_{i+1}, C_{tr,k}) > F_1(R_i, C_{tr,k})$
    - Evaluate $\{F_1, r, p\}(R^t_{i+1}, C_{val})$
    - If $F_1(R^t_{i+1}, C_{val}) > F_1(R_i, C_{val}) \Rightarrow R_{i+1} = R^t_{i+1}$
    -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

The Iterative Method



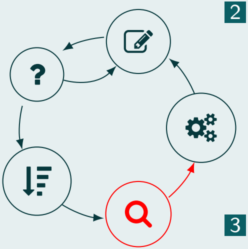
1. Evaluate $\{F_1, r, p\}(R_i, C_{tr,k})$
2. Repeat while $\exists m | R^t_{i+1} = R_i + \{m\}$
   $\Rightarrow F_1(R^t_{i+1}, C_{tr,k}) > F_1(R_i, C_{tr,k})$
   - Evaluate $\{F_1, r, p\}(R^t_{i+1}, C_{val})$
   - If $F_1(R^t_{i+1}, C_{val}) > F_1(R_i, C_{val}) \Rightarrow R_{i+1} = R^t_{i+1}$
   -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

The Iterative Method



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

The Iterative Method



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

Contents

1. Introduction
   - Overview
   - Motivation
2. Methodoloty
   - 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

Ranking and selection of EHR: Scoring Function

Documents are scored and ranked using the following scoring function:

$$score(d) = \sum_{i \epsilon K} N_i(d) * (1 - F_1(i)) * (1 - p_i)$$

$$p_i = \frac{\sum_{t \epsilon T} N_i(t)}{\sum_{i \epsilon K} \sum_{t \epsilon T} N_i(t)} \quad (1)$$

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

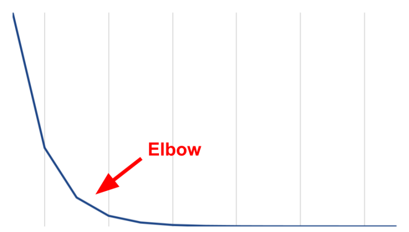


Figure: Schematic representation of the *elbow* point of an exponential function

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

Contents

1 Introduction
- Overview
- Motivation

2 Methodoloty
- 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...

Evaluation Corpora

Characteristics of the evaluation corpora (II)

Incoherent use of capitalization

"realitzarem innmovilitzaació, 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 inmovilització i control en una seetmana."* 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).

Evaluation Corpora

Characteristics of the evaluation corpora (II)

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 85º-0.50 +1.00 0.8 /+4.00. OI 115º-1.00 +0.25 0.9 /+3.50.AO 4DP BT en VL.Rx ¿OD NG. OI NG Ad/3.00.”

Evaluation Corpora

Characteristics of the evaluation corpora (III)

Inconsistent use of languages, since notes often combine Spanish and Catalan words, phrases or idioms

"M:febre de 39ºC 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

	Validation	Test	Resulting Corpus
PERSON	372	282	699
LOCATION	99	680	825
TELEPHONE	7	6	17
Notes	311	5000	1051
Notes with PHI	299	667	793

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 Methodoloty
- 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

Contents

1 Introduction
- Overview
- Motivation

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

3 Evaluation
- Evaluation Corpora
- Evaluation Framework
- Evaluation Results

4 Conclusions

Evaluation Results

Direct evaluation over each Iteration: Training

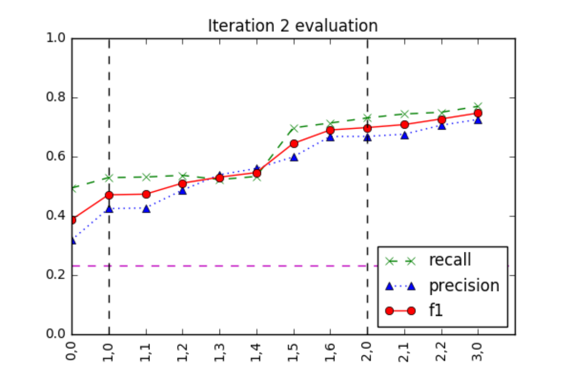


Figure: Evolution of precision, recall and F_1 score in the final training corpus over each iteration

Evaluation Results

Direct evaluation over each Iteration: Validation

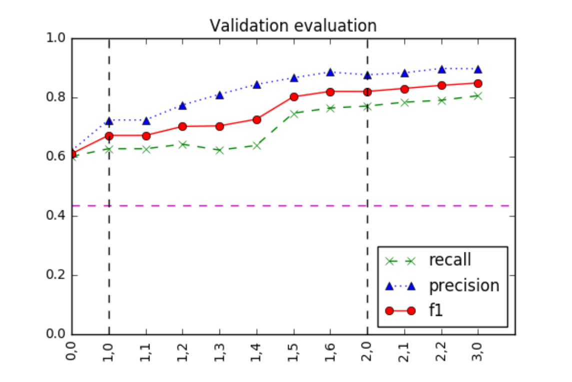


Figure: Evolution of precision, recall and F_1 score in the validation corpus over each iteration.

Evaluation Results

Final direct Evaluation

	Eval.	FreeLing	Ruleset
ALL	Recall	0.494	0.702
	Prec.	0.052	0.489
	F_1	0.094	0.576
PERSON	Recall	0.436	0.772
	Prec.	0.023	0.445
	F_1	0.044	0.564
LOCATION	Recall	0.517	0.371
	Prec.	0.064	0.809
	F_1	0.114	0.509

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

	Eval.	Cross-Val.	Res. Corpus
ALL	Recall	0.721 ± 0.027	0.699 ± 0.042
	Prec.	0.839 ± 0.026	0.769 ± 0.047
	F_1	0.774 ± 0.017	0.732 ± 0.039
PERSON	Recall	0.784 ± 0.064	0.759 ± 0.093
	Prec.	0.909 ± 0.041	0.730 ± 0.061
	F_1	0.840 ± 0.025	0.744 ± 0.057
LOCATION	Recall	0.695 ± 0.040	0.676 ± 0.056
	Prec.	0.812 ± 0.022	0.783 ± 0.061
	F_1	0.748 ± 0.037	0.726 ± 0.052

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.

Contents

1 Introduction
- Overview
- Motivation

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

3 Evaluation
- Evaluation Corpora
- Evaluation Framework
- Evaluation Results

4 Conclusions

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

Neamatullah, I., Douglass, M. M., Li-wei, H. L., Reisner, A., Villarroel, M., Long, W. J., Szolovits, P., Moody, G. B., Mark, R. G., and Clifford, G. D. (2008).
Automated de-identification of free-text medical records.
BMC medical informatics and decision making, 8(1):32.

Padró, L. and Stanilovsky, E. (2012).
Freeling 3.0: Towards wider multilinguality.
In *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*, Istanbul, Turkey. ELRA.

Stubbs, A., Filannino, M., and Uzuner, Ö. (2017).
De-identification of psychiatric intake records: Overview of 2016 cegs n-grid shared tasks track 1.
Journal of Biomedical Informatics.

References II

Stubbs, A. and Uzuner, Ö. (2015).
Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/uthealth corpus.
Journal of biomedical informatics, 58:S20–S29.

Uzuner, Ö., Luo, Y., and Szolovits, P. (2007).
Evaluating the state-of-the-art in automatic de-identification.
Journal of the American Medical Informatics Association, 14(5):550–563.