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

Contents

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
- Identification based on manual rules
  - Rules defined by non-programmers
  - Rules can be understood by health experts
  - Implemented using Augmented Transition Networks
- Iterative and interactive process:
  - Inspired by active learning
  - Expert evaluator defines new rules each iteration
  - The algorithm selects relevant examples to build such rules

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
- Identification based on manual rules
  - Rules defined by non-programmers
  - Rules can be understood by health experts
  - Implemented using Augmented Transition Networks
- Iterative and interactive process:
  - Inspired by active learning
  - Expert evaluator defines new rules each iteration
  - The algorithm selects relevant examples to build such rules

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
- Identification based on manual rules
  - Rules defined by non-programmers
  - Rules can be understood by health experts
  - Implemented using Augmented Transition Networks
- Iterative and interactive process:
  - Inspired by active learning
  - Expert evaluator defines new rules each iteration
  - The algorithm selects relevant examples to build such rules

Contents

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]

$\Rightarrow$ Most corpora is in **English**, multi-lingual 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**, multi-lingual corpora is needed

Motivation

Regulations and directives

- Different countries have different regulations:
  - **Spain:** *Ley Orgánica de Protección de Datos*
- 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

Motivation

Regulations and directives

- Different countries have different regulations:
  - **Spain:** *Ley Orgánica de Protección de Datos*
- 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

Motivation

Manual labelling costs

- Health notes usually have a very 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
- Labelling has to be done by experts
- Labels should be consensuated among various evaluators

Motivation

Manual labelling costs

- Health notes usually have a very 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
- Labelling has to be done by experts
- Labels should be consensuated among various evaluators

Motivation

Manual labelling costs

- Health notes usually have a very 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
- Labelling has to be done by experts
- Labels should be consensuated among various evaluators

Contents

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

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

The Iterative Method

Basic ideas about the method

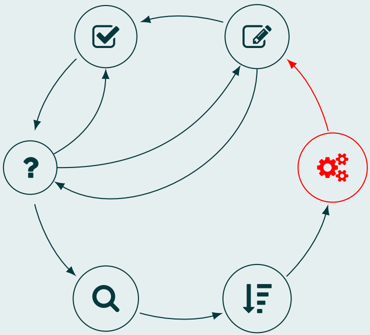
- Potential PHI in EHR are identified by using a set of rules
- Rules are implemented using Augmented Transition Networks (ATN)
- The rule set is iteratively updated
  - New rules are added
  - Existing ones are updated and grow in complexity
- New EHR are added to the training set in each iteration

The Iterative Method

Basic ideas about the method

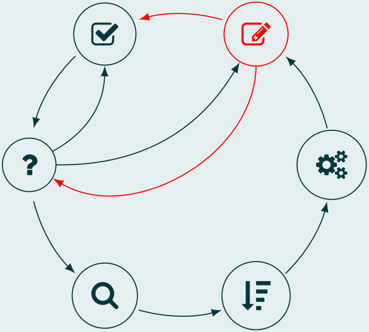
- Potential PHI in EHR are identified by using a set of rules
- Rules are implemented using Augmented Transition Networks (ATN)
- The rule set is iteratively updated
  - New rules are added
  - Existing ones are updated and grow in complexity
- New EHR are added to the training set in each iteration

The Iterative Method



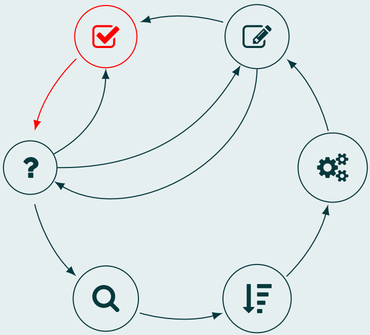
1. Run against C_{tr}
2. Define new rule so that $F_1'(C_{tr}) \geq F_1(C_{tr})$
3. Evaluate against C_{val}
4. Repeat from 2 unless
   - $r' \leq r$ and $F_1' < F_1 \Rightarrow$ Discard rule
   - $p' < p$ and $F_1' < F_1 \Rightarrow$ Update rule ($p' > p$)
5. Run against C_{unl}
6. Rank and select $f(C_{unl})$

The Iterative Method



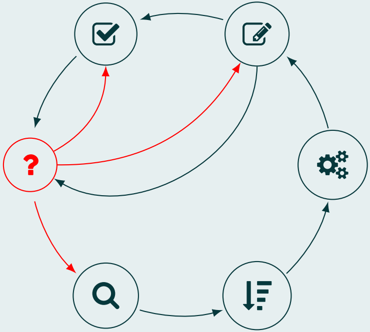
1. Run against C_{tr}
2. Define new rule so that $F_1'(C_{tr}) \geq F_1(C_{tr})$
3. Evaluate against C_{val}
4. Repeat from 2 unless
   - $r' \leq r$ and $F_1' < F_1 \Rightarrow$ Discard rule
   - $p' < p$ and $F_1' < F_1 \Rightarrow$ Update rule ($p' > p$)
5. Run against C_{unl}
6. Rank and select $f(C_{unl})$

The Iterative Method



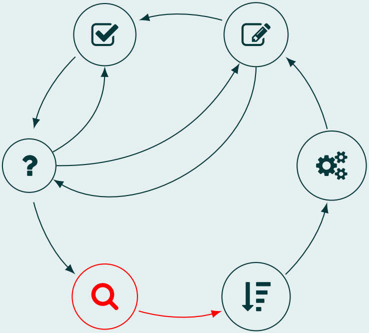
1. Run against C_{tr}
2. Define new rule so that $F_1'(C_{tr}) \geq F_1(C_{tr})$
3. Evaluate against C_{val}
4. Repeat from 2 unless
   - $r' \leq r$ and $F_1' < F_1$ $\Rightarrow$ Discard rule
   - $p' < p$ and $F_1' < F_1$ $\Rightarrow$ Update rule ($p' > p$)
5. Run against C_{unl}
6. Rank and select $f(C_{unl})$

The Iterative Method



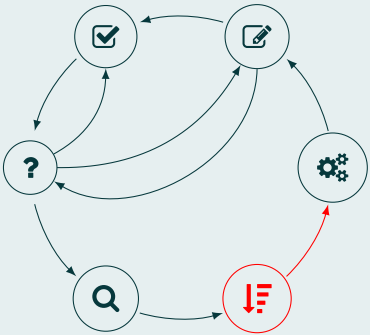
1. Run against C_{tr}
2. Define new rule so that $F_1'(C_{tr}) \geq F_1(C_{tr})$
3. Evaluate against C_{val}
4. Repeat from 2 unless
   - $r' \leq r$ and $F_1' < F_1 \quad \Rightarrow$ Discard rule
   - $p' < p$ and $F_1' < F_1 \quad \Rightarrow$ Update rule ($p' > p$)
5. Run against C_{unl}
6. Rank and select $f(C_{unl})$

The Iterative Method



1. Run against C_{tr}
2. Define new rule so that $F_1'(C_{tr}) \geq F_1(C_{tr})$
3. Evaluate against C_{val}
4. Repeat from 2 unless
   - $r' \leq r$ and $F_1' < F_1 \quad \Rightarrow$ Discard rule
   - $p' < p$ and $F_1' < F_1 \quad \Rightarrow$ Update rule ($p' > p$)
5. Run against C_{unl}
6. Rank and select $f(C_{unl})$

The Iterative Method



1. Run against C_{tr}
2. Define new rule so that $F_1'(C_{tr}) \geq F_1(C_{tr})$
3. Evaluate against C_{val}
4. Repeat from 2 unless
   - $r' \leq r$ and $F_1' < F_1 \Rightarrow$ Discard rule
   - $p' < p$ and $F_1' < F_1 \Rightarrow$ Update rule ($p' > p$)
5. Run against C_{unl}
6. Rank and select $f(C_{unl})$

The Iterative Method

Ranking and selection of EHR

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

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

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

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

Evaluation Framework

Direct and Indirect Evaluation

Direct Evaluation

Goal: Optimize the manual labeling process

- Evaluate using F_1 score achieved by the rule set
- Partial evaluation for boundary identification

Indirect Evaluation

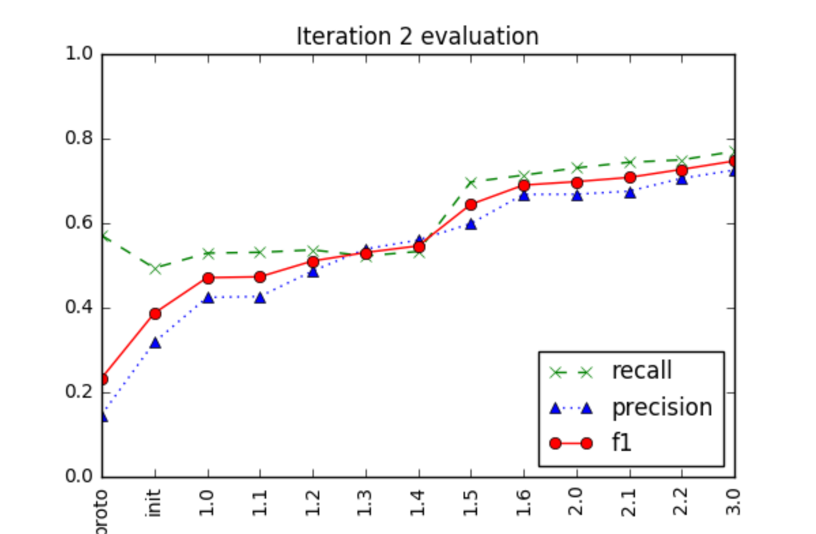
Goal: Optimize the resulting corpus

- Evaluate using F_1 score achieved by a tagger trained using the resulting corpus
- Strict evaluation for boundary identification

Contents

Evaluation Results

Direct evaluation over each Iteration



Evaluation Results

Final direct Evaluation

	Eval.	NERC	initial	final
ALL	Recall	0.052	0.147	0.702
	Prec.	0.494	0.208	0.489
	F_1	0.094	0.172	0.576
PERSON	Recall	0.436	0.676	0.772
	Prec.	0.023	0.196	0.445
	F_1	0.044	0.304	0.564
LOCATION	Recall	0.517	0.013	0.371
	Prec.	0.064	0.127	0.809
	F_1	0.114	0.024	0.509

Table: Evaluation results in the test set for the general-purpose *Freeling* NERC module, and for the initial and final sets of hand-crafted 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.

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.

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.

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.

References II

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.

Thank you for your attention!

Questions?