Combining Knowledge- and Data-driven Methods for De-identification of Clinical Narratives

Azad Dehghan[1], A. Kovačević[2], G. Karystianis[1], J A. Keane[1,4] and G Nenadic[1,3,4]
[1]School of Computer Science, University of Manchester, Manchester, UK
[2]Faculty of Technical Sciences, University of Novi Sad, Novi Sad, Serbia
[3]Health eResearch Centre, Manchester
[4]Manchester Institute of Biotechnoloy, University of Manchester

Contact: a.dehghan@manchester.ac.uk

Abstract

- The Problem
  - De-identification of clinical narratives
  - Personal health information: HIPAA
- Methods
  - Knowledge- and data-driven
- Results
  - F_1: (strict)90.65%;(lenient)94.80%;(HIPAA) 93.25%

MANCHESTER 1824

The University of Manchester

Task I: De-identification

- 2014 i2b2/UTHealth, Task I
- 25 entity types, 7 categories
  - AGE; DATE; CONTACT; LOCATION; ID; NAME; PROFESSION
- ~Longitudinal clinical narratives
  - Training: 790 and Test: 514 narratives

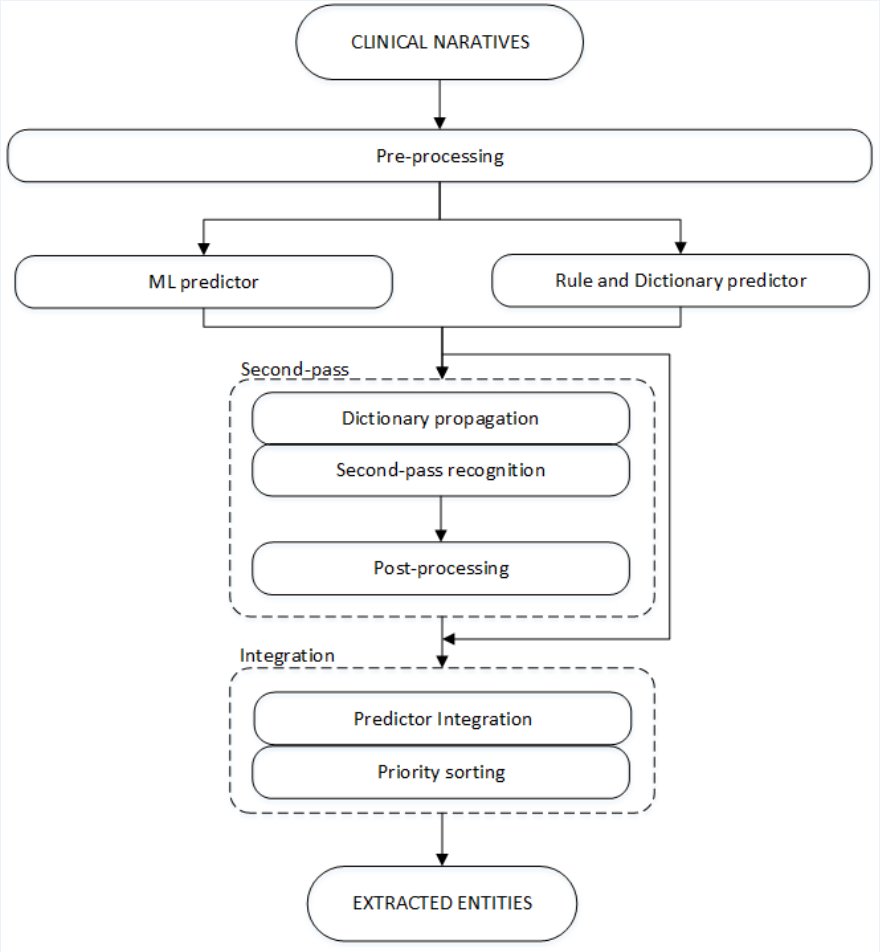
MANCHESTER 1824

The University of Manchester

Methodology

- Knowledge- and data-driven
  - Dictionary
  - Rule
  - machine learning
- Two-pass recognition
- Priority sorting

Methodology



ML predictor

- Initially, constructed models for all entity types
- Post validation: *City*, *Date*, *Patient*, *Hospital*, *Organization* and *Profession*
- *Six separate CRFs*
  - *280 features*
  - *Token-level CRF*
  - *Inside-Outside (I-O) schema*

Feature vector

- Lexical
  - Token, lemma, POS tag
  - Lemma and POS tags of surrounding tokens
  - Token location within the chunk (I-O)
- Orthographic
  - UpperInitial, allCaps, containNumber
  - Word pattern: BrightPoint -> “XxxxxxXxxxx”

MANCHESTER 1824

The University of Manchester

Feature vector

- Semantic
  - Dictionary and rules
    - US states and cities; calendar months; profession + cues:(e.g., “worked for”; “job as a”; “employed as”)
- Positional
  - Absolute line position containing the current token
  - Binary feature: presence of space character between current and next token

MANCHESTER
1824

The University of Manchester

Dictionary based predictor

- Lists or Gazetteers
- Longest match
- Post-processing rules (disambiguation)
- Sources: Internet, Wikipedia, deid tool
- **Dictionary for:** *Hospital*, *City*, *Country*, *Profession* and *Organization*

Rule based predictor

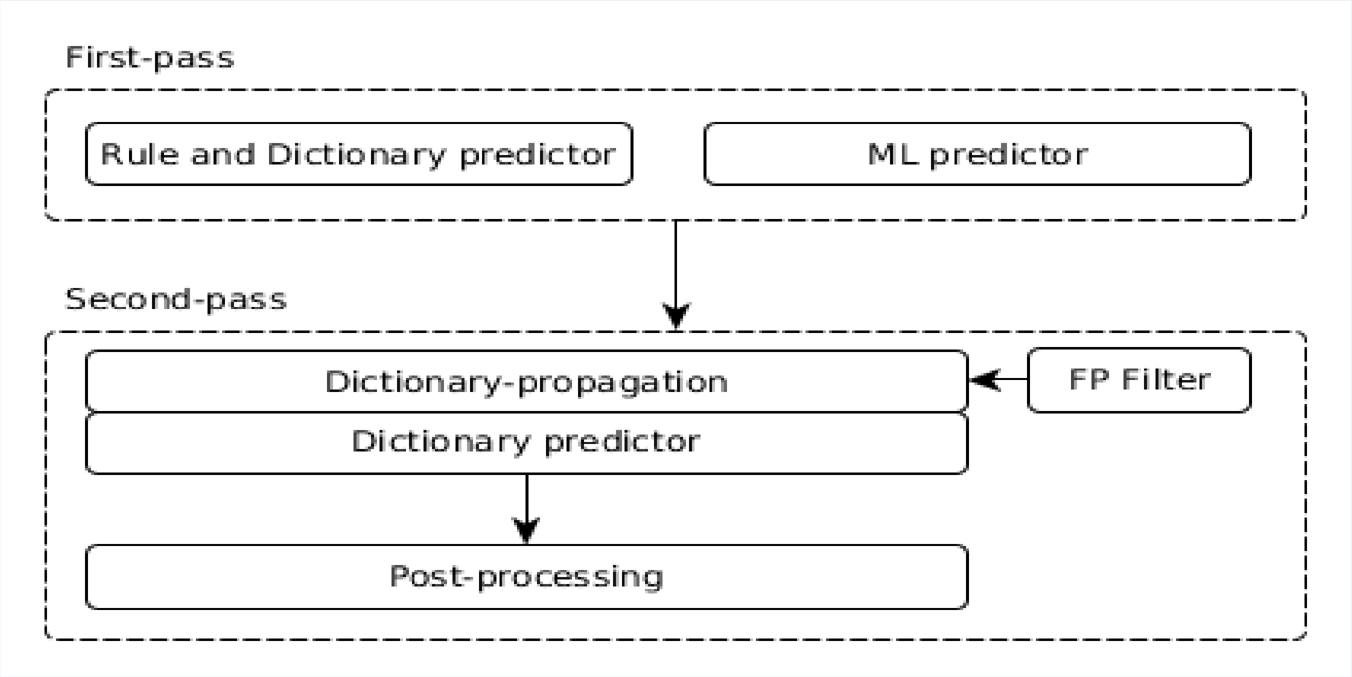
- CPSL and RegEx
- Largest number of predictors
- Small rule set; average: 5.6 rules / entity type
- **Exploited features:** orthographic, pattern, and contextual

Features

- Orthographic
  - upperInitial; allLowerCase; allCaps; Number; Punctuation
- Pattern
  - e.g., Date; Zip; Tel.
- Contextual cues
  - doctor and person titles (e.g., “Dr”; “Mr”)
  - symbols (e.g., brackets [*Username]*)
  - White space characters

Two-pass recognition

- Capture repeated information that lack contextual cues



Two-pass recognition

- Input:
  - "Mr John complained of a recurring upper chest pain; I have referred John to …."

1. First-pass recognition:
   - "Mr John complained of a recurring upper chest pain; I have referred John to …."
2. Second-pass recognition:
   - Dictionary-propagation: "John"
   - Dictionary matching…

- Output:
  - "Mr John complained of a recurring upper chest pain; I have referred John to …."

Two-pass recognition

- Rule and ML predictors
  - **Rule:** *Patient*, *Doctor*, *Date*, *Zip*, *Medicalrecord*, and *Idnum*
  - **ML:** *City*, *Hospital* and *Patient*
- **Improvements** (F_1):
  – e.g., Patient(~5%)
  – e.g., Date (~2%)
  – e.g., ZIP (~2%)

MANCHESTER 1824

The University of Manchester

Integration

- Predictor integration

- Dictionary, rules and ML
- Mention level union
- **Improvements** (F_1):
  - – e.g., ML + Rules
    - Date (~3%), Patient (~3%)

Priority sorting

- Priority sorting: disambiguate overlapping predictors
- Priority sorting
  - – Specific entity pairs, e.g.,
    - AGE over DATE;
    - DOCTOR over PATIENT;
    - ZIP over IDNUM
- **Improvement:** 1% (micro F_1)

MANCHESTER 1824

The University of Manchester

Results

Category	Entity type	Frequency	F-measure %
AGE	*Age*	764	94.47
DATE	*Date*	4980	95.55
CONTACT	*Email*	1	100.00
	Fax	2	40.00
	Phone	215	94.03
LOCATION	*City*	260	81.11
	Country	117	78.73
	Hospital	875	79.08
	Organization	82	27.42
	State	190	88.22
	Street	136	94.74
	Zip	140	97.06
ID	*Idnum*	195	84.07
	Medical record	422	93.82
NAME	*Doctor*	1912	89.72
	Patient	879	86.30
	Username	92	97.78
PROFESSION	*Profession*	179	57.47

Discussion and error analysis

- Well defined categories (F_1:88-98%):
  - *Age*, *Date*, *Email*, *Idnum*, *Medicalrecord*, *Phone*, *Street* and *Zip*
- Ambiguous and contextually dependent entities (F_1:78-86%):
  - *City*, *Country*, *Hospital and Patient*
- Lexically variable and infrequent entities
  - 57% (F_1) *Profession*
  - 27% (F_1) *Organization*

MANCHESTER 1824

The University of Manchester

Discussion and error analysis

- *Organization was* a relatively infrequent (124 mentions in the 790 gold standard narratives) and broadly defined type:
  - companies ("*IBM*", "*General Dynamics*");
  - universities ("*Vassar", "Yale*");
  - government organizations ("*army*", "*marines*"),
  - industry sectors ("*publishing*", "*catering business*")
  - general organization types ("*factory*", "*library*")
  - different communities ("*quilting group*", "*the Masons*", '*methodist church*'), specific places ("*weight room*").

The University of Manchester

MANCHESTER 1824

Summary

- De-identification of longitudinal clinical narratives
  - Hybrid approach (mainly knowledge-based)
  - Cheap in terms of labour
  - <u>Two-pass recognition</u> for longitudinal data
  - <u>Priority sorting</u> for overlapping predictors

Thanks!

MANCHESTER 1824

The University of Manchester

Task II

- Creation of vocabularies:
  - Acronyms, abbreviations, and variations
  - E.g., hyperlipidemia (hld, hyperlipidemia, dyslipidemia)
- Generic lexical expressions on text suggesting risk factors:
  - “He **underwent** CABG”.
- Expressions converted through MT markup language into rules.
- Rule combination with vocabularies to identify risk factors:
  - *“He was diagnosed with @hypertension”.*
  - @hypertension contains all synonyms, acronyms, abbreviations recognised in the training/development set and any complementary nouns from ICD-9.
- Different risk factor indicators, different sets of rules:
  - **Diabetes:** hemoglobin levels, glucose levels, diabetes mentions.
  - Same rules with different dictionaries used for all disease mentions.
- Time attribute:
  - Default rules due to the longitudinal nature of the records.
  - **Medication/disease mentions:** three default values (before, during, after DCT).
  - **Other indicators:** different defaults e.g., high blood pressure: one default value (before DCT).