NLM: Machine Learning Methods for Detecting Risk Factors for Heart Disease in EHRs

Kirk Roberts, Sonya Shooshan,
Laritza Rodriguez, Swapna Abhyankar,
Halil Kilicoglu, Dina Demner-Fushman

U.S. National Library of Medicine

Approach

Document Classification

Hypertension

Approach

Document Classification vs. Information Extraction

Hypertension

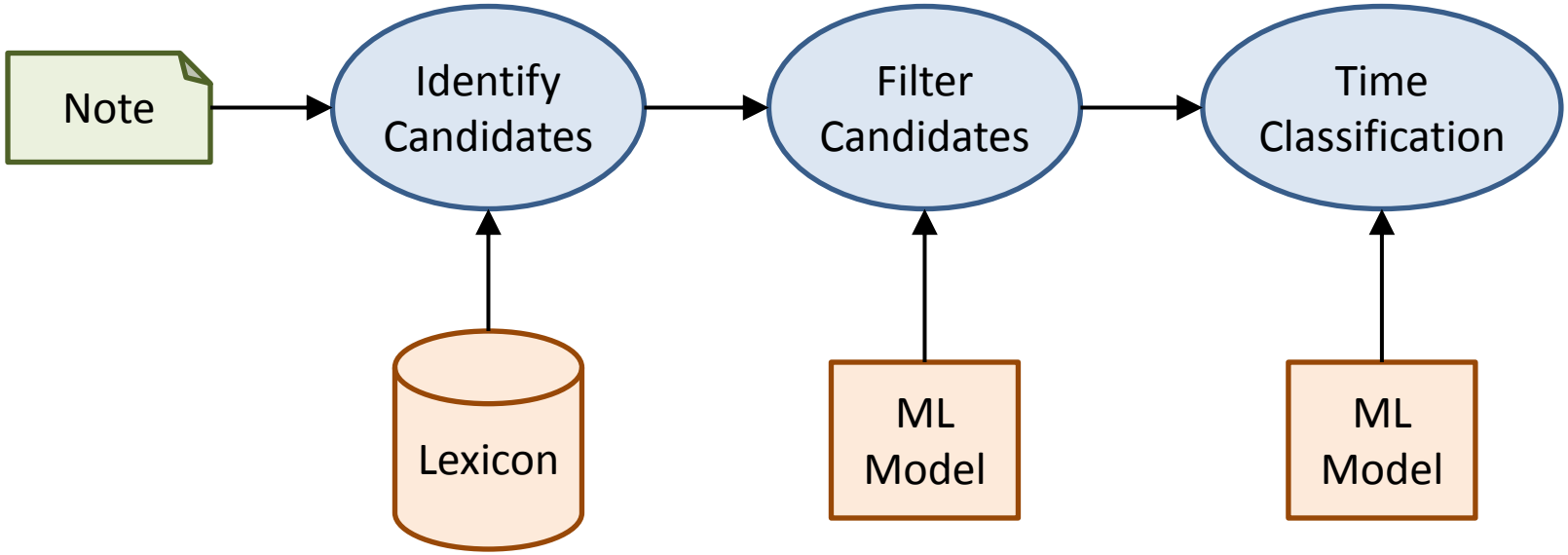
"…patient has severe hypertension…"

Approach

- In general, a single reference is enough for a document-level classification
- Relatively little lexical variety in concept references
  - hypertension, hypertensive, HTN, high blood pressure
  - A1C, HgA1C, glycohemoglobin

Approach

Machine Learning / Information Extraction



Problem

Provided annotations do not work with this method:

1. No consistent span boundaries, bad for context classification
   - Expand lexicon to account for variations
   - Assume any partial match with a lexicon item is sufficient
2. No negative annotations, lexicon matches that do not correspond to a manual annotation are either negative or unmarked positive
   - Assume all lexicon matches in a positive document are positive and likewise all matches in a negative document are negative
3. Unannotated instances are wasted potential training data
   - Above assumption could work as well

Solution

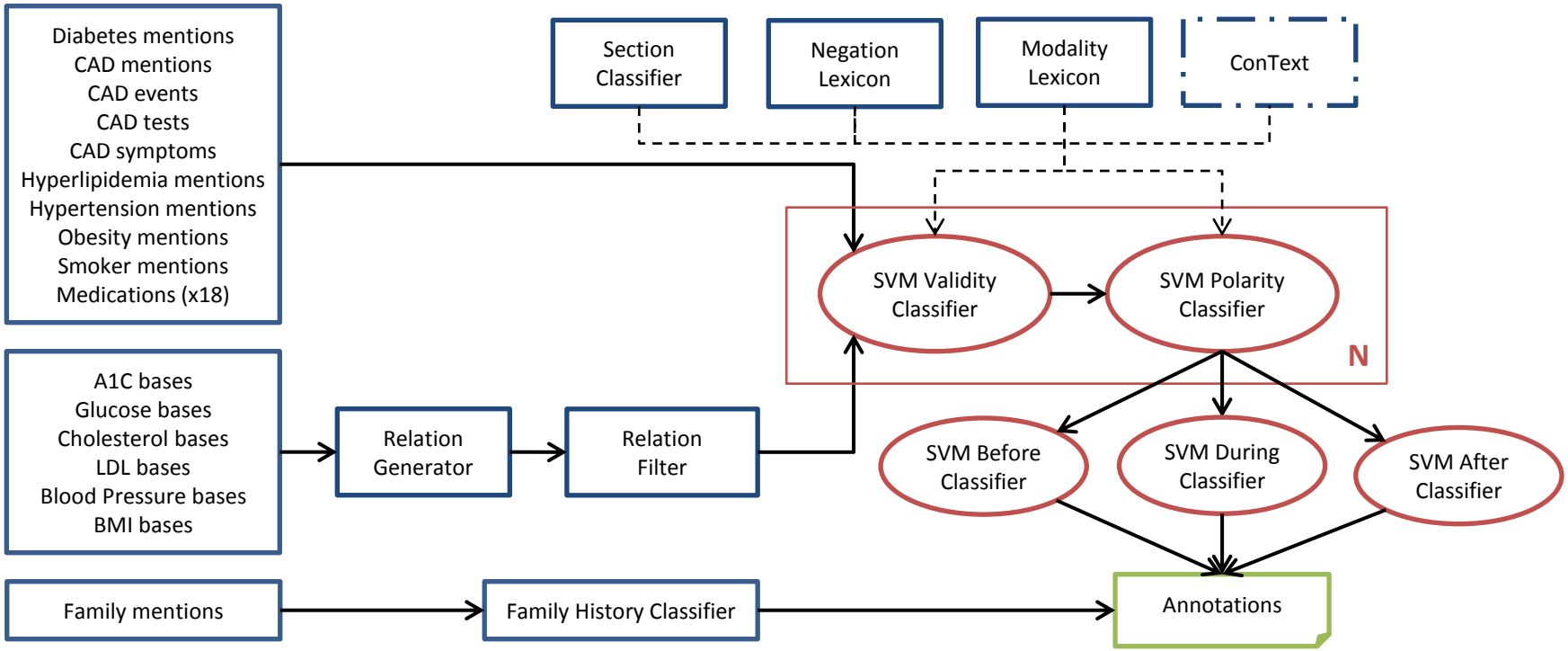
Manually annotate the data to fit our method

1. Decide on consistent annotation span boundaries for existing annotations
2. Use those annotations to build lexicon
3. Use lexicon to identify unannotated terms
4. Manually annotate these terms
   - Differentiate between *invalid* and *negative* terms

Annotation Process

- Annotators
  - Sonya E. Shooshan, MLS
  - Laritza Rodriguez, MD, PhD
  - Swapna Abhyankar, MD
  - Dina Demner-Fushman, MD, PhD
- Annotated first 2/3 of data
- Double-annotated & Resolved

Method


Diabetes mentions
CAD mentions
CAD events
CAD tests
CAD symptoms
Hyperlipidemia mentions
Hypertension mentions
Obesity mentions
Smoker mentions
Medications (x18)
A1C bases
Glucose bases
Cholesterol bases
LDL bases
Blood Pressure bases
BMI bases
Family mentions
Section Classifier
Negation Lexicon
Modality Lexicon
ConText
Relation Generator
Relation Filter
SVM Validity Classifier
SVM Polarity Classifier
N
SVM Before Classifier
SVM During Classifier
SVM After Classifier
Family History Classifier
Annotations

Candidate Identification

- Mentions: Simple lexicon look-up

> Patient does have a history of **coronary artery disease**.

- Measurements: Lexicon look-up for base term (e.g., "A1C") combined with rules for finding which numeric term the base corresponds with

> His hemoglobin **A1c** was **7.4** % a month ago.

Candidate Filtering

2 Filters:

- Binary Validity Classifier (SVM)
  - trained on {invalid} vs. {negated, positive} candidates
  - E.g., "**Ht**: 64 inches" → invalid
- Binary Polarity Classifier (SVM)
  - trained on {negated} vs. {positive} candidates
  - E.g., "father- **HTN**, 78 now" → negated

Candidate Filtering

Base Features

(1) Indexed Uncased Prev Words
(2) Indexed Uncased Next Words
(3) Generic Words within 5 Tokens
(4) Has Family Term within 5 Tokens
(5) Negation Word in Prev 10 Tokens
(6) Modality Word in Prev 10 Tokens
(7) ConText Negation Value
(8) ConText History Value
(9) ConText Hypothetical Value
(10) ConText Experiencer Value
(11) Section Name

Measurement Features

(All Base Features)
(12) Words between Base and Value
(13) Word Shapes between Base and Value
(14) Value Shape
(15) Base and Value on Same Line?
(16) # of Tokens between Base and Value
(17) Target Word in Prev 5 Tokens

Time Classification

3 Binary SVM Classifiers (before, during, after) with the following restrictions:

- Diabetes, CAD, hyperlipidemia, hypertension, and obesity mentions → [before, during, after]
- A1C, glucose, CAD event/test result/symptom, cholesterol, LDL, blood pressure, BMI → highest confidence of 3 classifiers
- Smoker → separate 5-way SVM classifier
- Medication → no restrictions

Time Features

(All Base Features)
(18) Annotation Type
(19) Medication Type

Results

Run 1: Basic System

Run 2: Basic System w/o Glucose, w/ Lexicon Pruning

Run #1	Micro	Macro
P	**0.8702**	**0.8694**
R	**0.9694**	**0.9682**
F_1	**0.9171**	**0.9162**

Run #2	Micro	Macro
P	**0.8951**	**0.8965**
R	**0.9625**	**0.9611**
F_1	**0.9276**	**0.9277**

Conclusion

- Relied on a rather simple ML architecture with straightforward features
- Manually annotated data to suit this architecture

Thank You

NLM: Machine Learning Methods for Detecting Risk Factors for Heart Disease in EHRs

Kirk Roberts, Sonya Shooshan,
Laritza Rodriguez, Swapna Abhyankar,
Halil Kilicoglu, Dina Demner-Fushman

U.S. National Library of Medicine