Biomedical Information Extraction from Semi-structured Data

Scott Russell Halgrim
Group Health Research Institute
halgrim.s@ghc.org

Fei Xia
Department of Linguistics
University of Washington
fxia@uw.edu

Imre Solti, Eithon Cadag
Biomedical & Health Informatics
University of Washington
{solti,ecadag}@uw.edu

Özlem Uzuner
Information Studies
College of Computing and Information
University of Albany, SUNY
ouzuner@uamail.albany.edu

Funded by grants 1K99LM010227-0110, 5 U54 LM008748, and U54LM008748 from the National Library of Medicine

Introduction

- Extracting information from narrative clinical records enables many applications.
- The 2009 i2b2 software development challenge was to extract medication information from discharge summaries.

From hospital discharge summaries…

Record #111999
TREATMENT:
After observing high blood sugar , patient was given 150 cc insulin once a day for one week.
DISCHARGE MEDICATIONS:
Tylenol 2 tabs q.d. p.o. headache

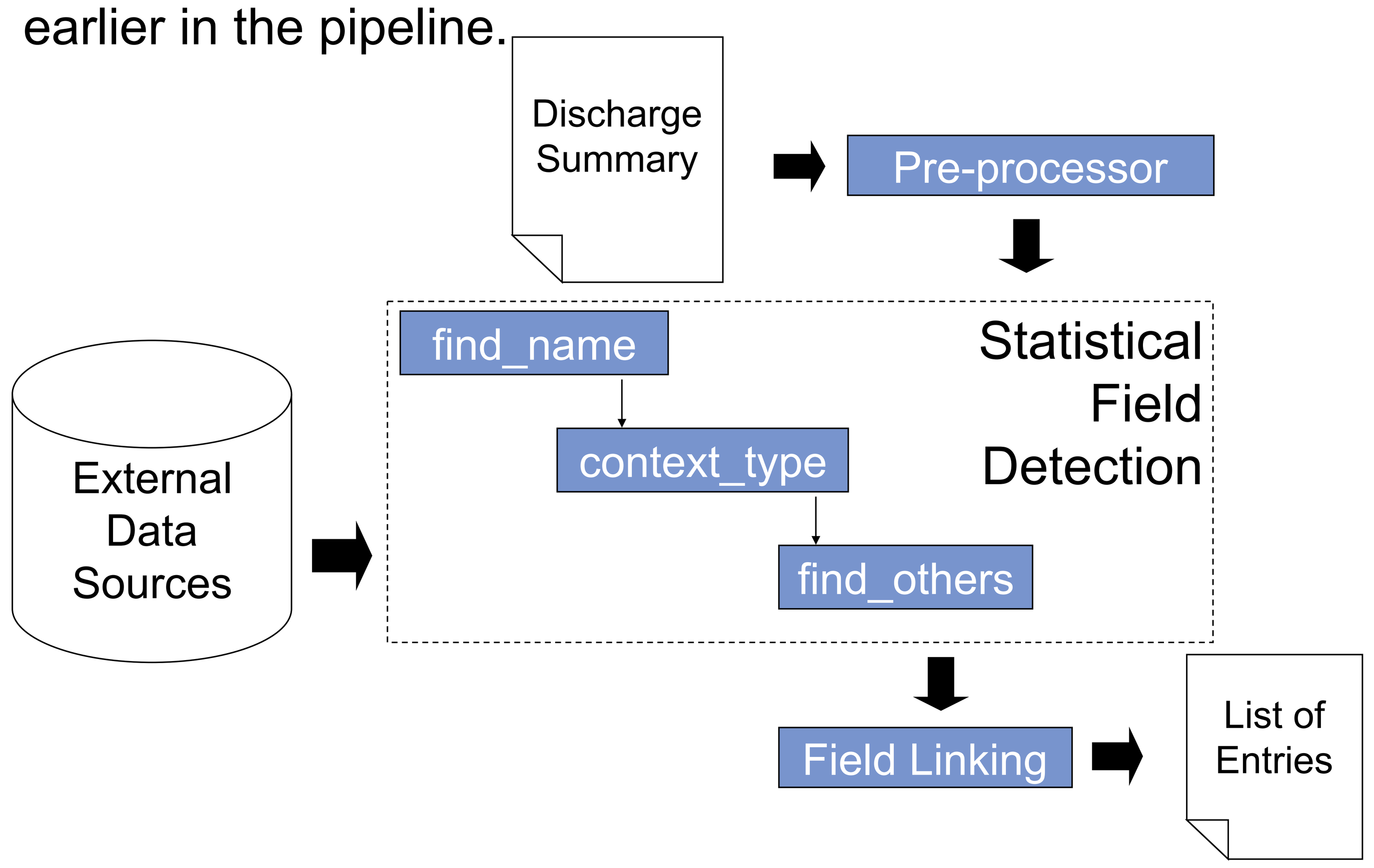
…extract six named entities and link into entries

m="insulin" || d="150 cc" || mo="nm" || f="once a day" || du="for one week" || r="high blood sugar" || ln="narrative"

m="tylenol" || d="two tabs" || mo="p.o." || f="q.d." || du="nm" || r="headache" ||ln="list"

System

- The core of our system is a pipeline of statistical classifiers.
- Modules have access to information produced by modules earlier in the pipeline.



Features

Group	Feature Types
F1	Normalized n-grams
F2	Affixes, token length, shape, and other compositional features of current and nearby tokens
F3	Class labels of previous tokens
F4	N-grams in external medications list

Link Fields

- Simple heuristics are applied to link named entities into entries.
- Non-name fields are paired with the preceding name field unless the following name field is much closer.
- The pairs are then converted to entries.

LASIX (FUROSEMIDE) 60 MG qam; 40 MG qpm PO
…
Patient was brought to the ICU on 3 mcg of epinephrine

n="LASIX (FUROSEMIDE)"||do="60 MG"||f="qam"||mo="PO"||du="nm"||r="nm"||ln="list"
n="LASIX (FUROSEMIDE)"||do="40 MG"||f="qpm"||mo="PO"||du="nm"||r="nm"||ln="list"
n="epinephrine"||do="3 mcg"||f="nm"||"mo="nm"||du="nm"||r="nm"||ln="narrative"

Data Sets

Data Set	Number of Files	Source
Training	110	University of Sydney
Development	35	University of Sydney
Test	251	i2b2 Community

- Results for our system alone are from the development set.
- Results comparing our system to others are from the test set.

Metrics

- Horizontal match:
  - First system entries are linked to gold entries.
  - Then count the fields (exact) and tokens within fields (inexact) that match across matched entries.

Gold entries:

1: n="LASIX (FUROSEMIDE)"||do="nm"||f="qam"||mo="PO"||du="for 3 days"||r="nm"
2: n="epinephrine"||do="3 mcg"||f="nm"||"mo="nm"||du="nm"||r="nm"

System entries:

1: n="LASIX"||do="nm"||f="nm"||mo="PO"||du="3 days"||r="nm"
2: n="epinephrine"||do="3 mcg"||f="nm"||"mo="nm"||du="nm"||r="nm"

- Six fields and 12 tokens in the gold standard.
- Five fields and seven tokens in the system output.
- Three fields match exactly -> Exact P/R/F = .600/.500/.545
- Seven tokens match -> Inexact P/R/F = 1.000/.583/.737

Results – Development Set

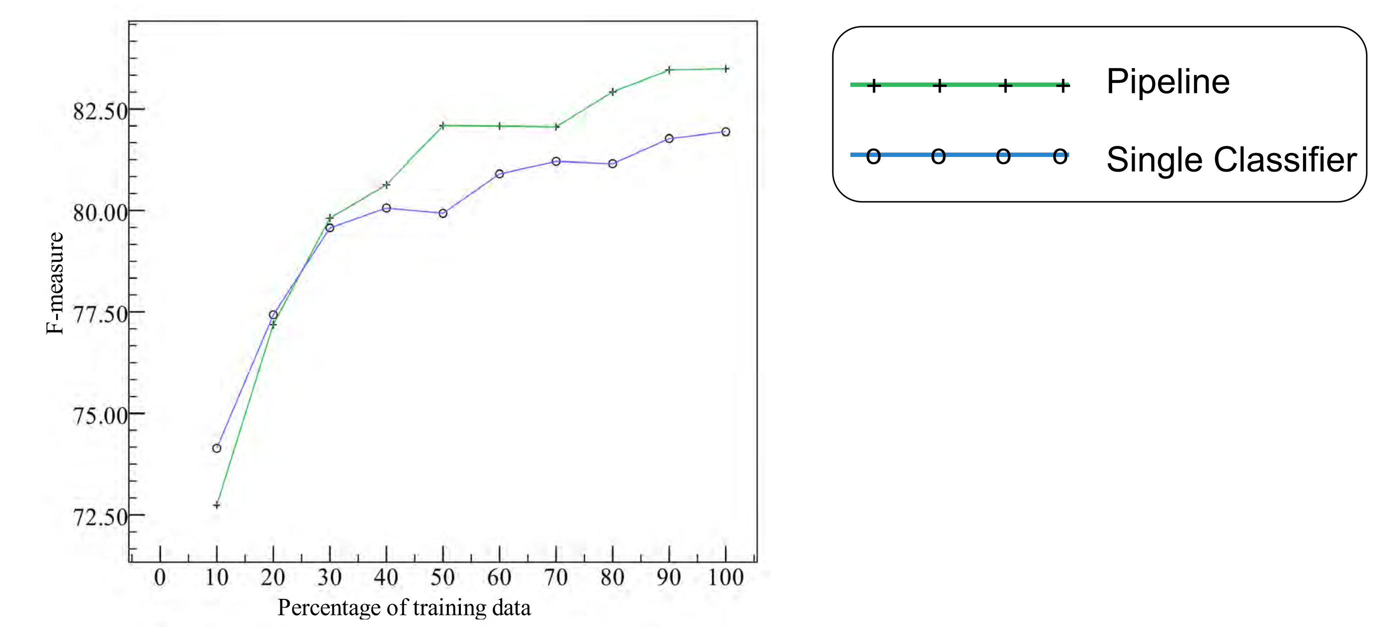
- Horizontal Exact results by feature set.

Features	Precision	Recall	F-score
F1	72.5	60.3	65.8
F1-F2	82.5	78.2	80.3
F1-F3	88.4	77.9	82.8
F1-F4	88.1	79.4	83.5

- The difference between each row is statistically significant at p<=0.01.
- The final row shows external resources help.

Pipeline v. Single Classifier

- With enough training data, the pipeline approach outperforms the single classifier.
- At 50% of training data and above, the differences are significant at p <= 0.05.



System Comparison

Our system compares favorably to those with many rules.

Exact Horizontal

Team	Prec.	Recall	F-score
Sydney	.896	.820	.857
Our system	**.886**	**.801**	**.841**
Vanderbilt	.840	.803	.821
Manchester	.864	.766	.812
NLM	.784	.823	.803
BME-Humboldt	.841	.758	.797

Inexact Horizontal

Team	Prec.	Recall	F-score
Sydney	.903	.801	.840
Our system	**.897**	**.788**	**.839**
Vanderbilt	.868	.783	.823
NLM	.898	.740	.812
OpenU	.858	.762	.807
BME-Humboldt	.850	.756	.800

Conclusion

- A machine learning approach compares favorably with rule-based approaches.
- External resources can be used to improve performance.
- A pipeline of classifiers outperforms a single classifier.