Teaching Computer to Read Medical Records

Eric Chang*, Yan Xu*+

Kai Hong#, Jianqiang Dong#, Zhaoquan Gu#

*Microsoft Research Asia

+Beihang University

#Tsinghua University

Transforming Data to Action

Data	Knowledge	Insight	Action

X-Ray
Blood pressure
Heart rate
Glucose level

Chronic high
blood pressure

High risk for
cardiovascular
disease

Cut salt intake
Lose weight

Motivation

- Translating narrative text to structural text in medical domain (concepts extraction, assertion classification, and relation identification)

- Good test case of adopting NLP to a specific domain

Challenges

- Traditional natural language processing (NLP) tools
  - Not designed for fragmented free text found in narrative clinical records
  - Does not perform well on this type of EMR data
  - Unique medical description of sentences and vocabularies

- Limited access to clinical records
  - Barrier to widespread development of medical language processing (MLP) technologies

- i2b2 research project
  - Provides de-identified medical records from 4 hospitals
  - Community work on different tasks annually

Objective

For **HTN** patient was started on **Norvasc** 10 mg daily on 1/18 .

Medical Problems, Treatments and Tests

- **Medical problems**
  - *He had an angiogram which shows **a severe stenosis** in the right distal area with a thrombosing right limb of **an aorto femoral graft***
- **Treatments**
  - *He was given **Flagyl** and had already apparently been on Levofloxacin outside the hospital since his recent discharge for questionable **pneumonia**.*
- **Tests**
  - ***an echocardiogram** revealed a pericardial effusion and possible tamponade clinically*

Assertions Made on Medical Problems

- **Present**

  the tumor was growing

- **Absent**

  patient denies pain

- **Possible**

  Patient may have had a heart attack

- **Conditional**

  Penicillin causes a rash

- **Hypothetical**

  Ativan 0.25 to 0.5 mg IV q 4 to 6 hours prn anxiety

- **Not associated with the patient**

  Brother had asthma

Medical Problems, Treatments and Tests

- Medical Problems & Treatments (TrP)
- Medical Problems & Tests (TeP)
- Medical Problems & Medical Problems (PP)

Medical Problems & Treatments

- **Treatment improves medical problem (TrIP)**

  *at that time with **anasarca** and congestive heart failure , responsive to **diuretics** and ACE inhibitors .*

- **Treatment worsens medical problem (TrWP)**

  the tumor** was growing despite the available **chemotherapeutic regimen

- **Treatment causes medical problem (TrCP)**

  *Also the risk of **ischemia** or infarct from **the internal carotid artery coil** could lead to thromboembolism*

- **Treatment is administered for medical problem (TrAP)**

  *He was given **Flagyl** and had already apparently been on Levofloxacin outside the hospital since his recent discharge for questionable **pneumonia** .*

- **Treatment is not administered because of medical problem (TrNAP)**

  *The patient 's **antibiotics** were discontinued with the thought that prolonged antibiotics only put her at more risk for **infection** .*

- **None of the above defined treatment-problem relationships (NoneTrP)**

  *In terms of his **liver abnormalities**, it was felt that viral hepatitis was in the differential as well as several opportunistic infections of the liver but also was felt that **Bactrim** could be a cause of these abnormalities .*

Medical Problems & Tests

- **Test reveals medical problem (TeRP)**

  an echocardiogram revealed a pericardial effusion and possible tamponade clinically

- **Test conducted to investigate medical problem (TeCP)**

  an VQ scan was performed to investigate pulmonary embolus

- **None of the above defined test-problem relationship (NoneTeP)**

  The patient has history of asthma and a new diagnosis of heart failure diagnosed by echocardiogram.

Medical Problems & Medical Problems

- **Medical problem indicates medical problem (PIP)**

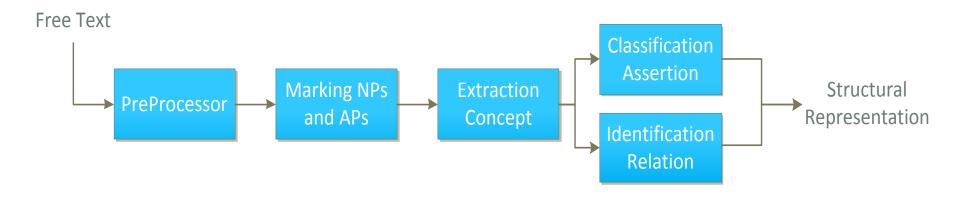 Azotemia presumed secondary to sepsis

- **None of the above defined problem-problem relationship (NonePP)**

  Significant for hypertension, hyperlipidemia

eHuatuo

Medical Record **Problem List**

Concept
- [ ] Problem
- [ ] Treatment
- [ ] Test

Assertion
- [ ] Present
- [ ] Absent
- [ ] Possible
- [ ] Conditional
- [ ] Hypothetical
- [ ] Not associated

Relation
- [ ] P indicate P
- [ ] Tr improve P
- [ ] Tr worsen P
- [ ] Tr cause P
- [ ] Tr administer P
- [ ] Tr not administer P
- [ ] Test reveal P
- [ ] Test conduct P

Clear

Start

Our Method

Free Text

PreProcessor → Marking NPs and APs → Extraction Concept → Classification Assertion / Identification Relation → Structural Representation

- Pre-processing sentences
- Marking noun phrases (NPs) and adjective phrases (APs)
- Extracting concepts
- Classifying assertions
- Identifying relations

Structural representation

Microsoft Research
微软亚洲研究院

Assertion

Feature
- Lexical
- Syntactic
- Rule-based

voting

Classification Assertion

Classifier
- SVM
- Boosting
- Logistic Regression
- Multi-LR

Rule-based

Relation

voting

Identification Relation

Classifier
- SVM
- Boosting
- Logistic Regression
- Multi-LR

Normalization

Feature
- Lexical
- Syntactic

Concept

Extraction Concept

LCS boundary

CRF

Feature
- Lexical
- Orthographic
- Syntactic
- Ontological

Medication Model   Other Model

Feature
- Lexical
- Orthographic
- Syntactic
- Ontological

Syntax analysis

Marking Noun Phrases and Adjective Phrases

POS → Chunk → Parser

Relation Parsing

LinkGrammar

Preprocessing (Remove Signs)

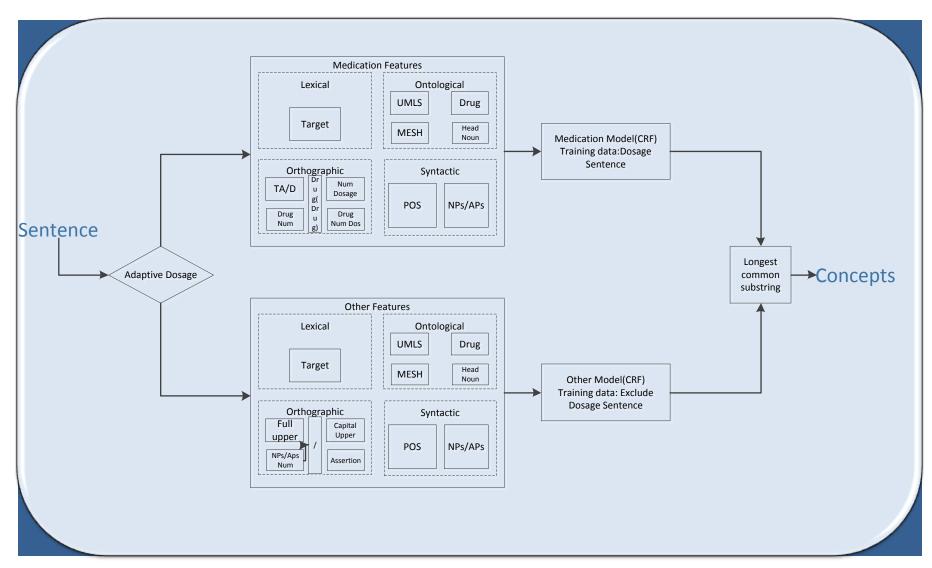| + | # | ( ) | : | \ | - | _ | = | ABC | Abc | & | * |

Narrative discharge summary

Extracting Concepts

- Determining types and models
  - Adaptive dosage dictionary

- Generating corresponding features
  - Building dictionaries including UMLS, MESH, Drug-name, Head noun
  - Features for "medication" Model
  - Features for "other" Model

- Extracting concepts and matching types
  - CRF++

- Obtaining correct boundaries for the concepts
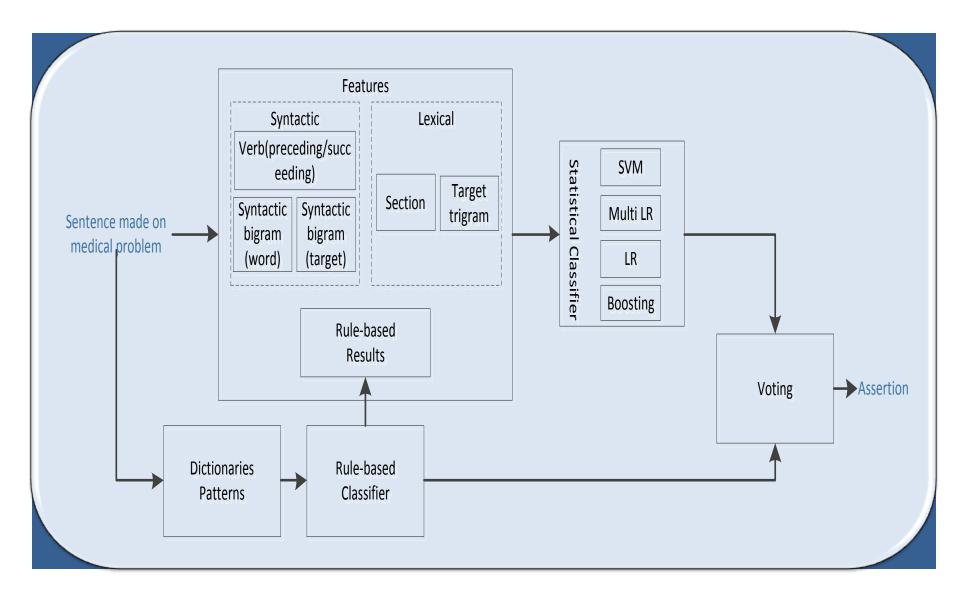 - Longest common substring algorithm (LCS)

Extracting Concepts

Classifying Assertions

- Generating dictionaries and rule-based patterns manually
- Classifying assertions by rule-based classifier
- Extracting features
    - Lexical context features
    - Syntactic context features
    - Results from rule-based classifier
- Classifying assertions by statistical classifiers
- Voting
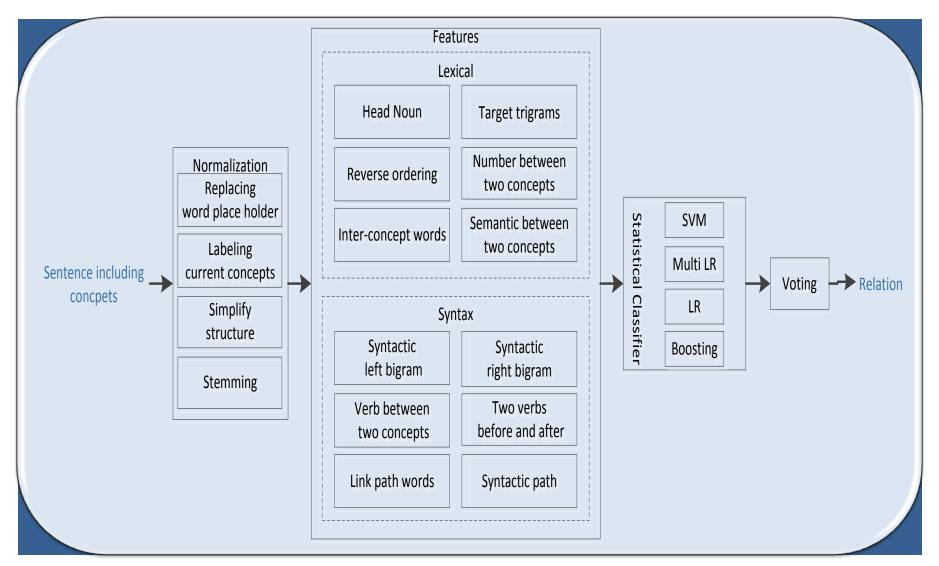
Classifying Assertions

Identifying Relations

- Normalizing sentences
  - Replacing each concept with one word place holder
  - Labeling current two concepts to be identified
  - Simplify the sentence structures
  - Stemming

- Extracting features
  - Lexical context features
  - Syntactic context features

- Identifying relations by statistical classifiers

- Voting

Identifying Relations

Submitted Results for Three Tasks

MSRA's submitted results for 2010 i2b2/VA NLP Workshop

Micro	precision	recall	F-measure	2010 Workshop F-measure max
Concept	0.7443	0.7905	0.7667	0.8523
Assertion	0.9210	0.9210	0.9210	0.9362
Relation	0.6198	0.6517	0.6354	0.7365

Latest Results for Three Tasks

Micro	precision	recall	F-measure	2010 Workshop F-measure max
Concept	0.8189	0.8589	0.8385	0.8523
Assertion	0.9403	0.9403	0.9403	0.9362
Relation	0.7227	0.7229	0.7228	0.7365

Submitted vs. Latest Results for Three Tasks

Micro	F-measure Submitted	F-measure Latest System	2010 Workshop F-measure max
Concept	0.7667	0.8385	0.8523
Assertion	0.9210	0.9403	0.9362
Relation	0.6354	0.7228	0.7365

Summary

- Carried out all 3 tasks of i2b2 2010 challenge
  - Invited to present based on good results
  - New work since July shows our system is among the top systems
- For assertion, rule based system worked well, with machine learning based system adding incrementally to final result
- Time consuming to craft rules and hard to scale
- Future work: explore active learning to help machine learning system take advantage of more data

Acknowledgements

- 2010 I2b2/VA challenge organizers
- Microsoft Research Asia
  - Weizhu Chen, Sigma Machine Learning Toolkit
  - Changning (Tom) Huang
  - Chin-Yew Lin
- Online resources

Thanks!

eric.chang@microsoft.com