Improving Transportability of NLP Solutions using CLAMP

Hua Xu, PhD

School of Biomedical Informatics,
University of Texas Health Science Center at Houston

The Transportability Problem of NLP

- Many NLP systems have been developed and shown good performance in different tasks
- However, performance drop when transporting these NLP tools
  - From one type of clinical notes to another
  - From one institute to another
  - From one application to another

An example of smoking status detection

- Mayo Clinical NLP System for smoking detection
  - I2b2 dataset, F-measure 85.5% (*Savova et al. JAMIA 2008*)
  - Vanderbilt dataset, F-measure 75% (*Liu et al. AMIA 2012*)
- Customize it to achieve an F-measure of 89%
  - Sentence boundaries
  - New lexicons (e.g., “Tob neg”)
  - Re-training of SVM classifiers by annotating local data
  - Add new rules

What is CLAMP - Clinical Language Annotation, Modeling, and Processing?

- A general purpose clinical NLP system – "**CLAMP CMD**"
  - Built on proven methods
  - Good performance, high speed
- An IDE (integrated development environment) for building customized clinical NLP pipelines via GUIs – "**CLAMP GUI**"
  - Annotating/analyzing clinical text
  - Training of ML-based modules
  - Specifying rules
- An enterprise solution for NLP needs in healthcare organizations – "**CLAMP Enterprise**"
  - Task management
  - Visual analytics

CLAMP CMD – built on proven methods

NLP Tasks		Ranking
Named entity recognition	2009 i2b2, medication	#2
	2010 i2b2 problem, treatment, test	#2
	2013 SHARe/CLEF abbreviation	#1
UMLS encoding	2014 SemEval, disorder	#1
Relation extraction	2012 i2b2 Temporal	#1
	2015 SemEval Disease-modifier	#1
	2015 BioCREATIVE Chemical-induced disease	#1

CLAMP CMD – performance

- Extract problems, treatments, and tests

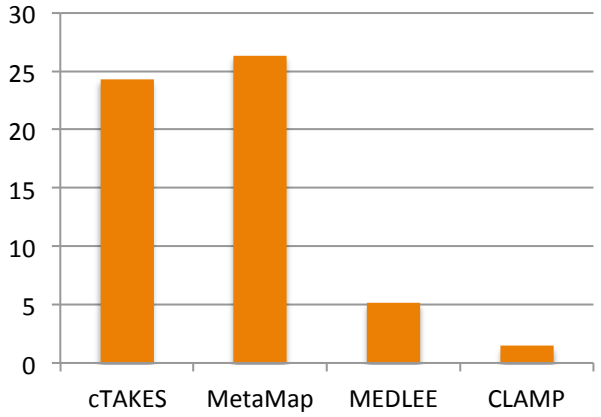
			Exact match			Relaxed match		
Corpus	Entity types	# entity	P	R	F1	P	R	F1
MTsamples	treatment, problem, test	25,531	0.841	0.811	0.826	0.921	0.890	**0.905**
i2b2	treatment, problem, test	72,846	0.891	0.861	0.876	0.958	0.925	**0.941**
UTNotes	treatment, problem, test	124,869	0.921	0.900	0.910	0.963	0.940	**0.951**
SemEval 2014	Disease_Disorder	10,077	0.861	0.791	0.824	0.870	0.799	**0.833**
SemEval 2015	Disease_Disorder	17,333	0.867	0.816	0.840	0.886	0.834	**0.859**

CLAMP CMD - speed

- Thread-safe

Pipeline	MAC		Linux		Windows	
	Single thread	Multi threads	Single thread	Multi threads	Single thread	Multi threads
clamp-ner	0.72	0.28	**1.25**	**0.17**	0.69	0.302
clamp-ner-attribute	0.93	0.38	1.59	0.24	0.90	0.422
disease-attribute	0.62	0.26	1.06	0.17	0.58	0.296
lab-attribute	0.62	0.26	0.99	0.16	0.56	0.286
medication-attribute	0.67	0.29	1.15	0.18	0.61	0.3

Test Data Mimic2 Data set (500 documents) **Number of multi-threads** 10

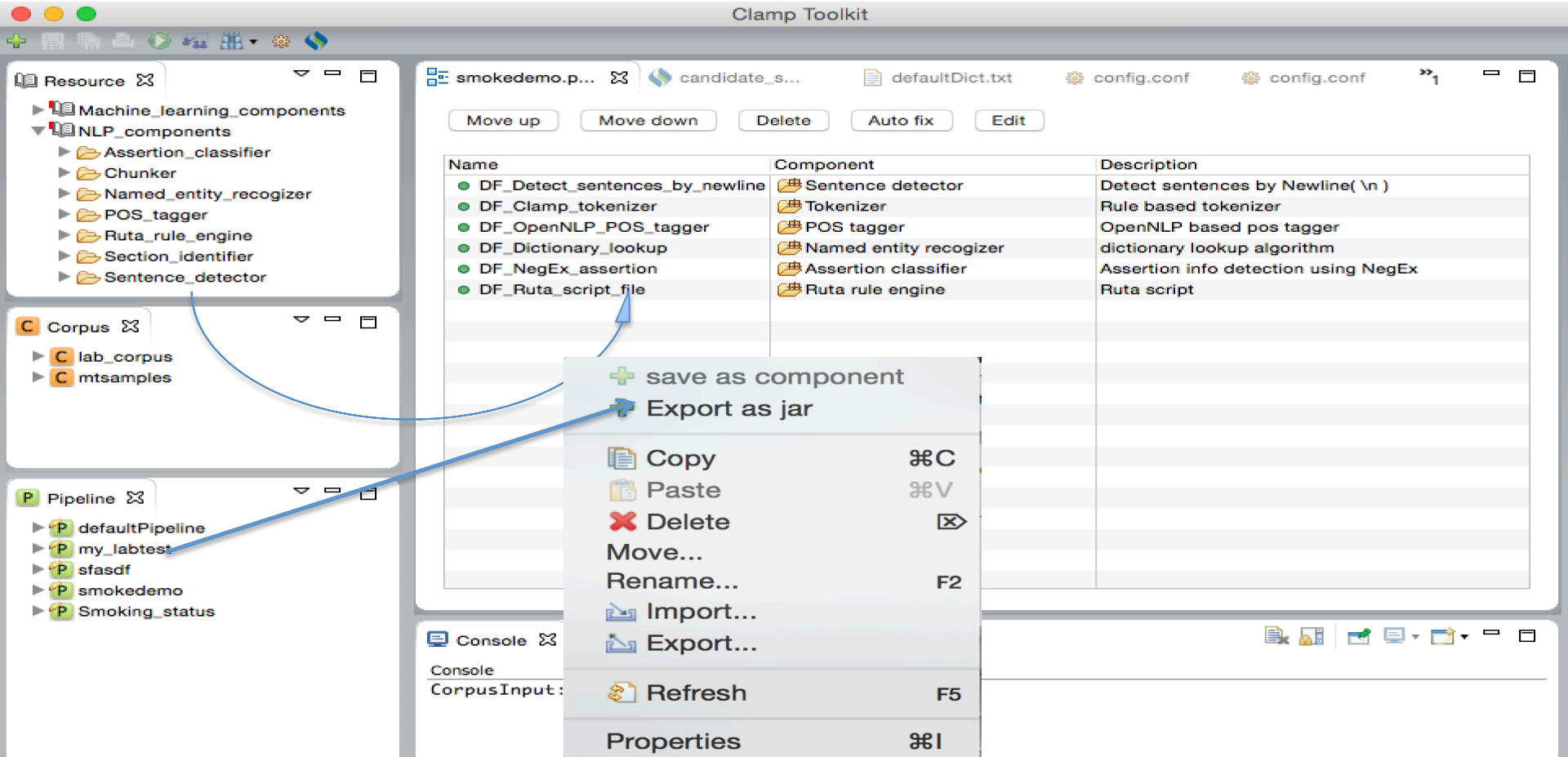


SemEval 2015 Corpus: 431 documents, avg doc size: 9.38k
SINGLE THREAD

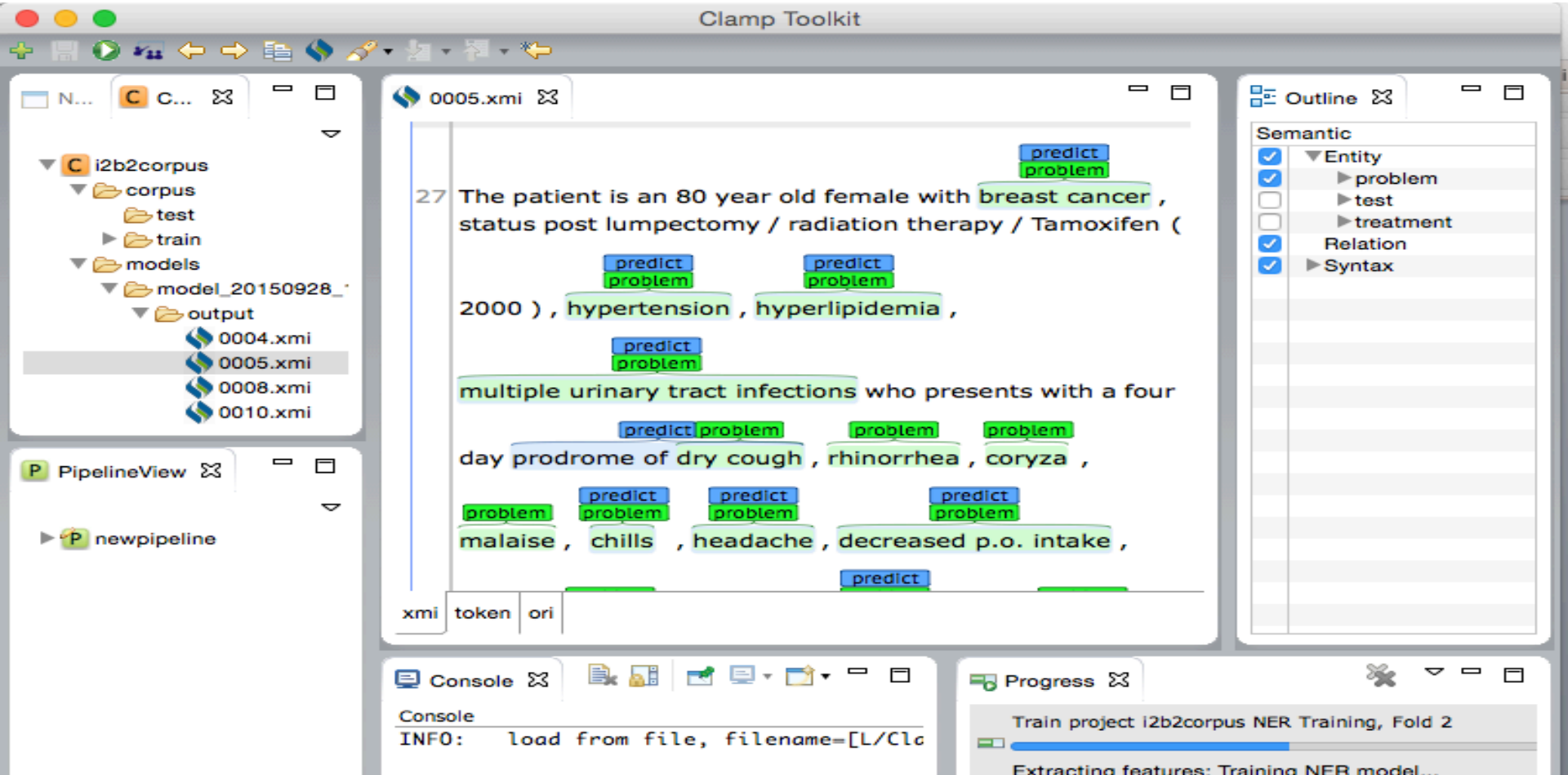
CLAMP GUI – two use cases

- Build a rule-based system to extract smoking status in clinical text and classify them into three categories: current smoker, past smoker, and non-smoker
- Build a hybrid (machine learning + rules) system for extracting lab tests and associated values from clinical text
- Videos: http://clamp.uth.edu/tutorial.php

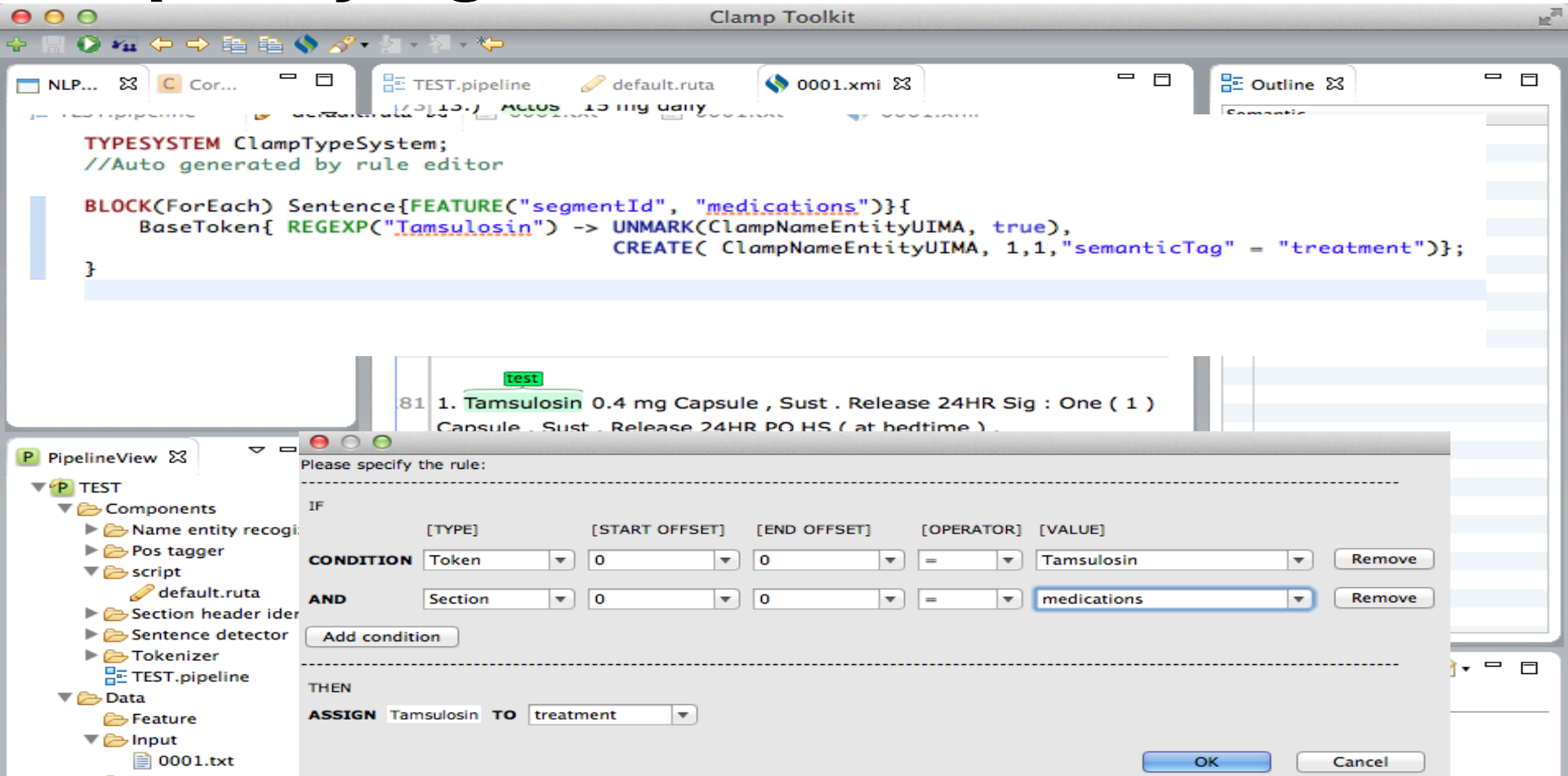
Building your own pipeline



Annotating/Re-training

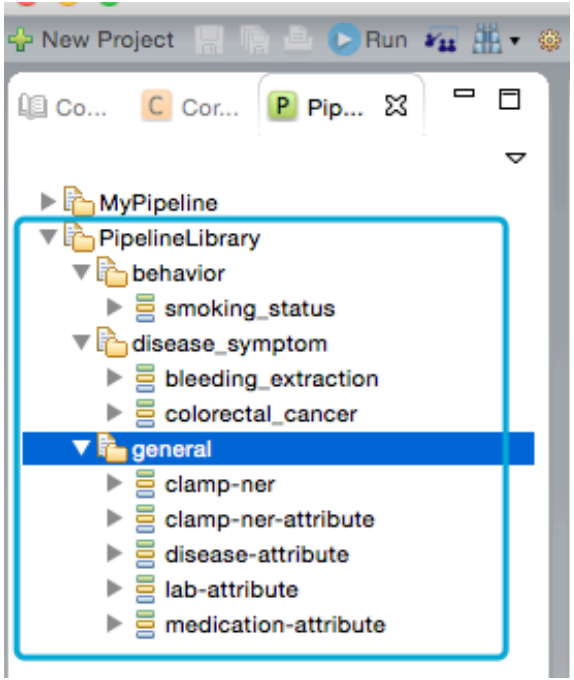


Specifying rules



A Library of NLP Pipelines

- General
  - Problem/treatment/test
  - Diseases with modifiers
  - Medications with signature
  - Lab tests and values
- Disease-specific
  - Colorectal cancer
  - Bleedings
  - …
- Behavior
  - Smoking status
  - …



Interoperate with existing systems

- Compatible with cTAKES – UIMA
- MetaMap2CLAMP – a command line program

MetaMap Machine Readable Outputs

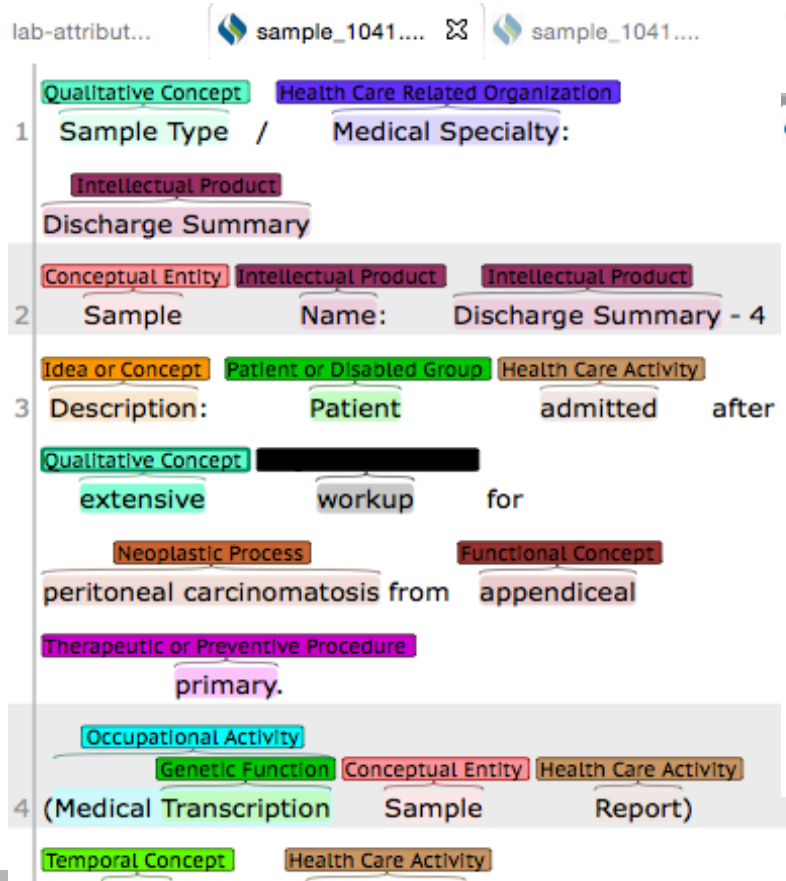
CLAMP XML inputs

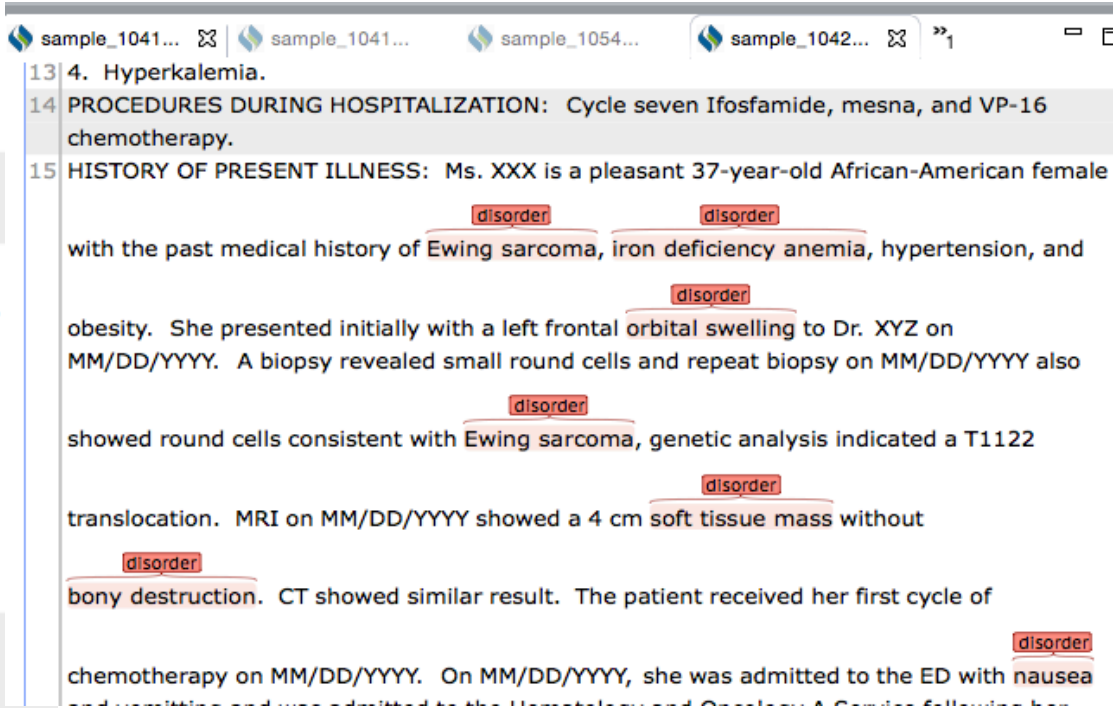
sample_1068.txt

```
args('metamap12.BINARY.Linux -L 2012 -Z 2012AA -y -qE /home/qwei1/metamap_data/
preprocessed/sample_1068.txt.pre /home/qwei1/metamap_data/output/sample_1068.txt.pre.out',
[lexicon_year-'2012',mm_data_year-'2012AA',word_sense_disambiguation-[],machine_output-
[],indicate_citation_end-[],infile-'/home/qwei1/metamap_data/preprocessed/
sample_1068.txt.pre',outfile-'/home/qwei1/metamap_data/output/sample_1068.txt.pre.out']).
aas([]).
neg_list([]).
utterance('00000000.tx.1',"Sample Type / Medical Specialty:  Discharge Summary",0/51,[]).
phrase('Sample Type / Medical Specialty',
[mod([lexmatch([sample]),inputmatch(['Sample']),tag(noun),tokens([sample])]),mod([lexmatch
([type]),inputmatch(['Type']),tag(noun),tokens([type])]),punc([inputmatch([/]),tokens([])]
),head([lexmatch([medical
specialty']),inputmatch(['Medical','Specialty']),tag(noun),tokens([medical,specialty])])],
0/31,[]).
candidates(9,0,0,9,[ev(-861,'C0037778','Specialty, Medical','Medical Specialities',
[medical,specialty],[bmod],[[[3,4],[1,2],0]],yes,no,
['AOD','CHV','CSP','MSH','MTH','NCI','SNOMEDCT'],[14/17],0),ev(-861,'C1552531','Medical
Specialty','Clinic / Center - Medical Specialty',[medical,specialty],[hcro,mnob],[[[3,4],
[1,2],0]],yes,no,['HL7V3.0','MTH'],[14/17],0),ev(-812,'C0037778','Specialty','Medical
Specialities',[specialty],[bmod],[[[4,4],[1,1],0]],yes,no,
['AOD','CHV','CSP','MSH','MTH','NCI','SNOMEDCT'],[22/9],
0),ev(-812,'C0205476','Medical','Medical',[medical],[ftcn],[[[3,3],[1,1],0]],yes,no,
['CHV','MTH','NCI','SNMI','SNOMEDCT'],[14/7],0),ev(-694,'C2347029','Sample
Type','Biospecimen Type',[sample,type],[qlco],[[[1,1],[1,1],0],[[2,2],[2,2],0]],no,no,
['MTH','NCI'],[0/11],0),ev(-645,'C0332307','Type','Type - attribute',[type],[qlco],
[[[2,2],[1,1],0]],no,no,['CHV','LNC','MTH','NCI','SNMI','SNOMEDCT'],[7/4],
0),ev(-645,'C0370003','Sample','Specimen',[sample],[sbst],[[[1,1],[1,1],0]],no,no,
['CHV','HL7V3.0','LCH','LNC','MTH','NCI','SNOMEDCT'],[0/6],
0),ev(-645,'C1547052','*Type','*Type - Kind of quantity',[type],[qnco],[[[2,2],[1,1],
0]],no,no,['HL7V2.5','MTH'],[7/4],0),ev(-645,'C2347026','Sample','Biospecimen',[sample],
```

```
urine throughout the admission.&#10;Keywords:&#10;discharge summary, iron deficiency
anemia, hypertension and obesity, iron deficiency, urine, anemia, hypertension,
chemotherapy, discharge, ewing, sarcoma,&#10;"/><tcas:DocumentAnnotation xmi:id="8"
sofa="1" begin="0" end="1956" language="x-unspecified"/><textspan:Sentence xmi:id="13"
sofa="1" begin="0" end="51" sentenceNumber="0"/><textspan:Sentence xmi:id="19" sofa="1"
begin="52" end="86" sentenceNumber="1"/><textspan:Sentence xmi:id="25" sofa="1" begin="87"
end="208" sentenceNumber="2"/><textspan:Sentence xmi:id="31" sofa="1" begin="209"
end="246" sentenceNumber="3"/><textspan:Sentence xmi:id="37" sofa="1" begin="247"
end="278" sentenceNumber="4"/><textspan:Sentence xmi:id="43" sofa="1" begin="279"
end="310" sentenceNumber="5"/><textspan:Sentence xmi:id="49" sofa="1" begin="311"
end="344" sentenceNumber="6"/><textspan:Sentence xmi:id="55" sofa="1" begin="345"
end="389" sentenceNumber="7"/><textspan:Sentence xmi:id="61" sofa="1" begin="390"
end="410" sentenceNumber="8"/><textspan:Sentence xmi:id="67" sofa="1" begin="411"
end="413" sentenceNumber="9"/><textspan:Sentence xmi:id="73" sofa="1" begin="415"
end="429" sentenceNumber="10"/><textspan:Sentence xmi:id="79" sofa="1" begin="430"
end="432" sentenceNumber="11"/><textspan:Sentence xmi:id="85" sofa="1" begin="434"
end="441" sentenceNumber="12"/><textspan:Sentence xmi:id="91" sofa="1" begin="442"
end="444" sentenceNumber="13"/><textspan:Sentence xmi:id="97" sofa="1" begin="446"
end="459" sentenceNumber="14"/><textspan:Sentence xmi:id="103" sofa="1" begin="460"
end="462" sentenceNumber="15"/><textspan:Sentence xmi:id="109" sofa="1" begin="464"
end="477" sentenceNumber="16"/><textspan:Sentence xmi:id="115" sofa="1" begin="478"
end="568" sentenceNumber="17"/><textspan:Sentence xmi:id="121" sofa="1" begin="569"
end="601" sentenceNumber="18"/><textspan:Sentence xmi:id="127" sofa="1" begin="602"
end="754" sentenceNumber="19"/><textspan:Sentence xmi:id="133" sofa="1" begin="756"
end="842" sentenceNumber="20"/><textspan:Sentence xmi:id="139" sofa="1" begin="844"
end="1016" sentenceNumber="21"/><textspan:Sentence xmi:id="145" sofa="1" begin="1018"
end="1092" sentenceNumber="22"/><textspan:Sentence xmi:id="151" sofa="1" begin="1094"
end="1119" sentenceNumber="23"/><textspan:Sentence xmi:id="157" sofa="1" begin="1121"
end="1188" sentenceNumber="24"/><textspan:Sentence xmi:id="163" sofa="1" begin="1190"
```

MetaMap concept filtering



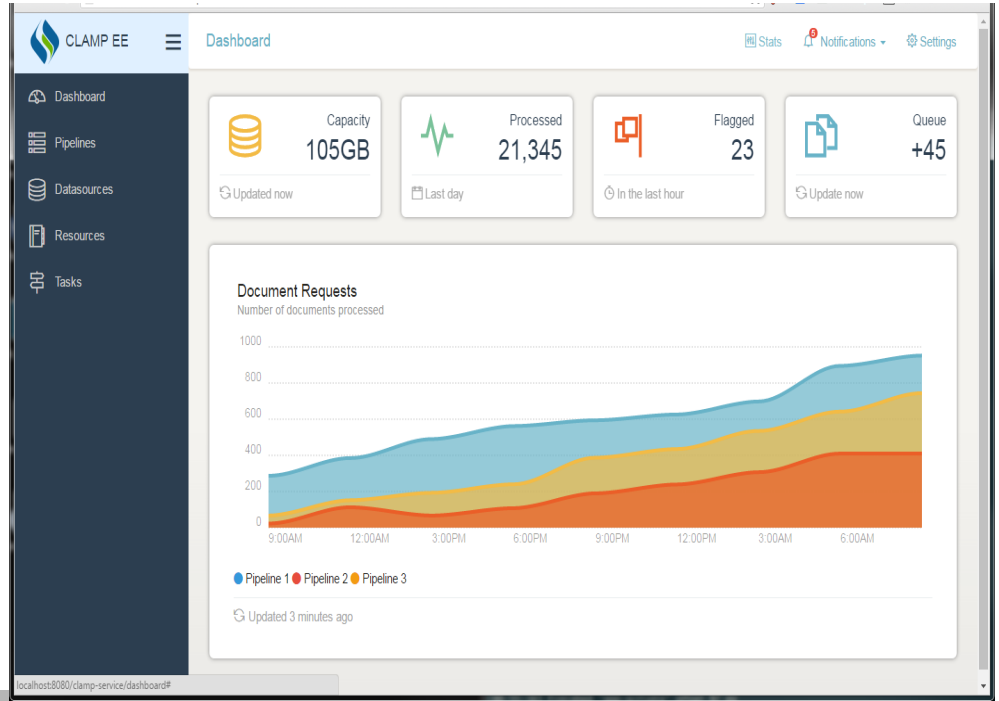


Additional Features for Developers

- CLAMP APIs for different components
  - Document object encapsulates the UIMA JCas object to overcome its complexity.
- CLAMP Extensions
  - Create custom components by implementing DocProcessor interface (UIMA compatible).

CLAMP Enterprise

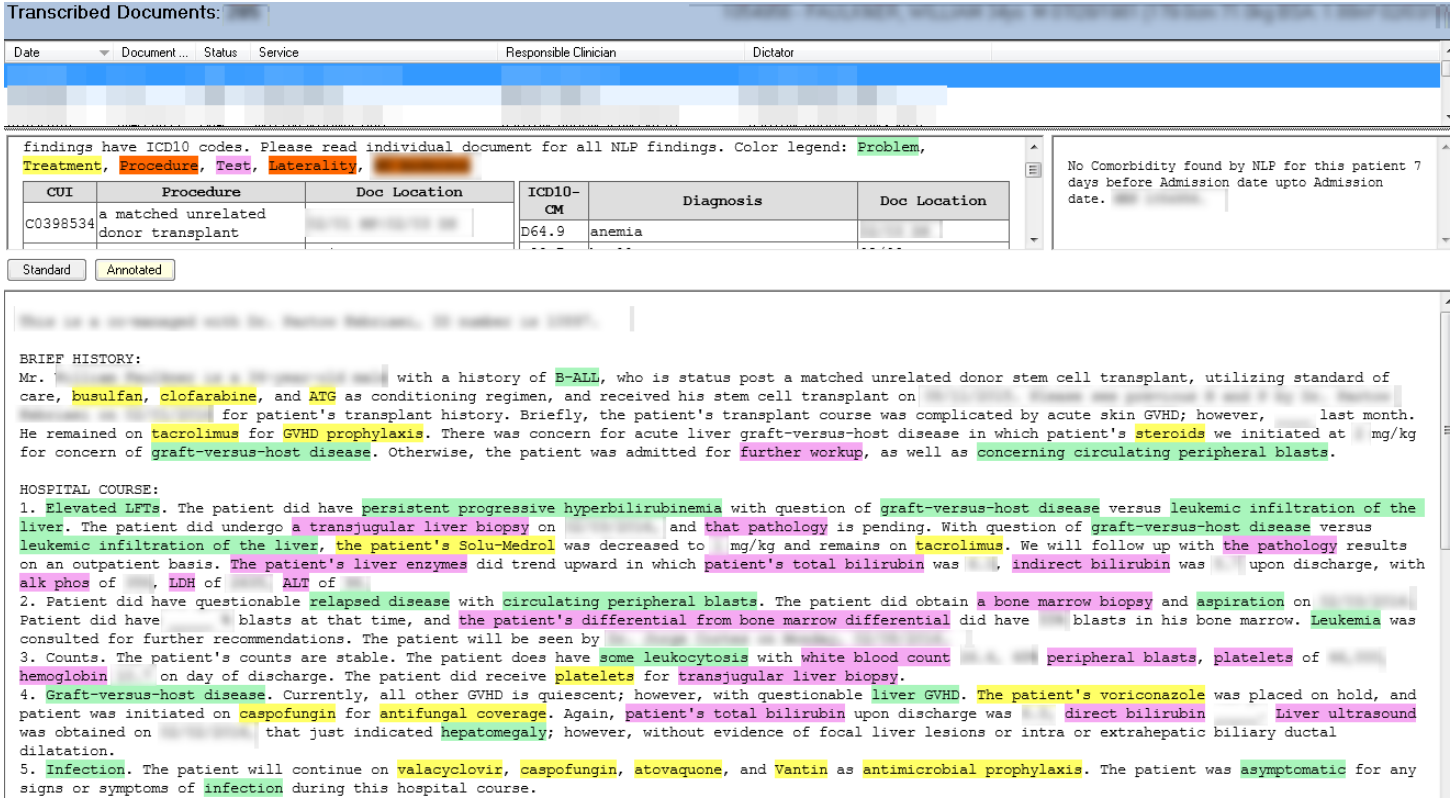
- Still under development



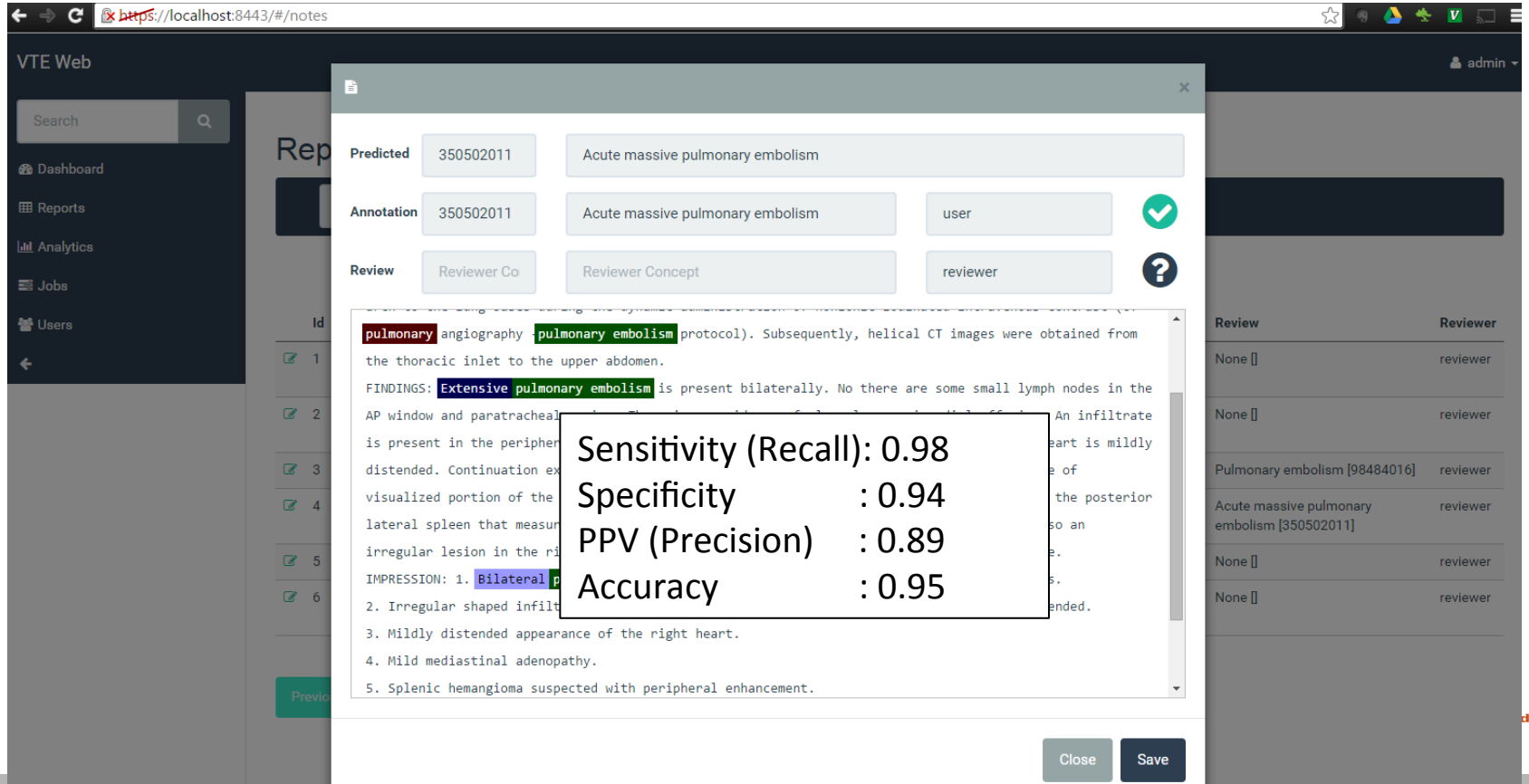
CLAMP users and use cases

- 20+ different organizations: academic + industry
- Many different use cases

Use case # 1 – computer assisted coding of ICD 10 at MD Anderson



Use case # 2 – Meaningful use quality measurement – VTE detection at Memorial Hermann Hospital



Use case #3 – support clinical research and informatics education

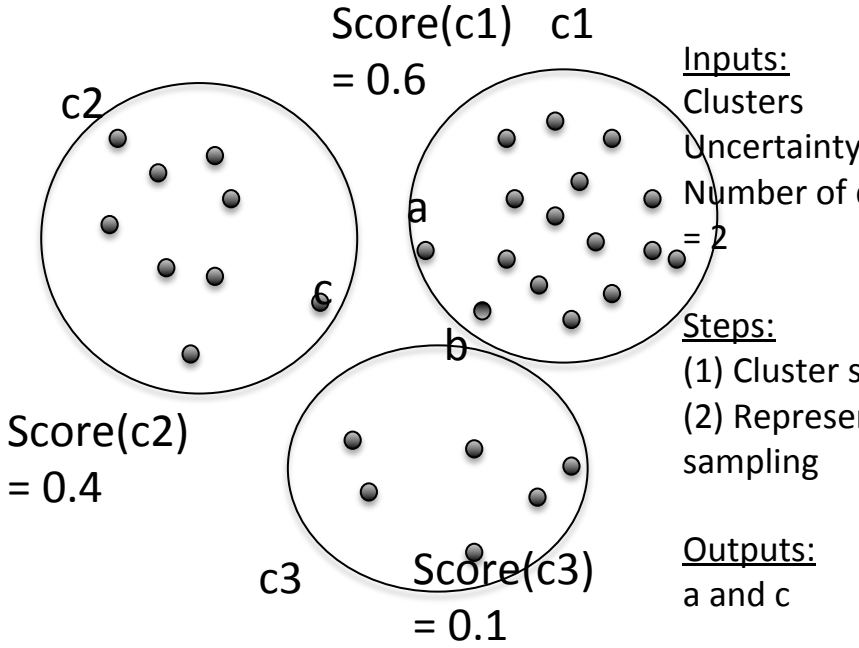
- EHR-based drug studies
  - Pharmacovigilance
  - Pharmacogenomics
  - Drug repurposing
- Domain-specific studies
  - Cancer epidemiology
  - Psychiatric studies
  - …….
- Education for biomedical NLP

More advanced algorithms to facilitate the transportability

- Active learning-based pre-annotation
- Deep learning-based NER
- Domain adaption
- Abbreviation handling

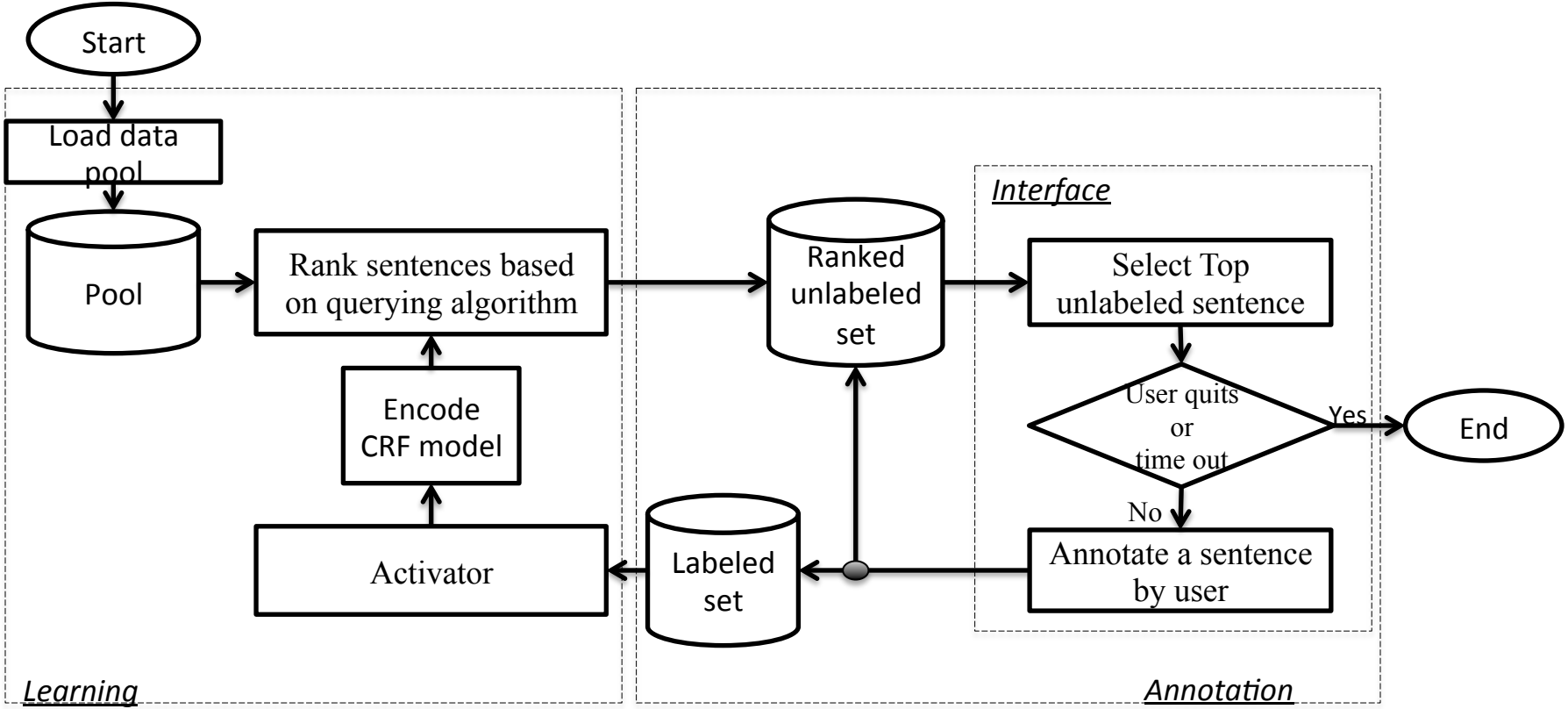
Active learning based pre-annotation

- Motivation
  - Annotation is costly
- Algorithm development
  - Clustering and uncertainty sampling engine (CAUSE)
  - Query the most uncertain and representative sentences

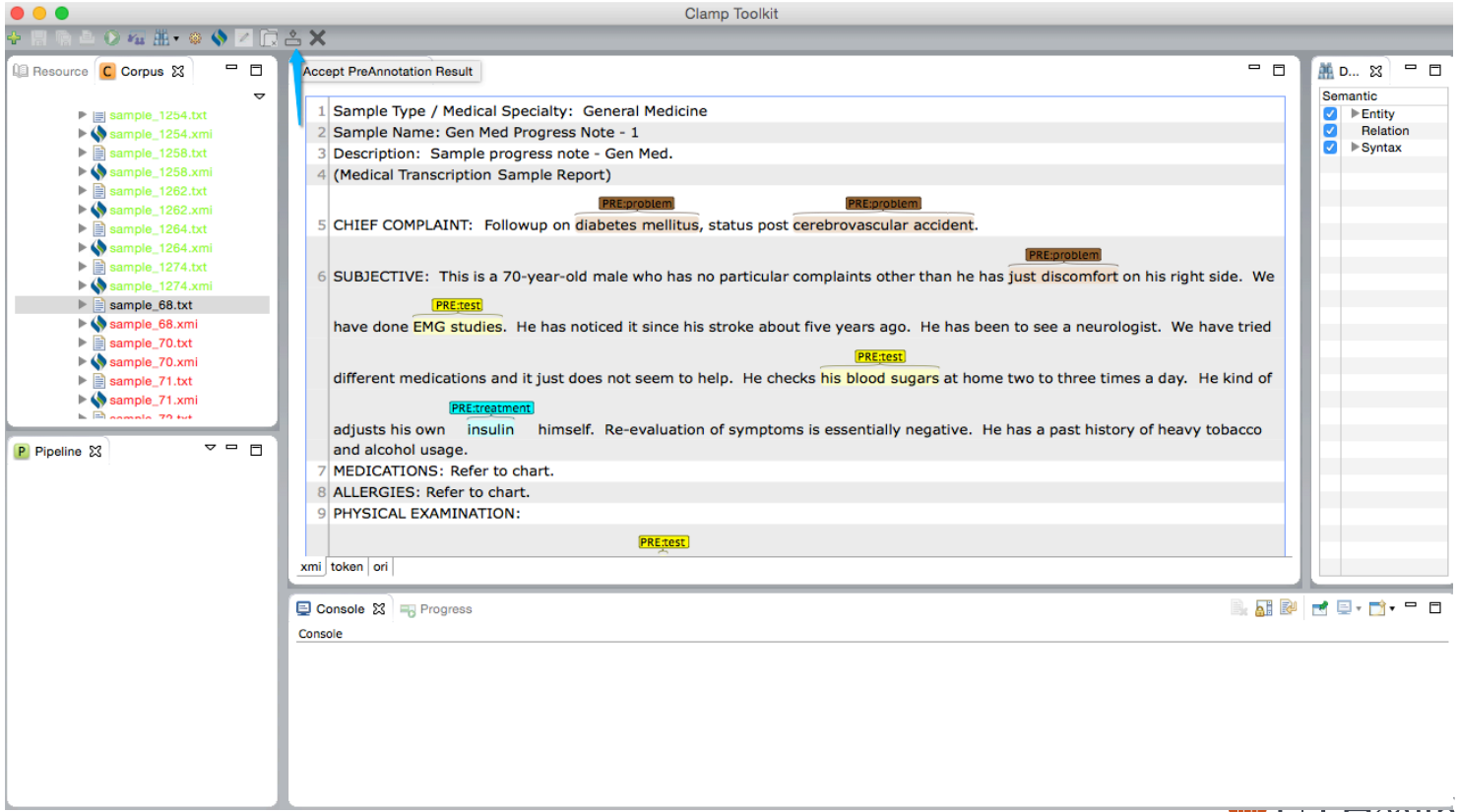


Chen Y, Dissertation, Vanderbilt, 2015.

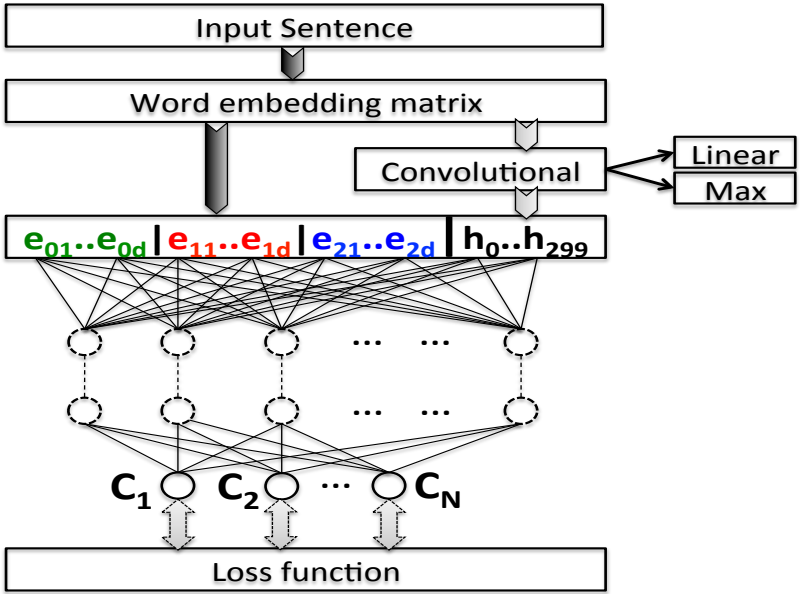
Active learning workflow



Pre-annotation interface



Deep neural networks for NER



The i2b2 corpus

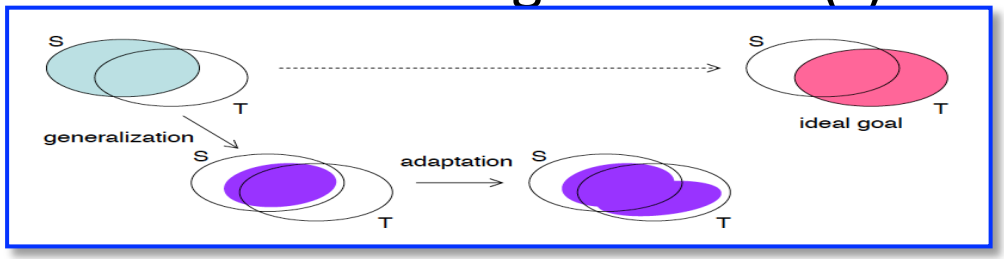
	Precision	Recall	F1-score
CRF-baseline	0.848	0.755	0.799
CRF-optimized	0.872	0.825	0.848
DNN	0.849	0.807	0.828

DNN Advantages:

- No human feature engineering
- No domain knowledge
- Simplified application system architecture
- Good baseline performance

Domain Adaptation

- **Motivation:** deploy models from fixed source domain across different target domain(s)



- Different algorithms
  - **Instance pruning:** remove most different instances from source
  - **Transfer self-training**: add most similar source instances to the target training set iteratively for retrain
  - **Feature augment**: amplifying target training features with weight-adapted source domain features.

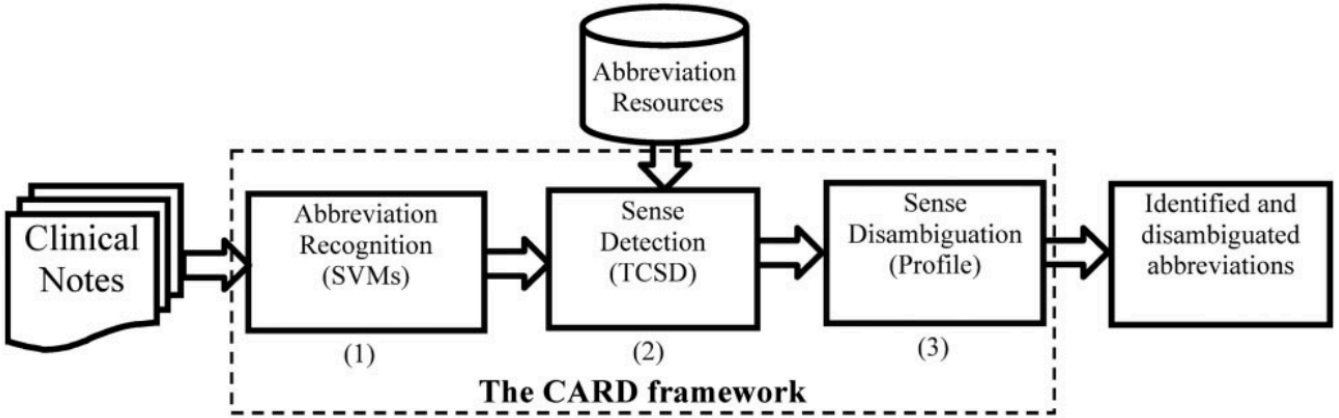
Domain Adaptation for semantic role labeling (SRL)

- significantly improved SRL performance
- reduced ~40% annotation cost
- directly using source domain dataset without domain adaptation may decrease the SRL performance
- different domain adaptation algorithms may be effective for different source domain datasets

Zhang Y et al, JAMIA 2015.

Handling abbreviations

- Motivation
  - Abbreviations are pervasive, important, ambiguous, dynamic
- CARD – Clinical Abbreviation Recognition and Disambiguation framework

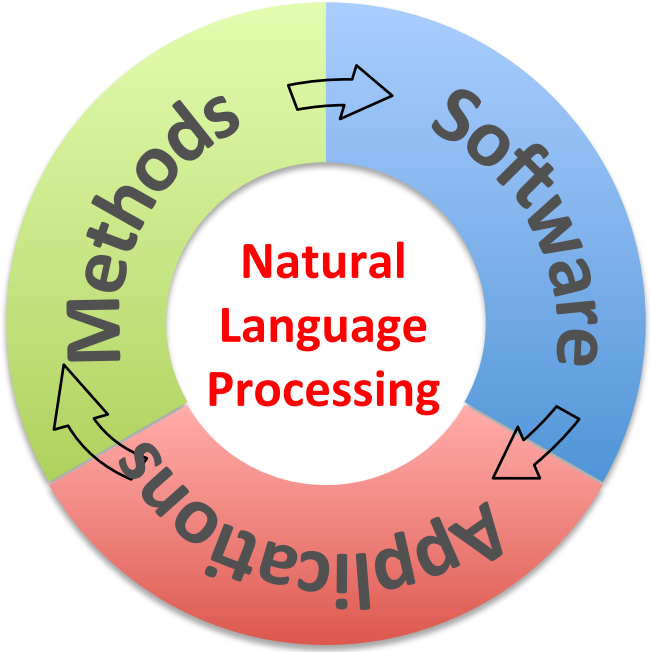


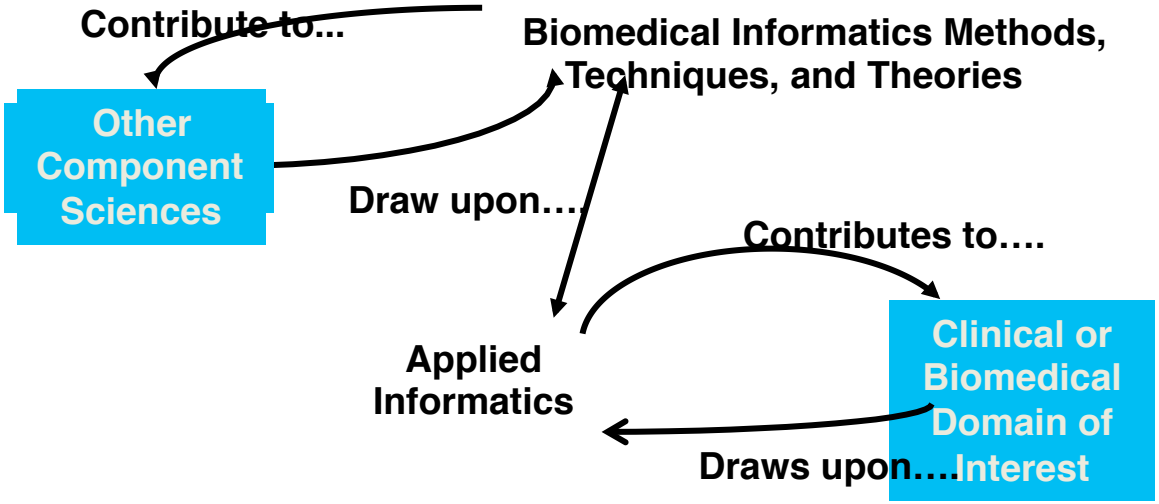
Wu et al. JAMIA 2016

Work with CLAMP

- License model
  - Research – free
    - User (without source code)
    - Contributor (with source code)
  - Operation/Commercial – paid
    - User (without source code)
    - Contributor (with source code)
- Support service
  - GitHub issue reporting
  - Supporting team: documentation, tutorial, QA…
- Available at http://clamp.uth.edu/

NLP and Biomedical Informatics





Acknowledgement

- Collaborators
  - Robert Murphy, MD
  - Josh Denny, MD
  - Trent Rosenbloom, MD
  - Randy Miller, MD
  - Hongfang Liu, PhD
  - Qiaozhu Mei, PhD
  - Serguei Pakhomov, PhD
  - Jason Hou, MD
  - Tom Lasko, MD, PhD

- Grants
  - CPRIT R1307
  - NIGMS R01 GM102282
  - NLM R01 LM010681

Team members:

- Jingqi Wang
- Min Jiang
- Ergin Soysal
- Sungrim Moon
- Jun Xu
- Yaoyun Zhang
- Anupama Gururaj
- Yonghui Wu
- Nina Slimi
- Kyle Nguyen
- Tolulola Dawodu
- Yukun Chen
- Qiang Wei
- Saied Pournejati
- Rui Li

Thank you!

Questions?

hua.xu@uth.tmc.edu