Extracting CHF information from clinical text using CLAMP

Hua Xu, PhD

pSCANNER 2016

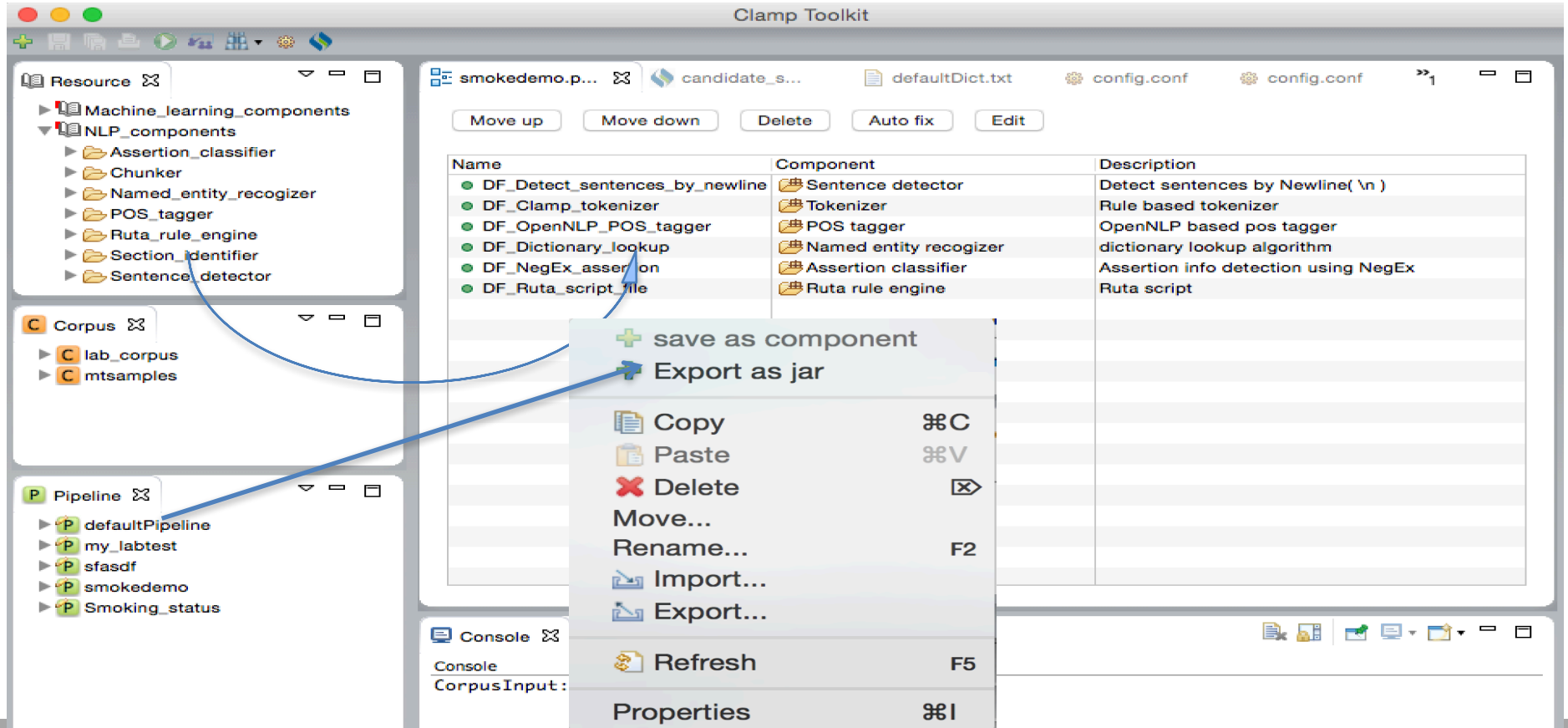
What is CLAMP - Clinical Language Annotation, Modeling, and Processing?

- 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

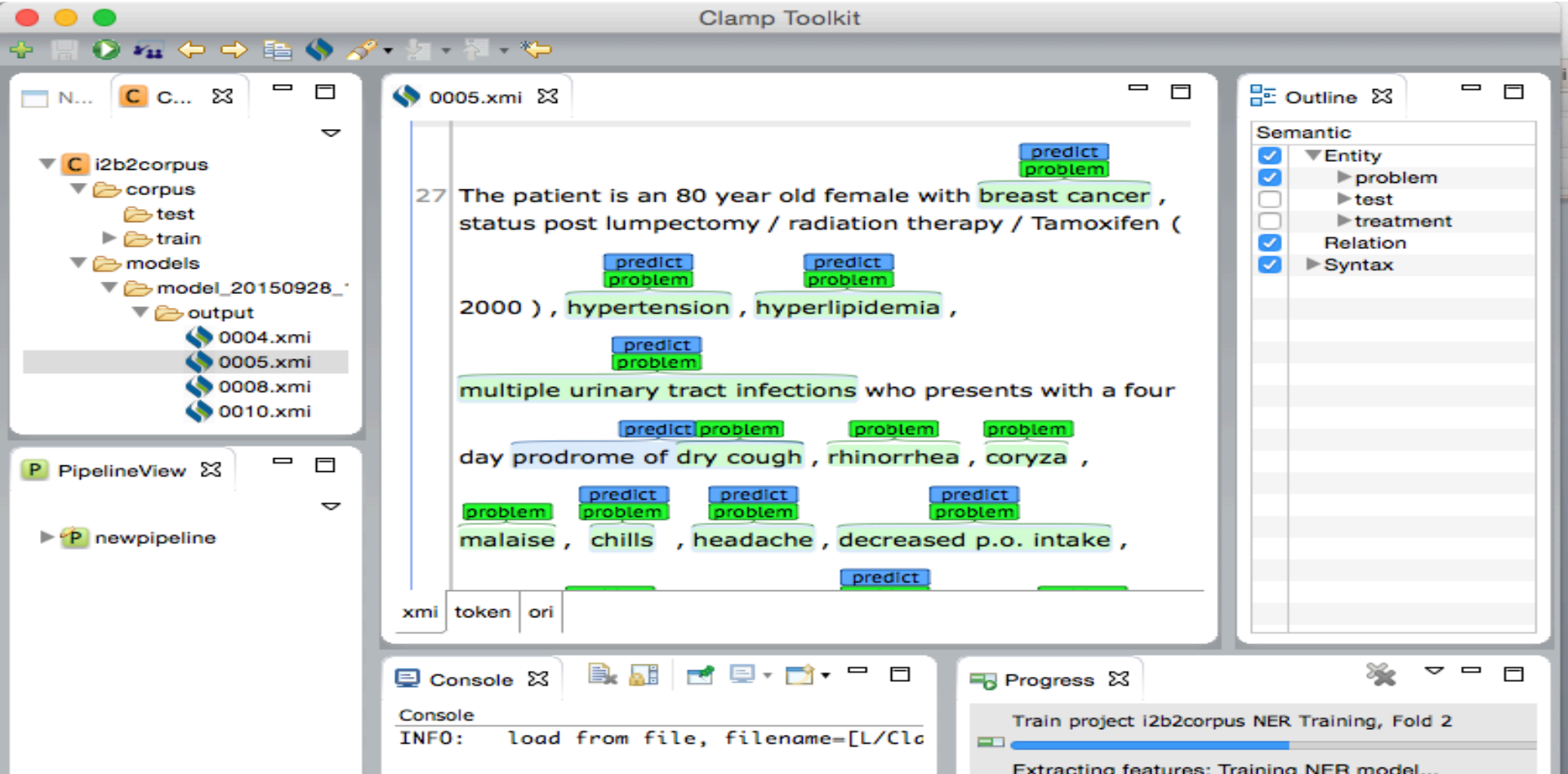
CLAMP CMD – built on proven methods

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

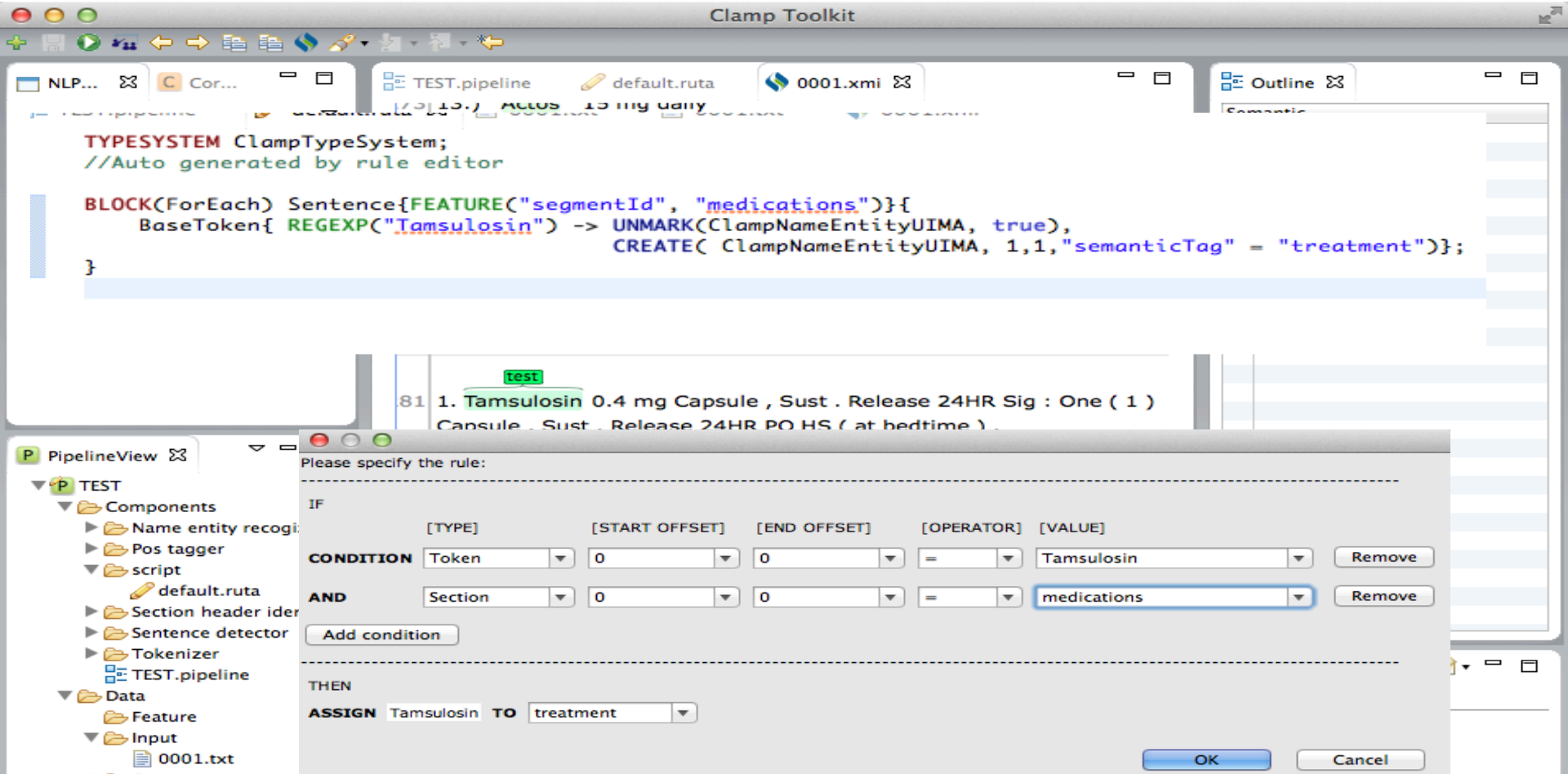
CLAMP GUI - Efficiently build customized NLP pipelines for individual applications!



Annotating/Re-training



Specifying rules



Extracting CHF information using CLAMP built-in components

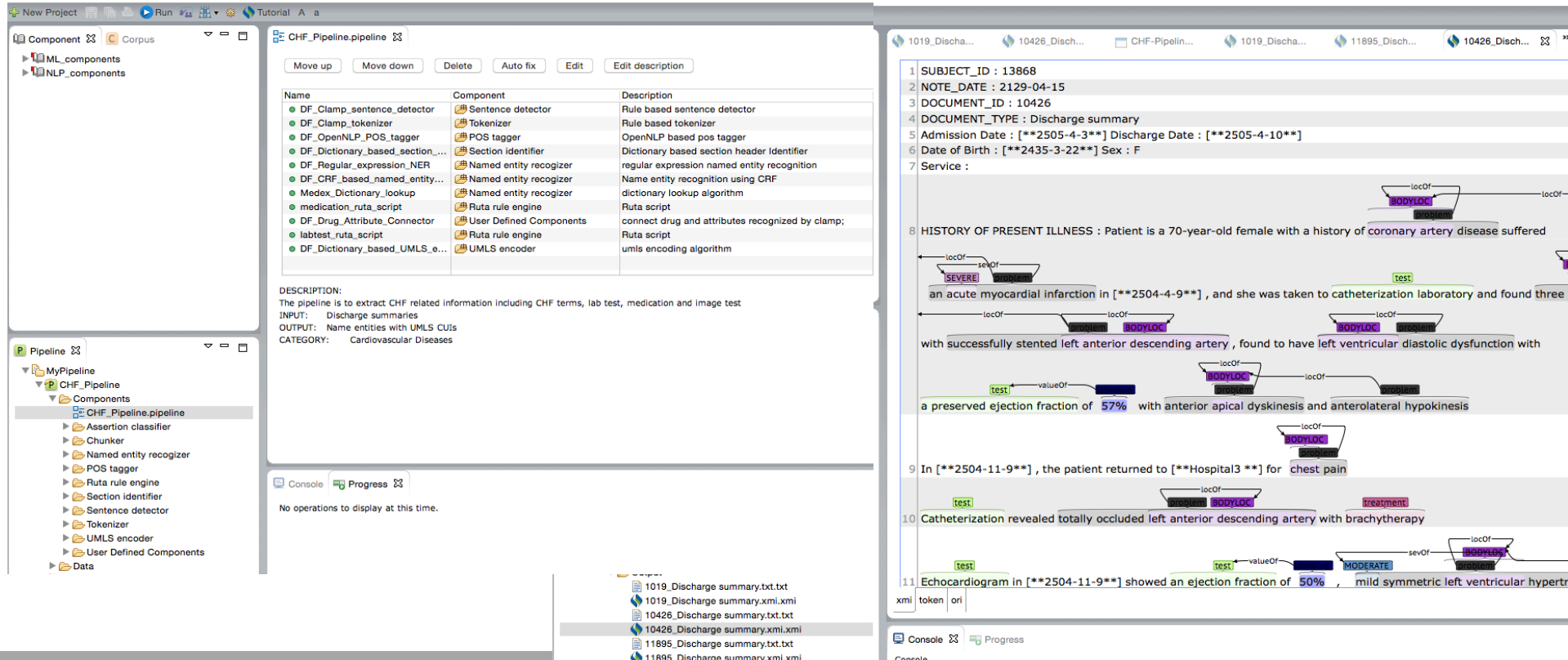
CHF information

- CHF Terms/Concepts
  - Any terms mentioned regarding to Congestive Heart Failure
  - E.g. “congestive heart failure”, “systolic heart failure” or “diastolic heart failure” etc.
- Lab tests
  - Laboratory test aiming to evaluate if patient has Congestive Heart Failure
  - E.g. “BUN”, “RBC” etc.
- Medications
  - Any medications to treat Congestive Heart Failure
  - E.g. ACE inhibitors, Angiotensin receptor blockers etc.
- Image tests
  - image tests to evaluate if patient has Congestive Heart Failure
  - E.g. “Ejection Fraction”, “chest x-ray” etc.

CLAMP built-in components

- Basic name entity recognition
  - Regular expression based NER - Numbers
  - Machine learning based NER - problems, treatments and tests
- Medication and signature identification
  - Wrap the rule-based MedEx-UIMA as a CLAMP component
- Lab test and value identification
  - Rule based module
- UMLS Encoding
  - Assign UMLS code to previously identified name entities
- **Customize it to CHF using post-processing rules**
  - Fltering by CUIs and names

CLAMP screen shots



Evaluation

- MIMIC-III corpus
  - 10,000+ patient with CHF diagnosis code
  - Randomly selected 30 discharge summaries from these patients
  - A nurse annotated CHF related information
- Six categories of name entities: CHF Term, Lab test name, Lab test name + value, Image name, Drug name, Drug name + signatures
- Matching criteria: exact vs. inexact
- Metrics: precision, recall, and F1

Preliminary Results

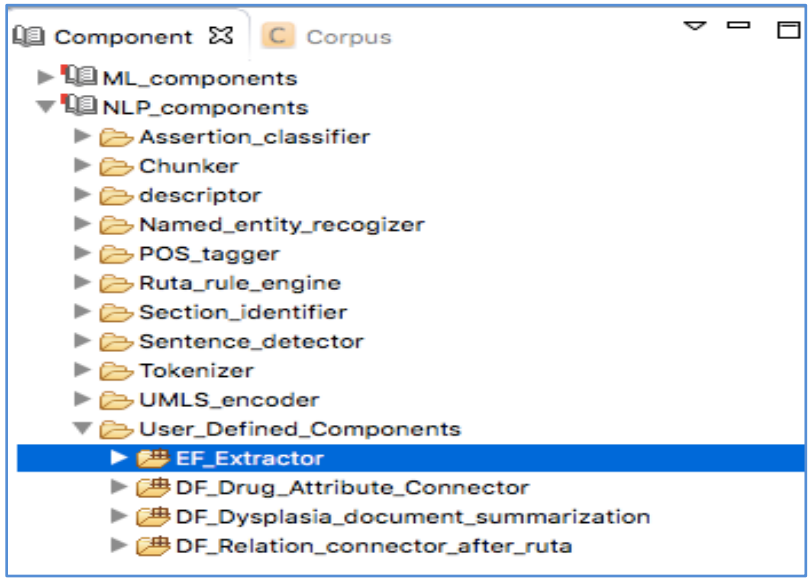
Category	# of Gold	# of Sys	Recall(%)	Precision(%)	F-1 measure(%)
CHF Term	80	79	80.0/93.7	81.0/94.9	80.5/94.3
Lab test name	318	331	89.6/95.6	86.1/91.8	87.8/93.6
Lab test name + value	318	331	84.5/92.7	81.2/89.1	82.8/90.8
Image	136	139	85.2/99.2	83.4/97.1	84.2/98.1
Drug name	388	392	83.5/96.6	82.6/95.6	83.0/96.0
Drug name + signature	637	642	81.4/95.7	80.8/95.0	81.1/95.3
All	1693	1691	84.7/94.3	84.8/94.5	84.8/94.4

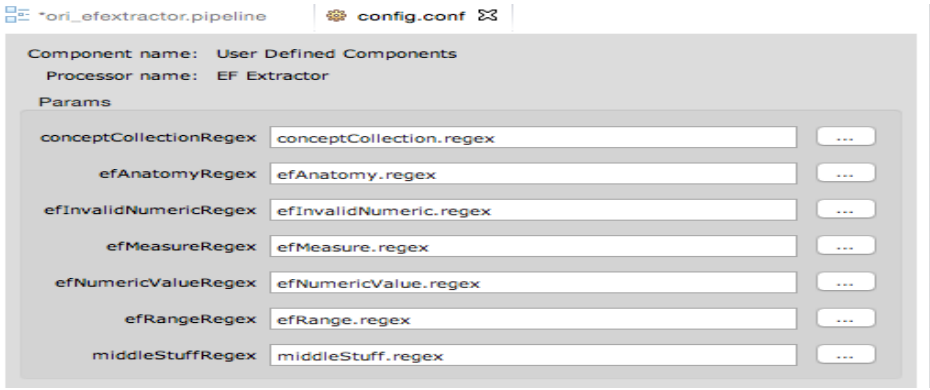
Adapting VA EF-Extractor to CLAMP

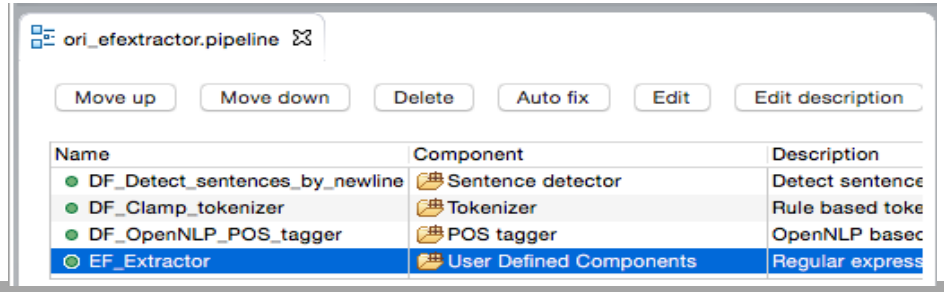
EF-Extractor by VA

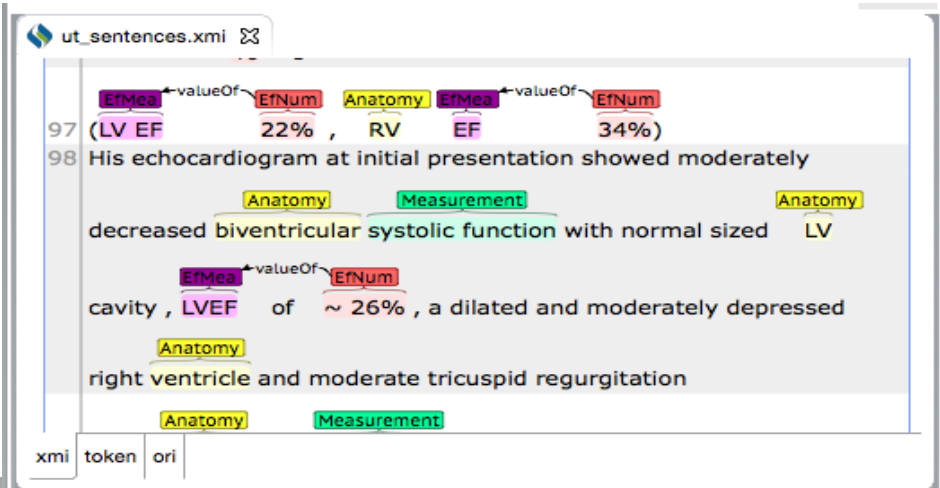
- Left ventricular ejection fraction (EF) is a key component of heart failure quality measures used within the Department of Veteran Affairs (VA)
- EF-Extractor use regular expressions and rules to capture the EF;
- It is UIMA based and can run on UIMA AS;
- We integrate the EF-Extractor as an ‘user-defined-component’ into CLAMP;

Integrate EF_Extractor into clamp





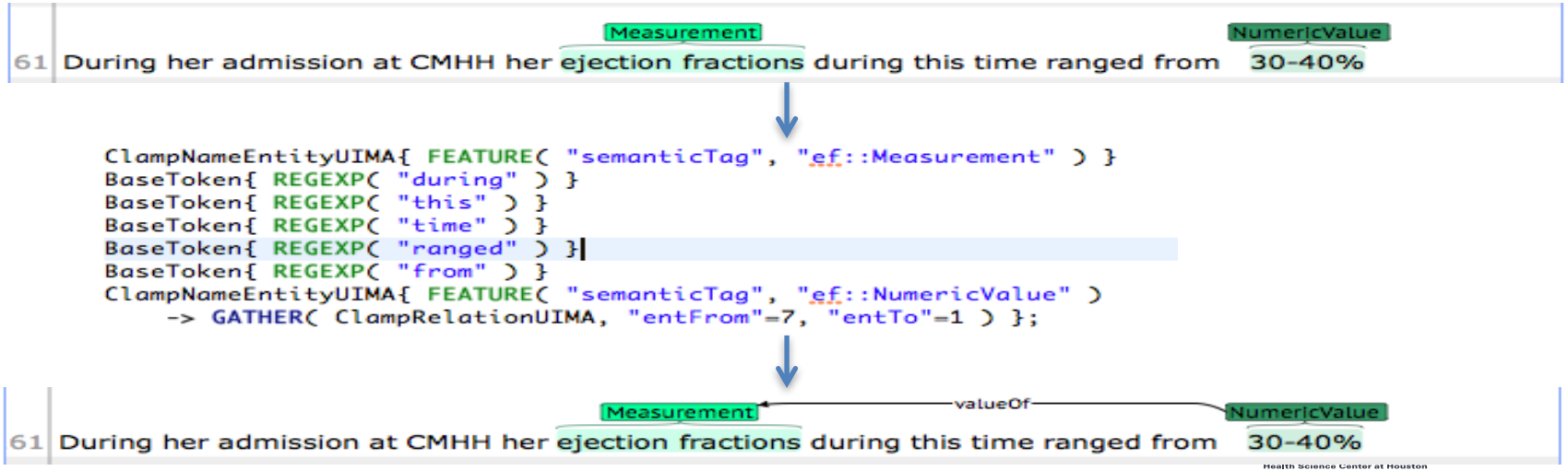




Further Improvement with clamp

Add more rules and Ruta script to further improve the performance;

Name	Component	Description
DF_Detect_sentences_by_newline	Sentence detector	Detect sentences by Newline(\n)
DF_Clamp_tokenizer	Tokenizer	Rule based tokenizer
DF_OpenNLP_POS_tagger	POS tagger	OpenNLP based pos tagger
EF_Extractor	User Defined Components	Regular expression based labtest-value extractor
DF_Ruta_script_file	Ruta rule engine	Ruta script
DF_Relation_connector_after_ruta	User Defined Components	set relation name after ruta script;



Test on UT dataset

- Extract 200 sentences from UT notes that contain keywords like ‘ejection fraction’ or ‘EF’..
- Evaluation:
  - Concept level: Recognize all EF mentions and related values;

Algorithm	Dataset	Precision	Recall	F1
Original	VA *	95%	88.9%	0.919
Original	UTHealth	83.0%	84.7%	0.838
Customized	UTHealth	98.2%	89.1%	0.934

**Garvin JH et al JAMIA 2012*

CLAMP – It's all about transportability

- Building customized pipelines for different applications using default components
- Integrating existing components/tools into CLAMP
- Exporting CLAMP pipelines to other NLP platforms