Observational Health Data Sciences and Informatics (OHDSI)

Natural Language Processing (NLP) Working Group

09/13/2017

AGENDA

- Presentations:
  - Xu Lab – Ergin Soysal, Jingqi Wang

CLAMP GUI for Population of OMOP NLP Tables

Xu Lab

The University of Texas Health Science Center at Houston

CLAMP

- 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**"
  - Fast deployment to various setting
  - Task management
  - Visual analytics

A track record of success in clinical NLP research

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

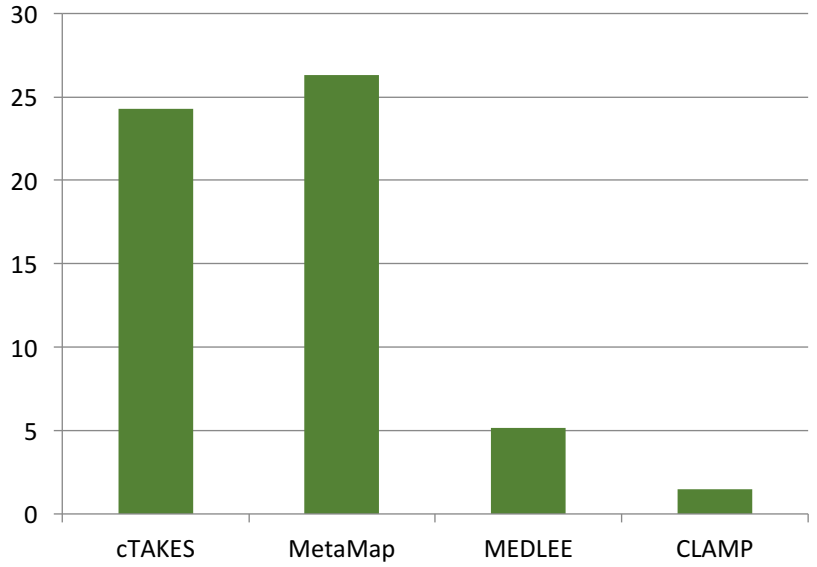
			Exact match			Relaxed match		
Corpus	Entity types	# entity	P	R	F1	P	R	F1
MTsamples	treatment, problem, test	25,531	0.841	0.811	0.826	0.921	0.890	**0.905**
i2b2	treatment, problem, test	72,846	0.891	0.861	0.876	0.958	0.925	**0.941**
UTNotes	treatment, problem, test	124,869	0.921	0.900	0.910	0.963	0.940	**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

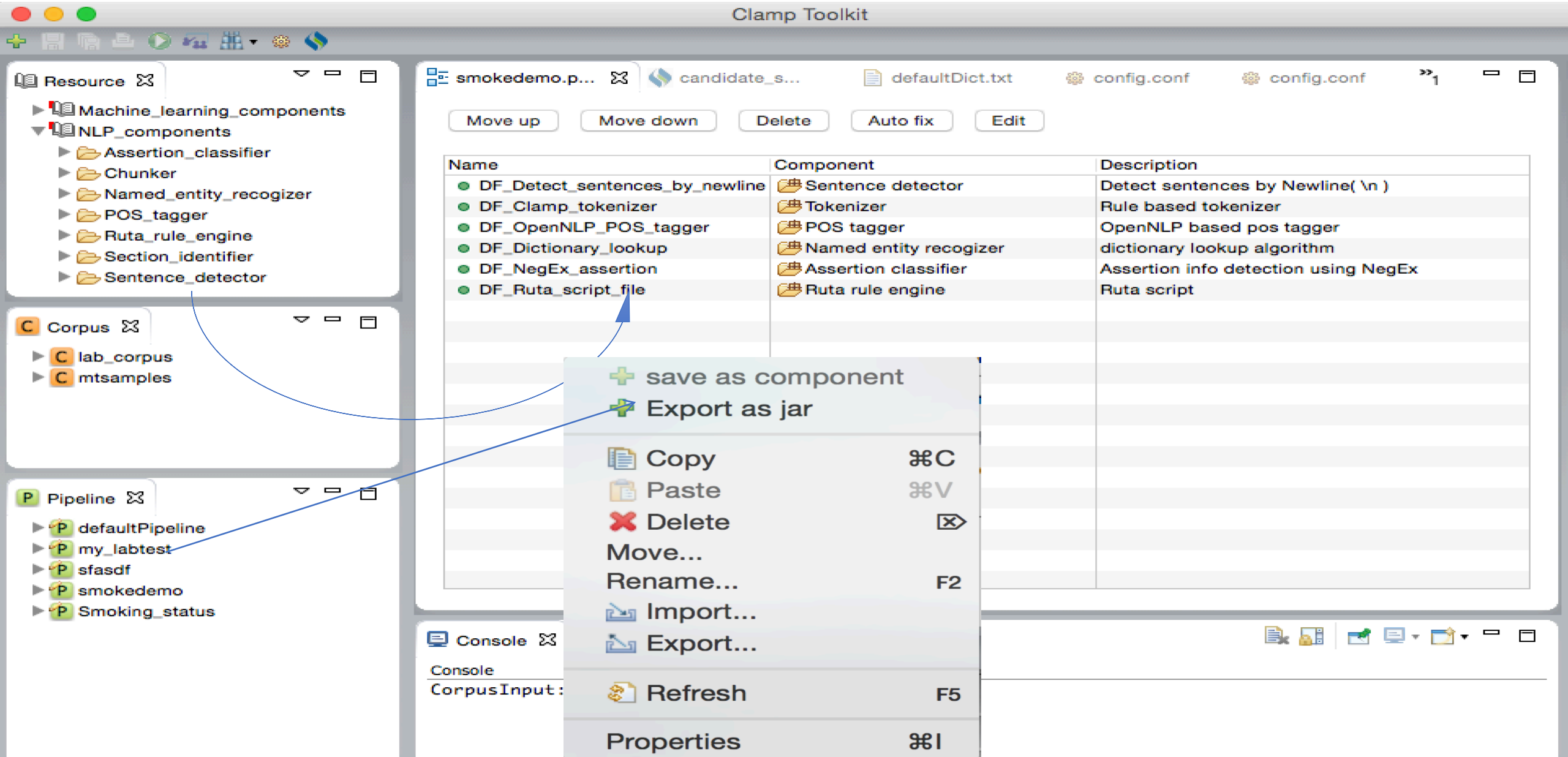
Pipeline	MAC		Linux		Windows	
	Single thread	Multi threads	Single thread	Multi threads	Single thread	Multi threads
clamp-ner	0.72	0.28	**1.25**	**0.17**	0.69	0.302
clamp-ner-attribute	0.93	0.38	1.59	0.24	0.90	0.422
disease-attribute	0.62	0.26	1.06	0.17	0.58	0.296
lab-attribute	0.62	0.26	0.99	0.16	0.56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

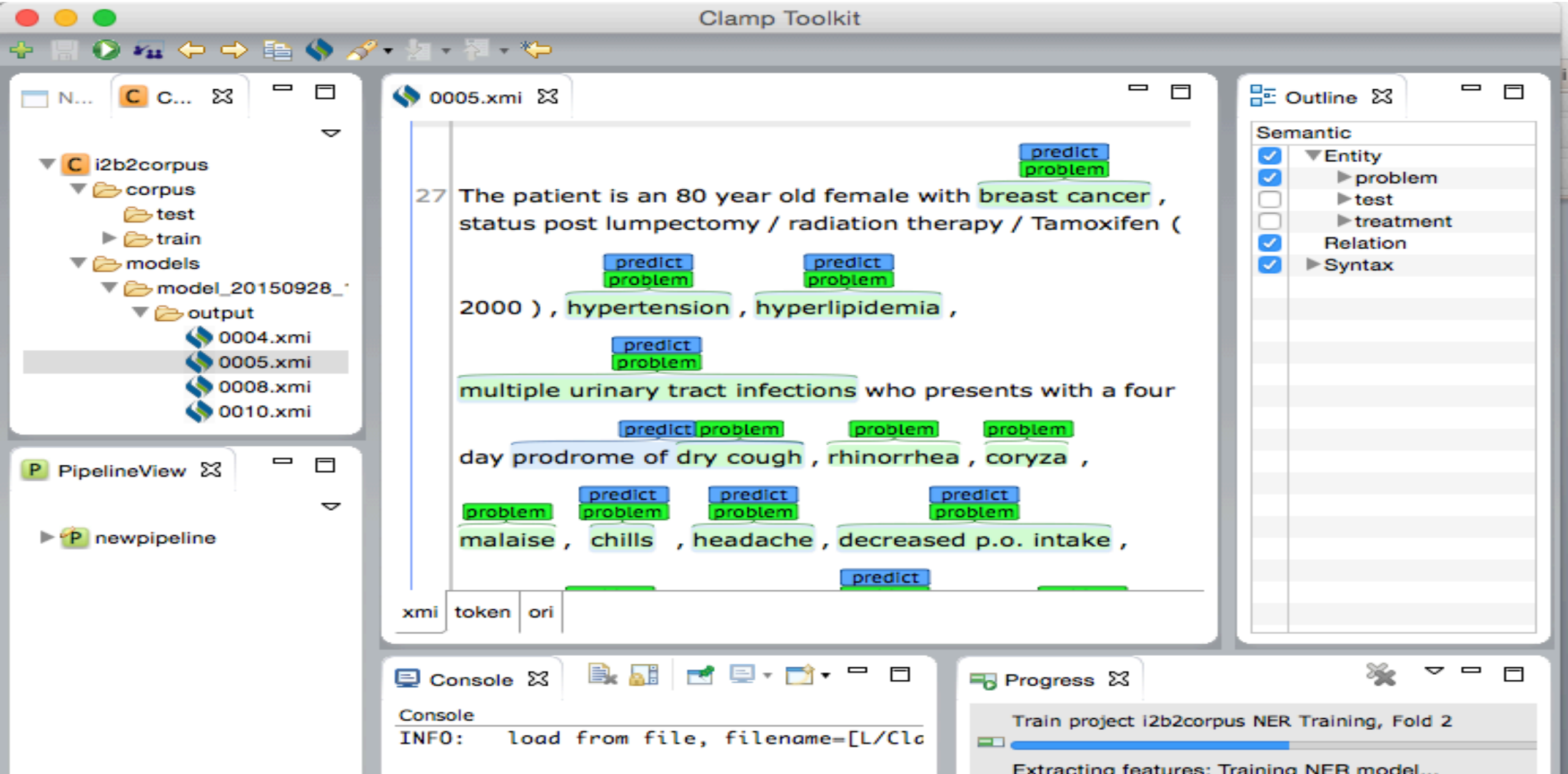


SemEval 2015 Corpus: 431 documents, avg doc size: 9.38k
SINGLE THREAD

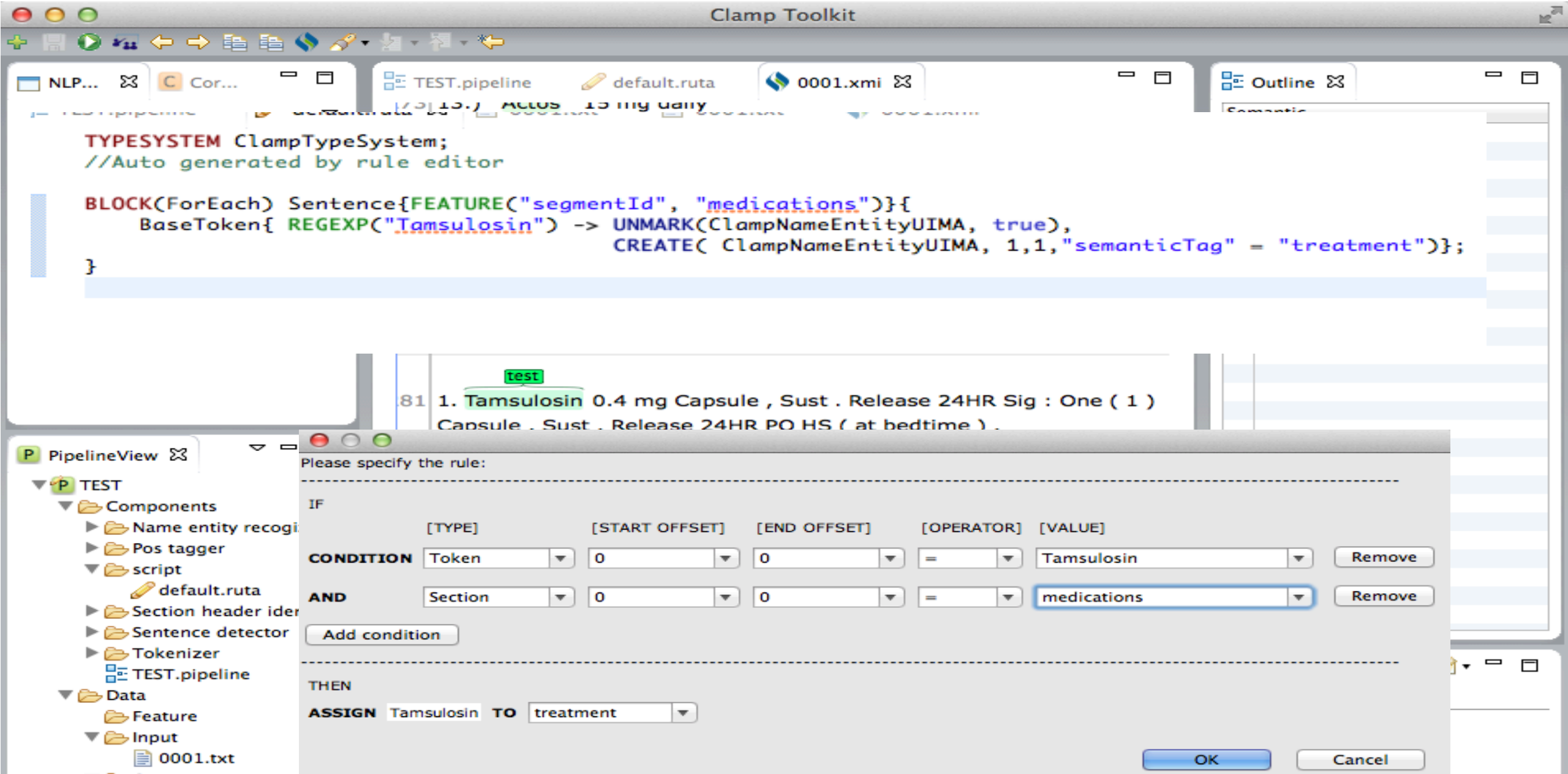
CLAMP-GUI: Building your own pipeline



CLAMP-GUI: Annotating/Re-training

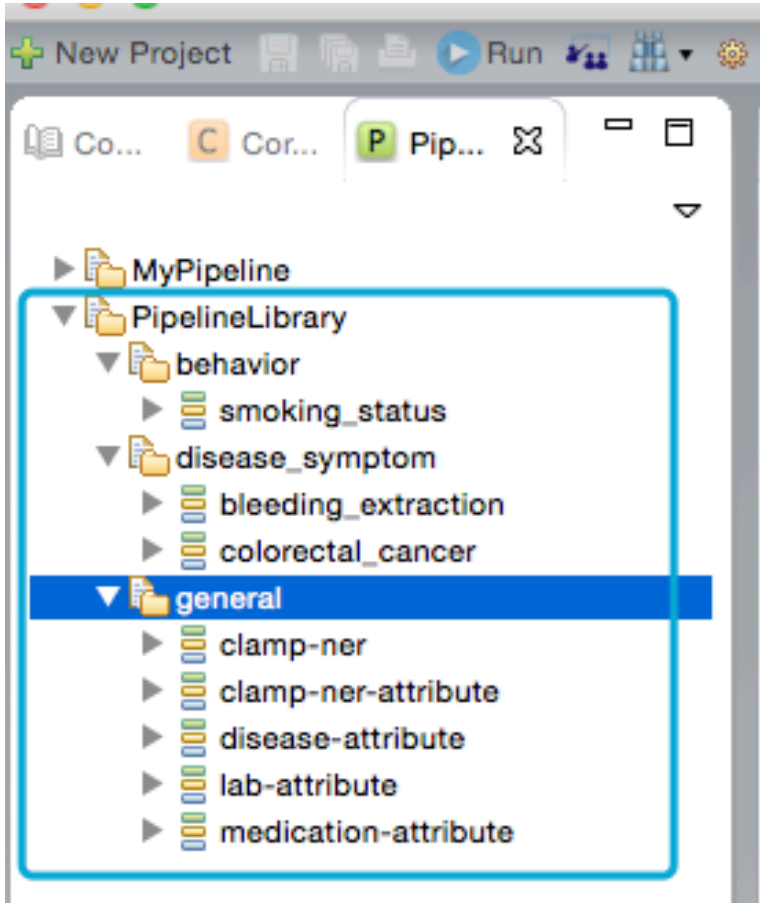


CLAMP-GUI: Specifying rules



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
  - …

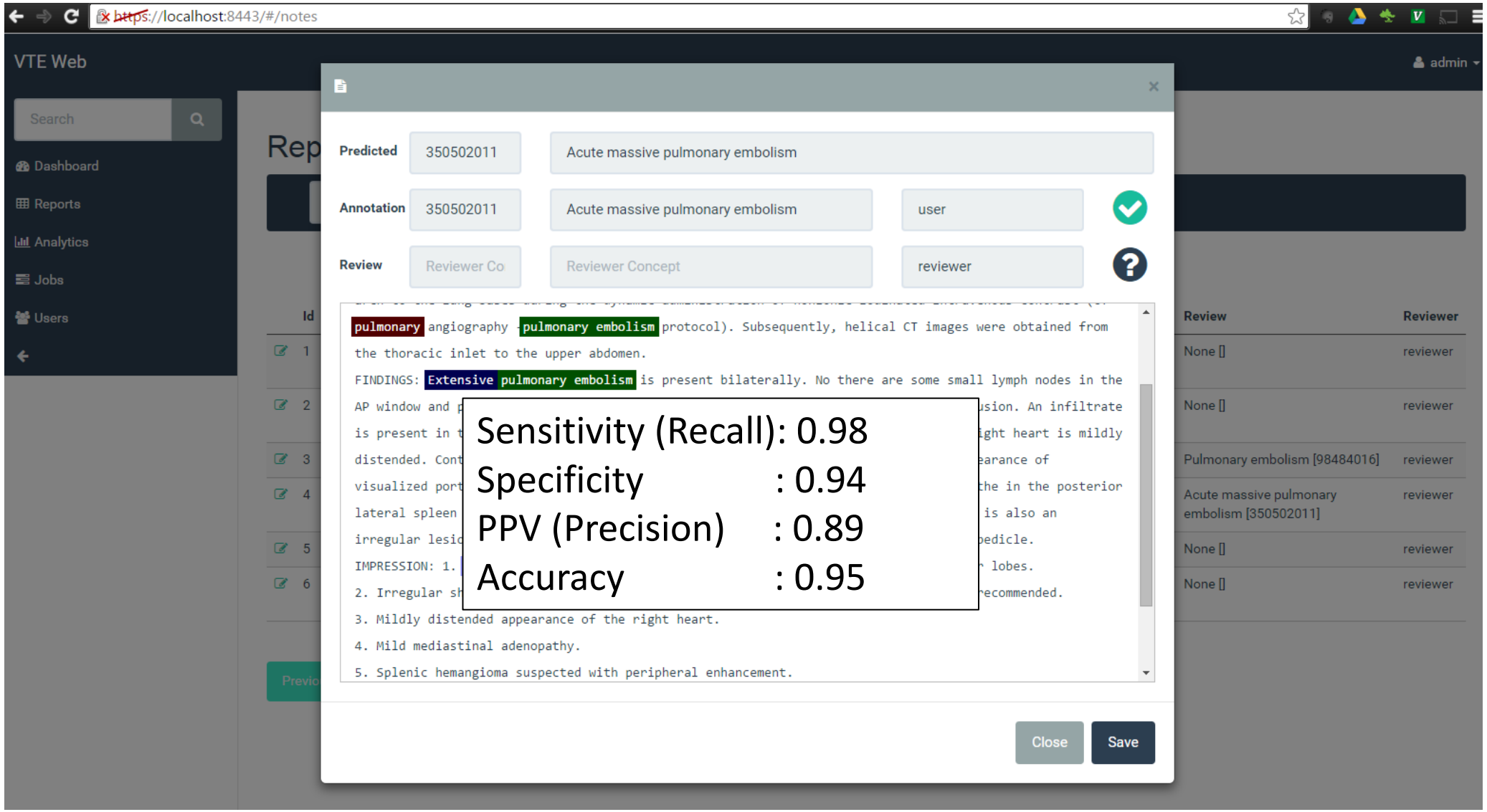


Some implementations

- ICD-10 Encoding
  - Assistance to medical coders
- Cancer Registry
  - ICD-O encoding at Cancer Registry
- Patient Safety Indicators:
  - Blood loss during surgery
  - Iatrogenic Organ Damage
  - Postop dehiscence

Applications

Meaningful use quality measurement – VTE detection at Memorial Hermann Hospital



CLAMP for CDM-NLP

NOTE

- The NOTE table captures unstructured information that was recorded by a provider about a patient in free text notes on a given date.
- Metadata

Field	Required	Type	Description
note_id	Yes	integer	A unique identifier for each note.
person_id	Yes	integer	A foreign key identifier to the Person about whom the Note was recorded. The demographic details of that Person are stored in the PERSON table.
note_date	Yes	date	The date the note was recorded.
note_datetime	No	datetime	The date and time the note was recorded.
note_type_concept_id	Yes	integer	A foreign key to the predefined Concept in the Standardized Vocabularies reflecting the type, origin or provenance of the Note.
note_class_concept_id	Yes	integer	A foreign key to the predefined Concept in the Standardized Vocabularies reflecting the HL7 LOINC Document Type Vocabulary classification of the note.
note_title	No	string(250)	The title of the Note as it appears in the source.
note_text	No	RBDMS dependent text	The content of the Note.
encoding_concept_id	Yes	integer	A foreign key to the predefined Concept in the Standardized Vocabularies reflecting the note character encoding type.
language_concept_id	Yes	integer	A foreign key to the predefined Concept in the Standardized Vocabularies reflecting the language of the note.
provider_id	No	integer	A foreign key to the Provider in the PROVIDER table who took the Note.
visit_occurrence_id	No	integer	Foreign key to the Visit in the VISIT_OCCURRENCE table when the Note was taken.

NOTE_NLP

- The NOTE table encodes all output of NLP on clinical notes.
- Each row represents a single extracted term from a note.

Field	Required	Type	Description
note_nlp_id	Yes	Big Integer	A unique identifier for each term extracted from a note.
note_id	Yes	integer	A foreign key to the Note table note the term was extracted from.
section_concept_id	No	integer	A foreign key to the predefined Concept in the Standardized Vocabularies representing the section of the extracted term.
snippet	No	string(250)	A small window of text surrounding the term.
offset	No	string(50)	Character offset of the extracted term in the input note.
lexical_variant	Yes	string(250)	Raw text extracted by the NLP tool.
note_nlp_concept_id	No	integer	Foreign key to Concept table. Represents the normalized concept for extracted term. Domain of the term is represented as part of the Concept table.
note_nlp_source_concept_id	No	integer	A foreign key to a Concept that refers to the code in the source vocabulary used by the NLP system.
nlp_system	No	string(250)	Name and version of the NLP system that extracted the term.
nlp_date	Yes	date	The date of the note processing.
nlp_date_time	No	datetime	The date and time of the note processing.
term_exists	No	Boolean	Term_exists is defined as a flag that indicates if the patient actually has or had the condition. Any of the following modifiers would make Term_exists false: Negation = true; Subject = [anything other than the patient]; Conditional = true; Rule_out = true; Uncertain = very low certainty or any lower certainties. A complete lack of modifiers would make Term_exists true. For the modifiers that are there, they would have to have these values: Negation = false; Subject = patient; Conditional = false; Rule_out = false; Uncertain = true or high or moderate or even low (could argue about low).
term_temporal	No	string(50)	Term_temporal is to indicate if a condition is “present” or just in the “past”. The following would be past: History = true; Concept_date = anything before the time of the report.
term_modifiers	No	string(2000)	Describes compactly all the modifiers extracted by nlp system. For example, “son has rash” → “negated=no,subject=family,certainty=undef,conditional=false,general=false”. Value will be saved as one of the modifiers.

1 Sample Type / Medical Specialty: Discharge Summary

3 Sample Name: Discharge Summary - 6

5 Description: A white male veteran with

problem
multiple comorbidities, who has a history of

BDL_Of
problem
timeof
BDL
bladder cancer diagnosed approximately

timeof
temporal
two years ago by the VA Hospital.

7 (Medical Transcription Sample Report)

9 ADMISSION DATE: MM/DD/YYYY

11 DISCHARGE DATE: MM/DD/YYYY

13 HISTORY OF PRESENT ILLNESS: Mr. ABC is a 60-year-old white male veteran with

problem
multiple comorbidities, who has a history of

BDL_Of
problem
timeof
BDL
bladder cancer diagnosed approximately

timeof
temporal
two years ago by the VA Hospital. He

treatment
underwent a resection there. He was to be

treatment
admitted to the Day Hospital for cystectomy. He was seen in Urology Clinic and Radiology Clinic on MM/DD/YYYY.

echocardiogram 1 of 6

section_concept_id	snippet	offset	lexical_variant	nlp_system	note_nlp_concept_id	nlp_date	nlp_date_tim			ers
description	: A white male veteran with multiple comorbidities , who has a history of bladder cancer	138-151	comorbidities	CLAMP	C0009488	09/12/2017	09/12/2017 16:24:23	True		
description	comorbidities , who has a history of bladder cancer diagnosed approximately two years ago	174-188	bladder cancer	CLAMP	C0005684	09/12/2017	09/12/2017 16:24:23	True	two years ago	BDL=[bladder], temporal=[two years ago]
history_present_illness	- old white male veteran with multiple comorbidities , who has a history of bladder cancer	432-445	comorbidities	CLAMP	C0009488	09/12/2017	09/12/2017 16:24:23	True		
history_present_illness	comorbidities , who has a history of bladder cancer diagnosed approximately two years ago	468-482	bladder cancer	CLAMP	C0005684	09/12/2017	09/12/2017 16:24:23	True	two years ago	BDL=[bladder], temporal=[two years ago]
history_present_illness	ago by the VA Hospital . He underwent a resection there . He was to be admitted to the	555-566	a resection	CLAMP	C0198907	09/12/2017	09/12/2017 16:24:23	True		
history_present_illness	to be admitted to the Day Hospital for cystectomy . He	621-631	cystectomy	CLAMP	C0010651	09/12/2017	09/12/2017 16:24:23	True		

Discussion

- Section standardization
  - Normalization is tricky
  - Different note types requires different hierarchy
- note_nlp_concept_id(integer) – Standardized Vocabulary note_nlp_source_concept_id(integer) -> CUI (string)
- term_exists: requires a special postprocessing
- term_temporal: present, past, future(?)
  - What's past?
  - Concrete point in time phrases
  - Relative events

Discussion

- term_modifiers
  - relationships?
    - Disease - body_location, medication – dose, lab - value
  - What if 2 terms are related to each other?
    - procedure – intent (disease)