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	2009 i2b2, medication	#2
recognition	2010 i2b2 problem, treatment, test	#2
	2013 SHARe/CLEF abbreviation	#1
UMLS encoding	2014 SemEval, disorder	#1
Relation	2012 i2b2 Temporal	#1
extraction	2015 SemEval Disease-modifier	#1
	2015 BioCREATIVE Chemical-induced disease	#1
	***	UTHealth School of Information

Biomedical

3

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

• •	Clamp Toolkit						
🕂 🗆 🗅 ڬ 🐼 🛲 • 👄 💊							
🕼 Resource 🔀 🛛 🗖 🗖	📴 smokedemo.p 🔀 🔇 candidate	_s 📄 defaultDict.txt	🏶 config.conf 🛛 🏶 config.conf 🎽 🗖				
Machine_learning_components IntP_components Acception classifier	Move up Move down D	Auto fix Edit					
Assertion_classifier	Name	Component	Description				
Named entity recogizer	DF_Detect_sentences_by_newline	😕 Sentence detector	Detect sentences by Newline(\n)				
▶ > POS_tagger	DF_Clamp_tokenizer	Tokenizer	Rule based tokenizer				
Ruta_rule_engine	DF_OpenNLP_POS_tagger	POS tagger	OpenNLP based pos tagger				
Section_identifier	 DF_Dictionary_lookup 	Mamed entity recogizer	dictionary lookup algorithm				
Sentence detector	DF_NegEx_asser/on	Assertion classifier	Assertion into detection using NegEx				
	• DF_Huta_script_file	Buta rule engine	Ruta script				
	🕂 save as o	component					
C mtsamples	Export as	s jar					
	🖹 Сору	жс					
	Paste	жv					
Pipeline 🛛 🗸 🗖 🗖	🔀 Delete	\mathbf{X}					
P defaultPipeline	Move						
P my_labtest	Rename	F2					
P smokedemo	🔄 🔤 Import						
▶ P Smoking_status	Export						
	E Console 🖾		🗟 🔜 📑 🖃 ד 🗂 ד 🗖 ד				
	Console & Refresh	F5					
	Properties	жı					

Annotating/Re-training



Specifying rules

e o o	-			Clamp	Toolkit					E A
□ NLP 😫 🖸 Cor □		TEST.pipeline	default.rut	a 🔇	0001.xmi	×		E Outline	×	- 8
TYPESYSTEM ClampTypeSystem; //Auto generated by rule editor										
<pre>BLOCK(ForEach) Sentence{FEATURE("segmentId", "medications")}{ BaseToken{ REGEXP("Tamsulosin") -> UNMARK(ClampNameEntityUIMA, true),</pre>							F;			
81 1. Tamsulosin 0.4 mg Capsule , Sust . Release 24HR Sig : One (1) Capsule , Sust . Release 24HR PO HS (at bedtime) .										
P PipelineView 🖾 🗸 🗸	Please specify	the rule:								
TEST	15									
 Components Components Name entity recogi: 	11	[TYPE]	START OFFSE	TI LEN	D OFFSET1	[OPERATO	R] [VALUE]			
► Costagger ▼ Costagger ▼ Costagger ♥ Costagg		Token	▼ 0 ▼ 0	▼ 0 ▼ 0	•		Tamsulosin medications	▼	Remove Remove	
Section header ider Section header ider Sentence detector	Add conditi	00								
 ▶ ➢ Tokenizer ₩ TEST.pipeline ▼ ➢ Data 	THEN)•
➢ Feature ▼ ➢ Input ◎ 0001.txt	ASSIGN Tam	nsulosin TO tro	eatment 💌					ОК	Cancel	

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

🐈 New Project 📳 📄 📄 🍉 Run 🚧 🏭 🔹 🥎	Tutorial A a			
Corpus V	E CHF_Pipeline.pipeline			🔇 1019_Discha 🔇 10426_Disch 🦳 CHF-Pipelin 🔇 1019_Discha 🔇 11895_Disch 🔇 10426_Disch 🕄 🕻
ML_components MLP_components	Move up Move down	Delete Auto fix Edit	Edit description	1 SUBJECT_ID : 13868
	Name	Component	Description	2 NOTE_DATE : 2129-04-15
	DF Clamp sentence detector	A Sentence detector	Bule based sentence detector	3 DOCUMENT_ID : 10426
	DF_Clamp_tokenizer	A Tokenizer	Rule based tokenizer	4 DOCUMENT_TYPE : Discharge summary
	DF OpenNLP POS tagger	POS tagger	OpenNLP based pos tagger	5 Admission Date : [**2505-4-3**] Discharge Date : [**2505-4-10**]
	DF_Dictionary_based_section	Bection identifier	Dictionary based section header Identifier	6 Date of Birth : [**2435-3-22**] Sex : F
	DF_Regular_expression_NER	Barned entity recogizer	regular expression named entity recognition	7 Service :
	DF_CRF_based_named_entity	Barned entity recogizer	Name entity recognition using CRF	
	Medex_Dictionary_lookup	Named entity recogizer	dictionary lookup algorithm	
	medication_ruta_script	/ Ruta rule engine	Ruta script	BODYLOC
	DF_Drug_Attribute_Connector	/ User Defined Component	s connect drug and attributes recognized by clamp;	United and the second se
	labtest_ruta_script	🕮 Ruta rule engine	Ruta script	8 HISTORY OF PRESENT ILLNESS : Patient is a 70-year-old female with a history of coronary artery disease suffered
	DF_Dictionary_based_UMLS_e		umls encoding algorithm	5
				selor
				SEVERE DOUGLAND (Lest)
	DESCRIPTION:			an acute myocardial infarction in [**2504-4-9**], and she was taken to catheterization laboratory and found three
	The pipeline is to extract CHF related in	nformation including CHF terms	, lab test, medication and image test	
INPUT: Discharge summaries				
P Pipeline 🕱 🗸 🗖	CATEGORY: Cardiovascular Disease	15 15		with successfully stented left anterior descending artery , found to have left ventricular diastolic dysfunction with
V Comparison MyPipeline				
CHF_Pipeline				test - value0f - sevenue - acco - acco - acco
Components				a presented election of 5704, with anterior acids duckingsis and anterplateral hundkingsis
CHF_Pipeline.pipeline				a preserved ejection naction of 37.70 with anterior apical dyskinesis and anterolateral hypokinesis
Assertion classifier				
Chunker				BODYLOC
Named entity recogizer				Brodens
POS tagger				9 In [**2504-11-9**], the patient returned to [**Hospital3 **] for chest pain
Huta rule engine	Console Rogress 🔀			locOf
Section identifier	No operations to display at this time			test BODYLOC treatment
Sentence detector	no operations to display at this time.			10 Catheterization revealed totally occluded left anterior descending artery with brachytherapy
Internizer				
Liser Defined Components				
Data				Test value0f sources MODERATE
- Duita		1	·	11 Echogardingram in [##2504-11-9##] showed an election fraction of 50%
			1019_Discharge summary.txt.txt	In Echocal dogram in [250+11-9] showed an ejection inaction of 50% , This symmetric fet vehiclean hypera
			1019_Discharge summary.xmi.xmi	xmi token ori
			10426_Discharge summary.txt.txt	
			V 10426_Discharge summary.xmi.xmi	Console X Progress
			11895_Discharge summary.txt.txt	G

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-definedcomponent' into CLAMP;



Integrate EF_Extractor into clamp

Component S Components Component Components Components Components Components Compon				tori	_efextractor.pipeline	In the second secon		
Image: Section - Legender Components Image: Section - Legender Component Image: Section - Legender Component Description Image:	🕼 Component 🖾 [Cor	rpus	~	Com	ponent name: User D	efined Components		
• ****************************	▶ 💷 ML_components			Pro	cessor name: EF Ext	ractor		
 Assertion_classifier Assertion_classifier Assertion_classifier Chunker Chunker Chunker Chunker Chunker Chunker Chunker Chunker PoS_tagger Ruta_rule_engine Section_identifier Section_ident	Image: NLP_components			Par	ams			
Chunker Chu	Assertion_classifi	ier		con	ceptCollectionRegex	conceptCollection.regex		
Addescriptor	🕨 🗁 Chunker				efAnatomyRegex	efAnatomy.regex		
Named_entity_recogizer POS_tagger POS tagger	descriptor			efI	nvalidNumericRegex	efInvalidNumeric.regex		
POS_tagger POS_tagge	Named_entity_rec	cogizer			-			
<pre> File RangeRegex efRange.regex File Components File Extractor File Component File Edit description File Components File Edit description File Components File Edit description File Component File Components File Edit description File File Components File Components File Edit description File File Components File Edit description File File Components File Component File Component File Component File Components File Component File Components File Compon</pre>	POS_tagger				ermeasurekegex	etMeasure.regex		
<pre>erRangeRegex erRangeRegex middleStuff.regex erRangeRegex middleStuff.regex midd</pre>	Ruta_rule_engine			ef	NumericValueRegex	efNumericValue.regex		
Sentence_detector Some concert Tokenizer WildleStuff.regex middleStuff.regex middle	Section_identifier	r			efRangeRegex	efRange.regex		
 Concerning the second of the se	Sentence_detect	or			middleStuffRegex	middleStuff.regex		
 Wills_encoder User_Defined_Components EF_Extractor DF_Dysplasia_document_summarization DF_Relation_connector_after_ruta Move up Move down Delete Auto fix Edit <l< td=""><td>🕨 🗁 Tokenizer</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></l<>	🕨 🗁 Tokenizer							
 Components Components Components Component connector_after_ruta Component Component Description Component Rule based toke Component	UMLS_encoder							
 	User_Defined_Co	mponents		tu 🍫 🛛	t_sentences.xmi 🔀			
 ^A DF_Drug_Attribute_Connector ^A DF_Dysplasia_document_summarization ^A DF_Relation_connector_after_ruta ^A DF_Relation_connector_after_ruta ^A atomy ^A atomy	EF_Extractor					-		
 	DF_Drug_Attri	bute_Connector			EfMea valueOf Ef	Num Anatomy EfMea	alueOf EfNum	
> DF_Relation_connector_after_ruta > P B His echocardiogram at initial presentation showed moderately > ori_efextractor.pipeline X Move up Move down Delete Auto fix Edit Edit description OF_Detect_sentences_by_newline Sentence detector DF_Clamp_tokenizer Detect sentences DF_OpenNLP_POS_tagger POS tagger DF_Cettractor Post tagger OpenNLP basec Xmit Xmit token	▶ 🗁 DF_Dysplasia_	document_summarization		97	(LV EF 2	2%, RV EF	34%)	
Anatomy Measurement Anatomy Move up Move down Delete Auto fix Edit Edit description ame Component Description DF_Detect_sentences_by_newline Sentence detector Detect sentence DF_Clamp_tokenizer Image: Component of the co	▶ 🗁 DF_Relation_c	onnector_after_ruta		98	His echocardiog	am at initial presenta	tion showed moderate	ly
wri_efextractor.pipeline X Move up Move down Delete Auto fix Edit Edit description ame Component Description DF_Detect_sentences_by_newline Sentence detector Detect sentences DF_Clamp_tokenizer @ Tokenizer Rule based toke DF_OpenNLP_POS_tagger @ POS tagger OpenNLP basec DEF_Extractor @ User Defined Components Regular express				_	Ana	tomy Measurement		natomy
Move up Move down Delete Auto fix Edit Edit description DF_Detect_sentences_by_newline Sentence detector Detect sentence DF_Clamp_tokenizer Brokenizer Rule based toke DF_OpenNLP_POS_tagger BPOS tagger OpenNLP basec DF_Extractor Buser Defined Components Regular express	ori, efextractor pipeline				decreased biven	tricular systolic funct	on with normal sized	ÎV
Move up Move down Delete Auto fix Edit Edit description ame Component Description DF_Detect_sentences_by_newline Bestence detector Detect sentence DF_Clamp_tokenizer Bestence Description DF_OpenNLP_POS_tagger BPOS tagger OpenNLP basec DF_EF_Extractor Buser Defined Components Regular express	on_elextractor.pipeline Ki							
ame Component Description DF_Detect_sentences_by_newline Bestime cave and the sentence DF_Clamp_tokenizer Bestime cave and the sentence DF_OpenNLP_POS_tagger Bestime components DF_Extractor Bestime components	Move up Move down D	elete Auto fix Edit	Edit description		EfMea	EfNum		
Amage Component Description DF_Detect_sentences_by_newline Bestence detector Detect sentence DF_Clamp_tokenizer Bestence Bestence DF_OpenNLP_POS_tagger Bestence OpenNLP basect DF_EF_Extractor Buser Defined Components Regular express					cavity , LVEF	of ~ 26% , a dilated	and moderately depre	essed
DF_Detect_sentences_by_newline	Name	Component	Description		Anatomy			
DF_Clamp_tokenizer (#) Tokenizer Rule based toke DF_OpenNLP_POS_tagger (#) POS tagger OpenNLP basec DF_Extractor (#) User Defined Components Regular express	DF_Detect_sentences_by_newline	Bentence detector	Detect sentence		right ventricle a	nd moderate tricuspid	regurgitation	
DF_OpenNLP_POS_tagger Components Regular express EF_Extractor Components Regular express	DF_Clamp_tokenizer	进 Tokenizer	Rule based toke		Anatomy	Measurement		
EF_Extractor III User Defined Components Regular express xmi token ori	DF_OpenNLP_POS_tagger	POS tagger	OpenNLP basec			measurement		
	EF_Extractor	Ber Defined Components	Regular express	xmi	token ori			

Further Improvement with clamp

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

Name	Component	Description
DF_Detect_sentences_by_newline	A 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

