OHDSI NLP WG Monthly Meeting

03/13/2019

Agenda

- Introduction of New Members
- **Leveraging and Enriching Common Data Model towards Portable Clinical NLP System** – Yuan Luo
- Ongoing projects
- Other issues

PRESENTATION

Leveraging and Enriching Common Data Model towards Portable Clinical NLP System

Yuan Luo

Leveraging and Enriching Common Data Model towards Portable Clinical NLP System

Yuan Luo

Assistant Professor
Department of Preventive Medicine
Departments of IEMS and EECS (Courtesy)
Northwestern University
yuan.luo@northwestern.edu
@yuanhypnosluo

NORTHWESTERN UNIVERSITY FEINBERG SCHOOL OF MEDICINE

3/13/2019

Northwestern University

Introduction

- We introduce portability to NLP-driven phenotyping of unstructured clinical records
- We present a portable phenotyping system that facilitates portability across different institutions and data systems
- The portability is introduced by storing key components of rule-based NLP systems' and standard NLP pipelines' results as annotations using the format defined in OMOP CDM
- Experimental results on i2b2's Obesity Challenge show the feasibility of our system

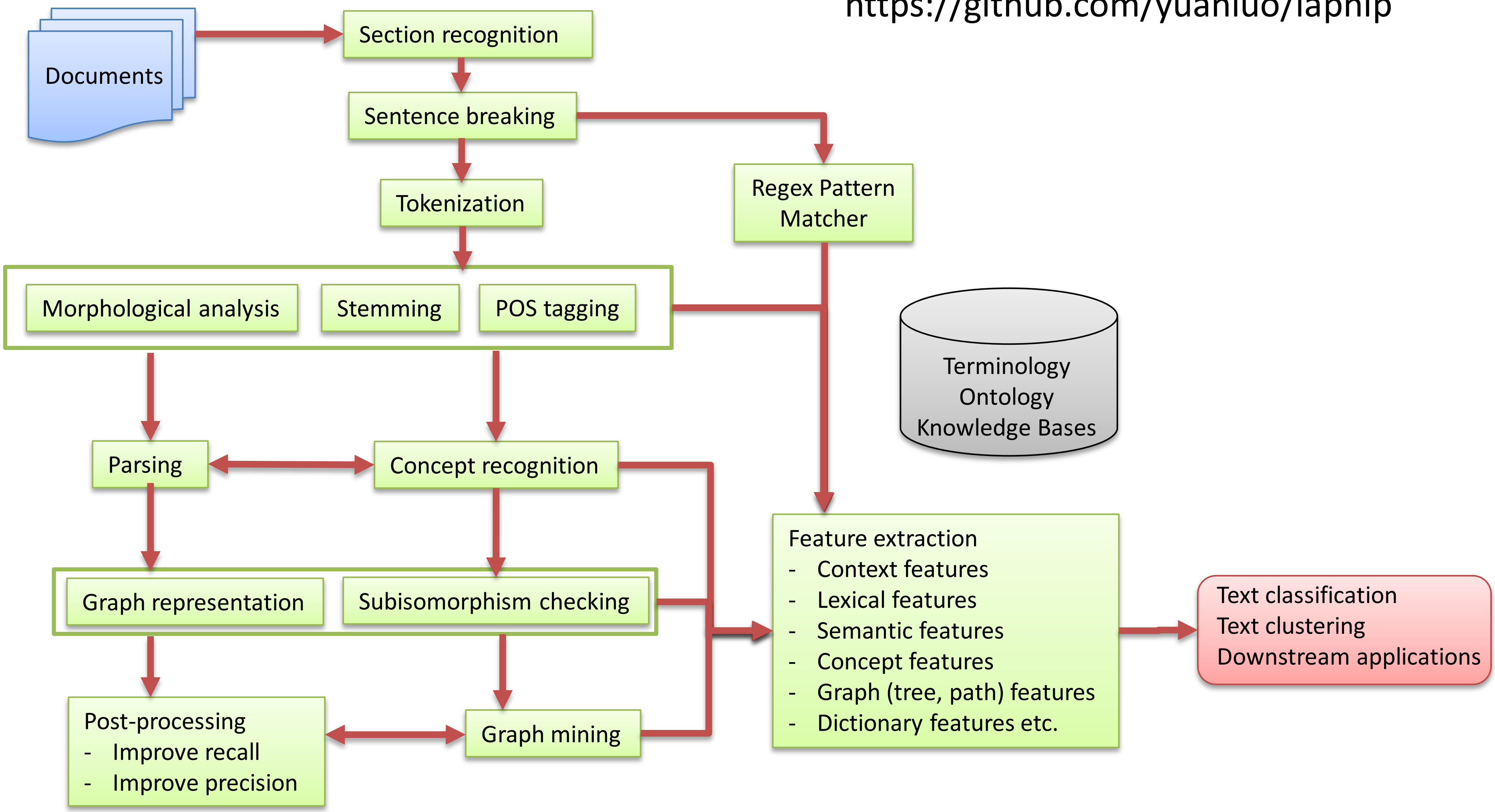
O. Uzuner, "Recognizing Obesity and Comorbidities in Sparse Data," *Journal of the American Medical Informatics Association*, vol. 16, no. 4, pp. 561-570, 2009.

Clinical Note Processing

- Deabbreviation: all abbreviations are translated back to full terms
- Section and boundary detection: record the start and end position of each section
- Rule-based components annotation: annotate the key components by rule-based methods
- Annotation Feature Extraction and Mapping: parse the files by MetaMap to extract CUIs
- Annotation storing: store annotations in OMOP CDM tables (Note and Note_NLP tables)

Bigger Picture - NLP Workflow

https://github.com/yuanluo/lapnlp



Introduction Inline vs. Stand-off Annotation

In-line annotation

The$_{3}$ patient$_{11}$ underwent$_{21}$ an$_{24}$ ECHO$_{29}$ and$_{33}$ endoscopy$_{43}$ at$_{46}$ <PHI TYPE="Hospital">Beth$_{51}$ Israel$_{58}$ Deaconess$_{68}$ Medical$_{76}$ Center$_{83}$</PHI> on$_{86}$ <PHI TYPE="Date">April$_{92}$ 28$_{95}$</PHI>.

Stand-off annotation

Start	End	Annotation Type	Annotation Attribute
48	83	PHI	Type=Hospital
88	95	PHI	Type=Date
…	…	…	…

Y Luo, P Szolovits . Efficient queries of stand-off annotations for natural language processing on electronic medical records. *Biomedical informatics insights. 2016 Jan;8:BII-S38916.*

Selected CUIs Related Clinical Tasks

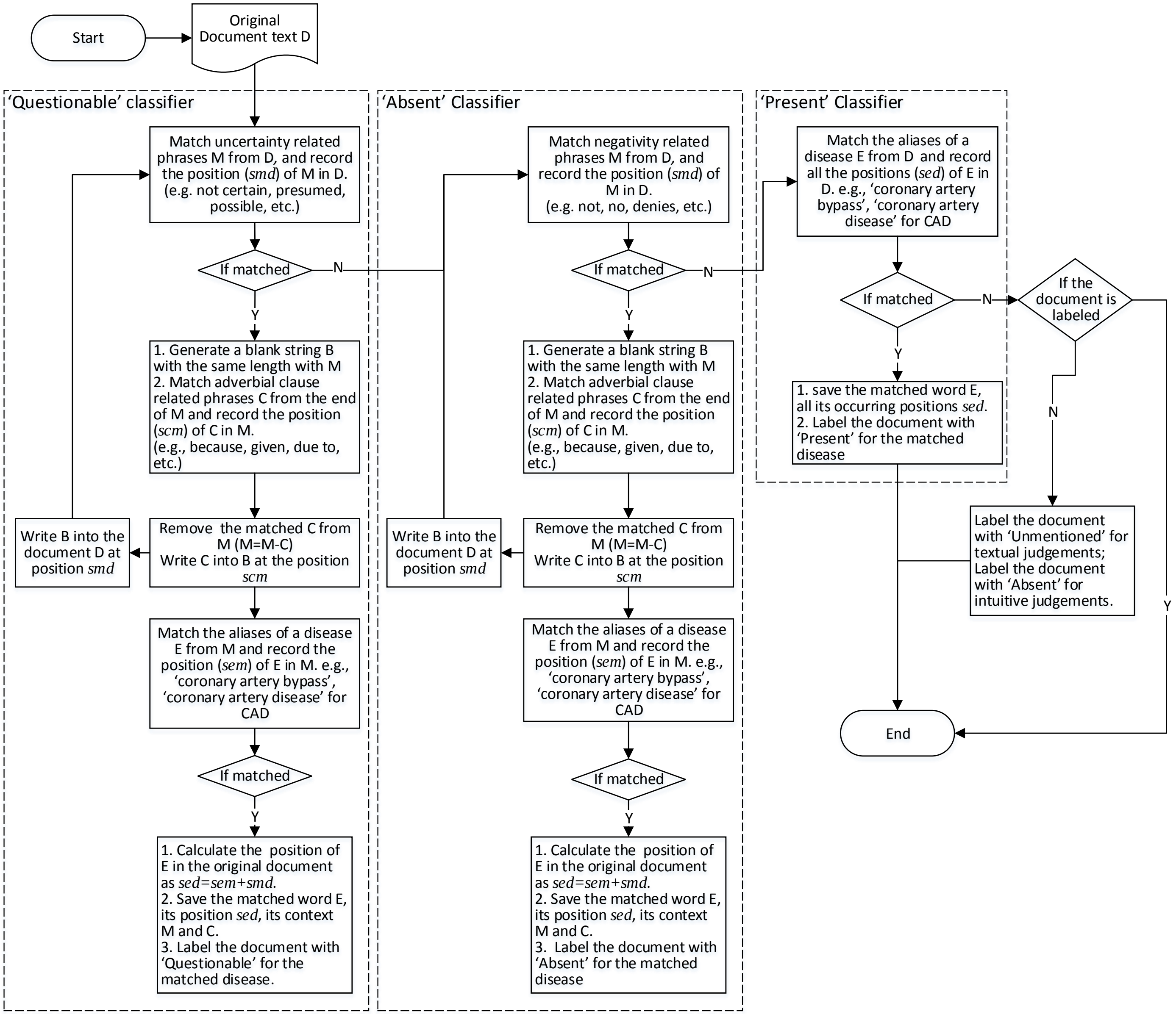
TUI	Semantic group	Semantic type description
T017	Anatomy	Anatomical Structure
T022	Anatomy	Body System
T023	Anatomy	Body Part, Organ, or Organ Component
T033	Disorders	Finding
T034	Phenomena	Laboratory or Test Result
T047	Disorders	Disease or Syndrome
T048	Disorders	Mental or Behavioral Dysfunction
T049	Disorders	Cell or Molecular Dysfunction
T059	Procedures	Laboratory Procedure
T060	Procedures	Diagnostic Procedure
T061	Procedures	Therapeutic or Preventive Procedure
T121	Chemicals & Drugs	Pharmacologic Substance
T122	Chemicals & Drugs	Biomedical or Dental Material
T123	Chemicals & Drugs	Biologically Active Substance
T184	Disorders	Sign or Symptom

W.-H. Weng, K. B. Wagholikar, A. T. McCray, P. Szolovits, and H. C. Chueh, "Medical subdomain classification of clinical notes using a machine learning-based natural language processing approach," *BMC Medical Informatics and Decision Making*, vol. 17, no. 1, 2017.

Note_NLP Table Data Elements

Column name	Description
note_nlp_id	A unique identifier for each term extracted from a note. A randomly generated auto-incremented number.
note_id	A foreign key. The note_id from the Note table from the note the term was extracted from.
section_concept_id	The representation of the section that extracted concept belongs to.
snippet	A threshold (e.g., +/- 100 characters from the end/start of the phrase)
offset	Provided by the MetaMap in the output file.
lexical_variant	The actual phrase text that MetaMap generates.
note_nlp_concept_id	The concepts or CUIs.
nlp_system	NLP tool.
nlp_date_time	Date and Time of creation/running

Anchoring Regular Expression Matches as Stand-Off Annotations



Key Components Annotation

disease	dis_pos	dis_alias	sen_pos	sentence
CHF	(50, 53)	chf	(1558, 1611)	the patient was presumed to have pneumonia versus chf

→ Questionable

disease	dis_pos	dis_alias	sen_pos	sentence
CAD	(15, 38)	coronary	(797, 836)	no evidence of coronary artery disease

→ Absent

disease: The name of the disease.
sentence: The key sentence or phrase that indicates the classification.
sen_pos: The position of the key sentence or phrase in the original record.
dis_alias: The matched alias name of the disease.
dis_pos: The matched position of this match (in the corresponding key sentence).

disease	dis_pos	dis_alias
Venous Insufficiency	(2839, 2852)	venous stasis
Venous Insufficiency	(8918, 8931)	venous stasis
OA	(3466, 3480)	osteoarthritis
Diabetes	(293, 301)	diabetes
Diabetes	(464, 472)	diabetes
Diabetes	(1676, 1684)	diabetes
Diabetes	(7874, 7882)	diabetes
Diabetes	(1647, 1655)	diabetic
CHF	(500, 524)	congestive heart failure
CHF	(1586, 1610)	congestive heart failure

→ Present

Experiments

Classifiers and parameters for grid search

Classifier	Parameter grid
LR	'C':[0.01,0.1,1,10,100]
SVM	'C':[0.01,0.1,1,10,100], 'kernel':['linear', 'rbf']
DT	'criterion':['gini','entropy']
RF	'n_estimators':[5,10,30,50,80,100], 'criterion':['gini','entropy']

LR: Logistic Regression; SVM: Support Vector Machine; DT: Decision Tree; RF: Random Forest

Experiments

- The number of each CUI represents the frequency of occurrence of the CUI in a medical record and serves as a feature of the record.
- Only using machine learning approaches on the features of the records for classification
  - Use multi-class classification algorithms to all classes (4 classes on obesity data. Y, N, Q, U)
- Integrating rule-based and machine learning based approaches for classification.
  - For major classes, use machine learning methods.
  - For minor classes, use Solt's rule-based methods [1].

I. Solt, D. Tikk, V. Gal, and Z. T. Kardkovacs, "Semantic Classification of Diseases in Discharge Summaries Using a Context-aware Rule-based Classifier," *Journal of the American Medical Informatics Association*, vol. 16, no. 4, pp. 580-584, 2009.

Experimental Results

The classification results for all classes on all CUIs corresponding to the original records

Intuitive						
	P-Micro	P-Macro	R-Micro	R-Macro	F-Micro	F-Macro
LR	0.8719	0.5792	0.8719	0.5509	0.8719	0.5618
SVM	0.8727	0.5776	0.8727	0.5537	0.8727	0.5632
DT	**0.9281**	**0.6113**	**0.9281**	**0.6116**	**0.9281**	**0.6115**
RF	0.8524	0.5626	0.8524	0.5349	0.8524	0.5454
Textual						
	P-Micro	P-Macro	R-Micro	R-Macro	F-Micro	F-Macro
LR	0.8846	0.4379	0.8846	0.4195	0.8846	0.4268
SVM	0.8886	0.4384	0.8886	0.4243	0.8886	0.4300
DT	**0.9436**	**0.5127**	**0.9436**	**0.5115**	**0.9436**	**0.5121**
RF	0.8621	0.4220	0.8621	0.4044	0.8621	0.4112

*the best results are bolded.

Experimental Results

The classification results for all classes on all CUIs corresponding to the records without family history

Intuitive						
	P-Micro	P-Macro	R-Micro	R-Macro	F-Micro	F-Macro
LR	0.8716	0.5794	0.8716	0.5503	0.8716	0.5615
SVM	0.8735	0.5780	0.8735	0.5546	0.8735	0.5640
DT	**0.9331**	**0.6159**	**0.9331**	**0.6149**	**0.9331**	**0.6154**
RF	0.8627	0.5685	0.8627	0.5462	0.8627	0.5551
Textual						
	P-Micro	P-Macro	R-Micro	R-Macro	F-Micro	F-Macro
LR	0.8836	0.4372	0.8836	0.4189	0.8836	0.4262
SVM	0.8895	0.4391	0.8895	0.4248	0.8895	0.4306
DT	**0.9475**	**0.5284**	**0.9475**	**0.5199**	**0.9475**	**0.5238**
RF	0.8618	0.4210	0.8618	0.4049	0.8618	0.4112

*the best results are bolded.

Experimental Results

The classification results for all classes on 15 types of selected CUIs corresponding to the records without family history

Intuitive						
	P-Micro	P-Macro	R-Micro	R-Macro	F-Micro	F-Macro
LR	0.9024	0.6040	0.9024	0.5763	0.9024	0.5874
SVM	0.9077	0.6055	0.9077	0.5831	0.9077	0.5924
DT	**0.9299**	**0.6131**	**0.9299**	**0.6129**	**0.9299**	**0.6130**
RF	0.8784	0.5849	0.8784	0.5559	0.8784	0.5671
Textual						
	P-Micro	P-Macro	R-Micro	R-Macro	F-Micro	F-Macro
LR	0.9145	0.4560	0.9145	0.4410	0.9145	0.4472
SVM	0.9227	**0.5832**	0.9227	0.4532	0.9227	0.4607
DT	**0.9452**	0.4878	**0.9452**	**0.4785**	**0.9452**	**0.4807**
RF	0.8830	0.4353	0.8830	0.4195	0.8830	0.4258

*the best results are bolded.

Experimental Results

The classification results for major classes on all CUIs corresponding to the original records

Intuitive						
	P-Micro	P-Macro	R-Micro	R-Macro	F-Micro	F-Macro
LR	0.8709	0.6457	0.8709	0.5733	0.8709	0.5960
SVM	0.8724	0.6444	0.8724	0.5770	0.8724	0.5981
DT	**0.9311**	**0.6804**	**0.9311**	**0.6374**	**0.9311**	**0.6488**
RF	0.8466	0.6226	0.8466	0.5559	0.8466	0.5765
Textual						
	P-Micro	P-Macro	R-Micro	R-Macro	F-Micro	F-Macro
LR	0.8882	0.7846	0.8882	0.7085	0.8882	0.7397
SVM	0.8930	0.7858	0.8930	0.7135	0.8930	0.7434
DT	**0.9545**	**0.8167**	**0.9545**	**0.7636**	**0.9545**	**0.7854**
RF	0.8882	0.7846	0.8882	0.7085	0.8882	0.7397

*the best results are bolded, the shaded results can be among the top 10 results reported in [2].

Experimental Results

The classification results for major classes on all CUIs corresponding to the records without family history

Intuitive						
	P-Micro	P-Macro	R-Micro	R-Macro	F-Micro	F-Macro
LR	0.8723	0.6473	0.8723	0.5741	0.8723	0.5970
SVM	0.8732	0.6448	0.8732	0.5780	0.8732	0.5989
DT	**0.9339**	**0.6829**	**0.9339**	**0.6392**	**0.9339**	**0.6509**
RF	0.8559	0.6317	0.8559	0.5623	0.8559	0.5838
Textual						
	P-Micro	P-Macro	R-Micro	R-Macro	F-Micro	F-Macro
LR	0.8886	0.7854	0.8886	0.7083	0.8886	0.7398
SVM	0.8938	0.7865	0.8938	0.7139	0.8938	0.7439
DT	**0.9546**	**0.8164**	**0.9546**	**0.7640**	**0.9546**	**0.7855**
RF	0.8640	0.7665	0.8640	0.6934	0.8640	0.7233

*the best results are bolded, the shaded results can be among the top 10 results reported in [2].

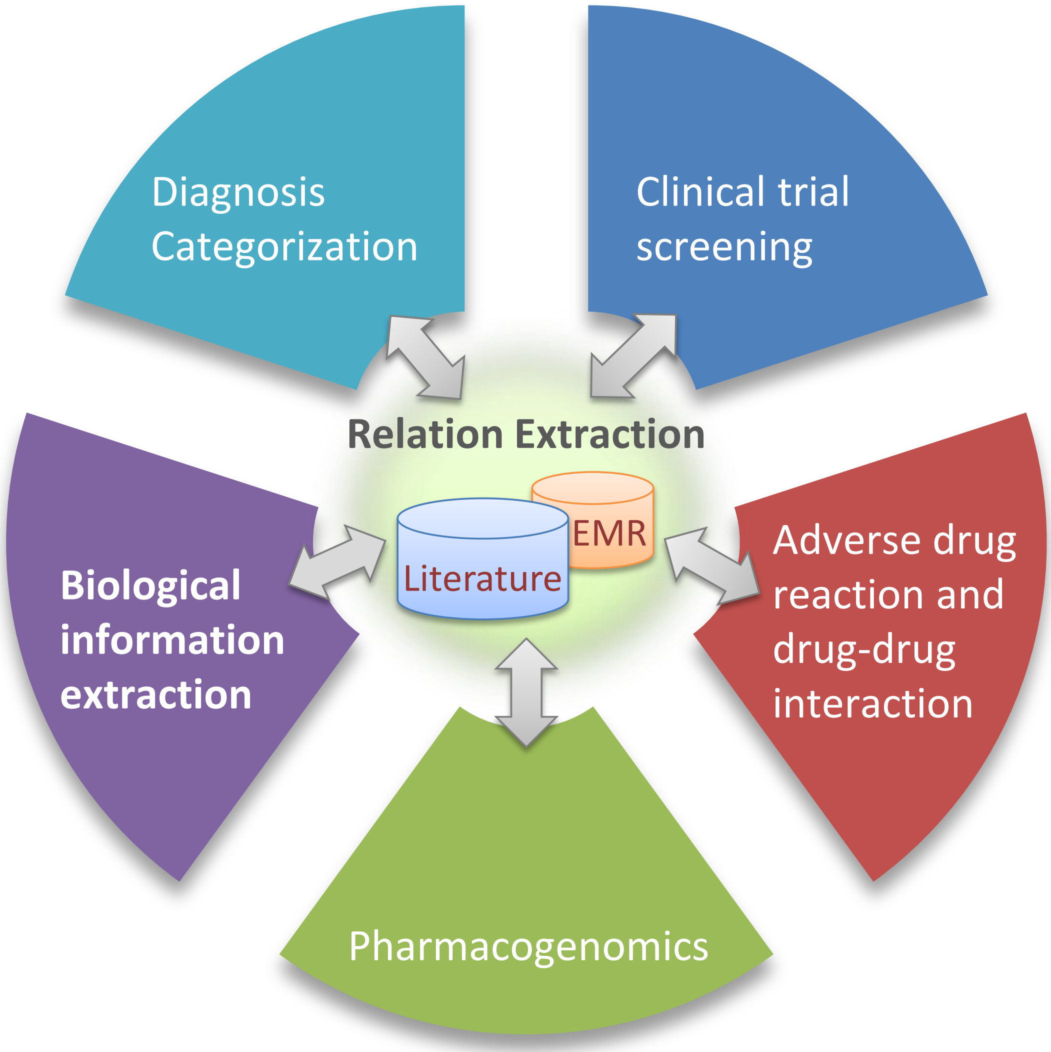
Experimental Results

The classification results for major classes on 15 types of selected CUIs corresponding to the records without family history

Intuitive						
	P-Micro	P-Macro	R-Micro	R-Macro	F-Micro	F-Macro
LR	0.9001	0.6695	0.9001	0.5979	0.9001	0.6206
SVM	0.9074	0.6725	0.9074	0.6065	0.9074	0.6274
DT	**0.9285**	**0.6783**	**0.9285**	**0.6355**	**0.9285**	**0.6467**
RF	0.8690	0.6417	0.8690	0.5740	0.8690	0.5952
Textual						
	P-Micro	P-Macro	R-Micro	R-Macro	F-Micro	F-Macro
LR	0.9188	0.8037	0.9188	0.7303	0.9188	0.7608
SVM	0.9273	0.8060	0.9273	0.7388	0.9273	0.7669
DT	**0.9538**	**0.8160**	**0.9538**	**0.7633**	**0.9538**	**0.7849**
RF	0.8864	0.7823	0.8864	0.7081	0.8864	0.7386

*the best results are bolded, the shaded results can be among the top 10 results reported in Uzuner et al. The 15 types of selected CUIs are considered most related to clinical tasks in Weng et al.

Why Graph Representation of Narrative Sentences?



Y Luo, Ö Uzuner, P Szolovits. Bridging Semantics and Syntax with Graph Algorithms - State-of-the-Art of Extracting Biomedical Relations. *Briefings in Bioinformatics 2016 18 (1), 160-178. PMCID: 5221425*

Graph Representation of Narrative Sentences

- "Immunostains show the large atypical cells are strongly positive for CD30 and negative for CD15, CD20, BOB1, OCT2 and CD3."

Graph Representation of Narrative Sentences

- "Immunostains show the large atypical cells are strongly positive for CD30 and negative for CD15, CD20, BOB1, OCT2 and CD3."
- The sentence tells relationships among procedures, cells, and immunologic factors

Graph Representation of Narrative Sentences

- “Immunostains show the large atypical cells are strongly positive for CD30 and negative for CD15, CD20, BOB1, OCT2 and CD3.”
- The sentence tells relationships among procedures, cells, and immunologic factors
- Feature choices
  - Words
  - UMLS (Unified Medical Language System) concepts, e.g. LCA and CD45

Graph Representation of Narrative Sentences

- "Immunostains show the large atypical cells are strongly positive for CD30 and negative for CD15, CD20, BOB1, OCT2 and CD3."
- The sentence tells relationships among procedures, cells, and immunologic factors
- Feature choices
  - Words
  - UMLS (Unified Medical Language System) concepts, e.g. LCA and CD45
- Can we do better? Relations?

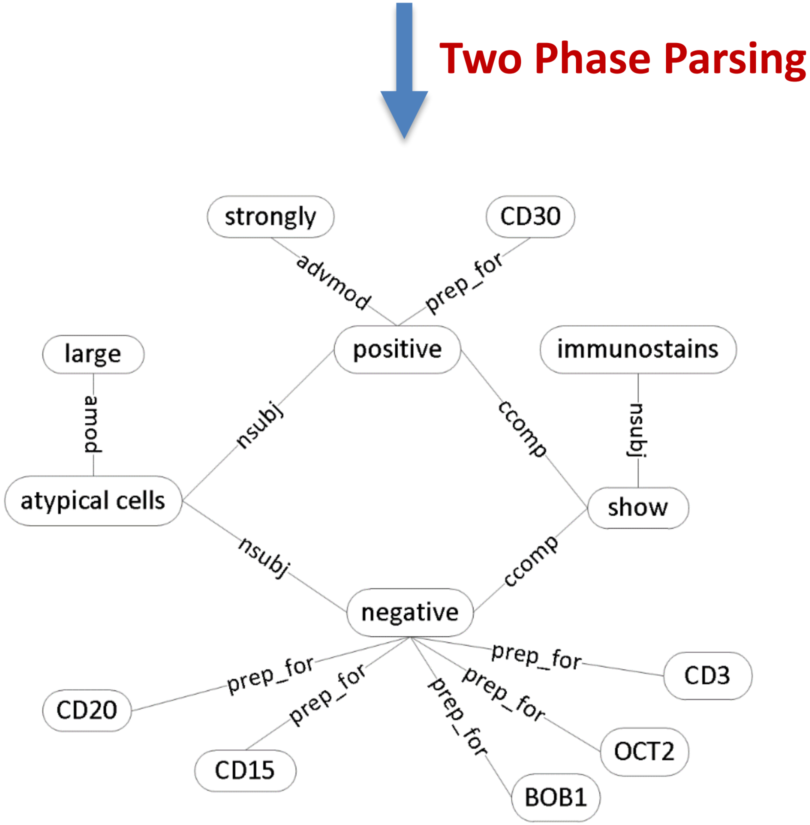
Graph Representation of Narrative Sentences

- "Immunostains show the large atypical cells are strongly positive for CD30 and negative for CD15, CD20, BOB1, OCT2 and CD3."
- The sentence tells relationships among procedures, cells, and immunologic factors
- Feature choices
  - Words
  - UMLS (Unified Medical Language System) concepts, e.g. LCA and CD45
- Can we do better? Relations?

Graph representation is the universal language for modeling relationships among flexible number of concepts

Graph Representation of Narrative Sentences

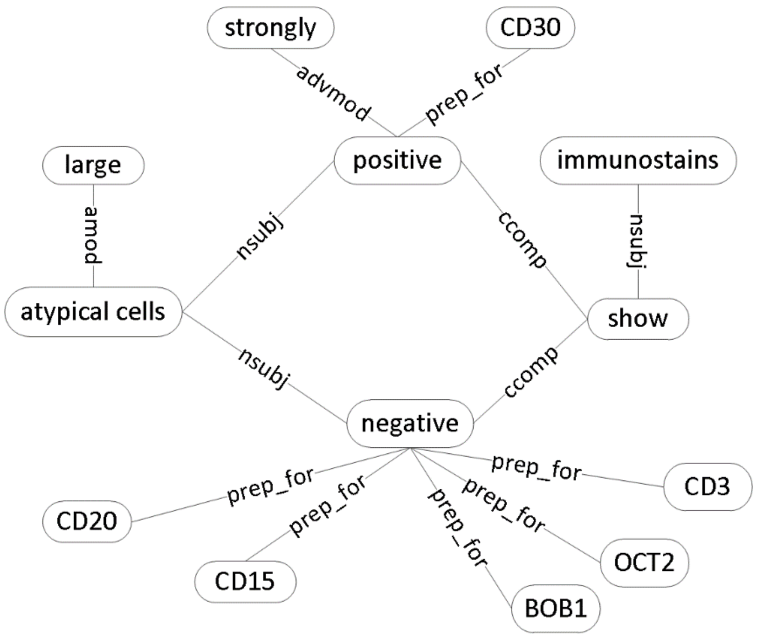
- "Immunostains show the large atypical cells are strongly positive for CD30 and negative for CD15, CD20, BOB1, OCT2 and CD3."



Graph Representation of Narrative Sentences

- "Immunostains show the large atypical cells are strongly positive for CD30 and negative for CD15, CD20, BOB1, OCT2 and CD3."

Two Phase Parsing

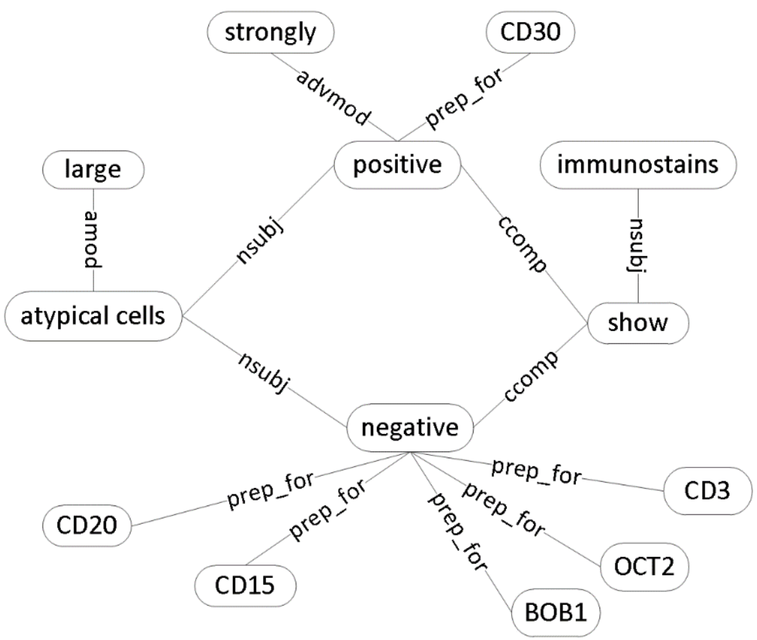


Important relations are likely to be repeated in pathology daily practice:
large atypical cells are positive for CD30 ⇒ sign of Hodgkin lymphoma etc. ⇒ frequently ordered test

Graph Representation of Narrative Sentences

- "Immunostains show the large atypical cells are strongly positive for CD30 and negative for CD15, CD20, BOB1, OCT2 and CD3."

Two Phase Parsing



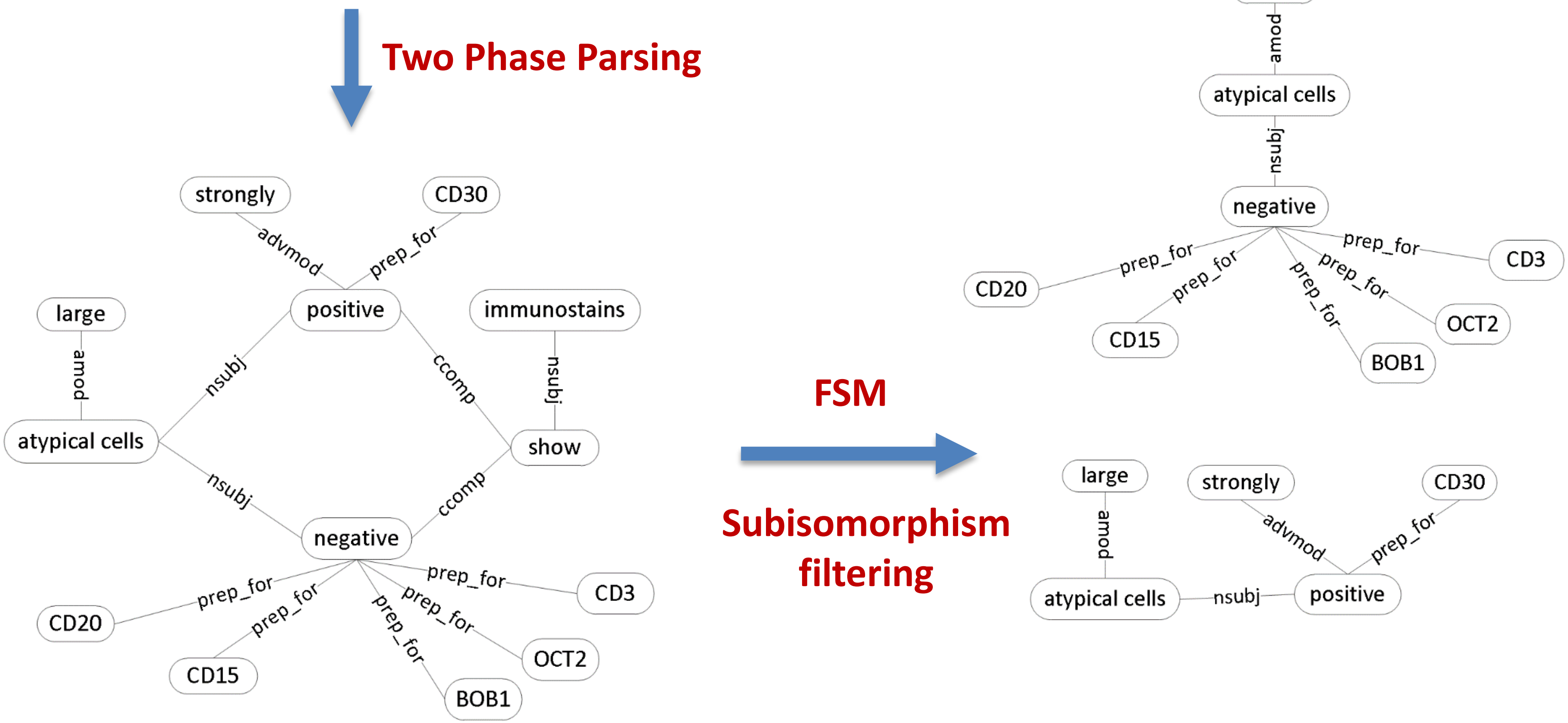
FSM

Subisomorphism filtering

FSM: frequent subgraph mining

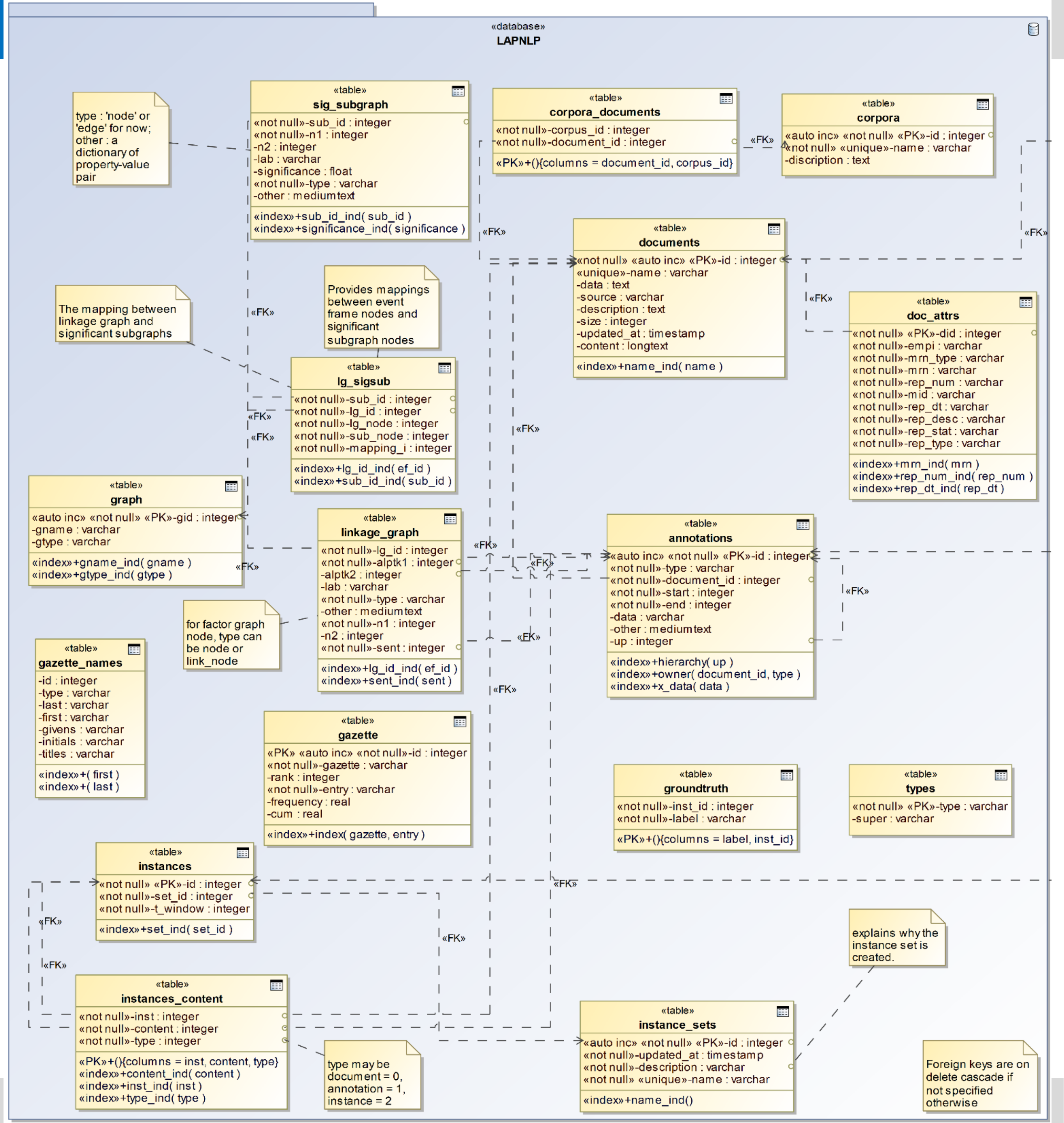
Graph Representation of Narrative Sentences

- "Immunostains show the large atypical cells are strongly positive for CD30 and negative for CD15, CD20, BOB1, OCT2 and CD3."



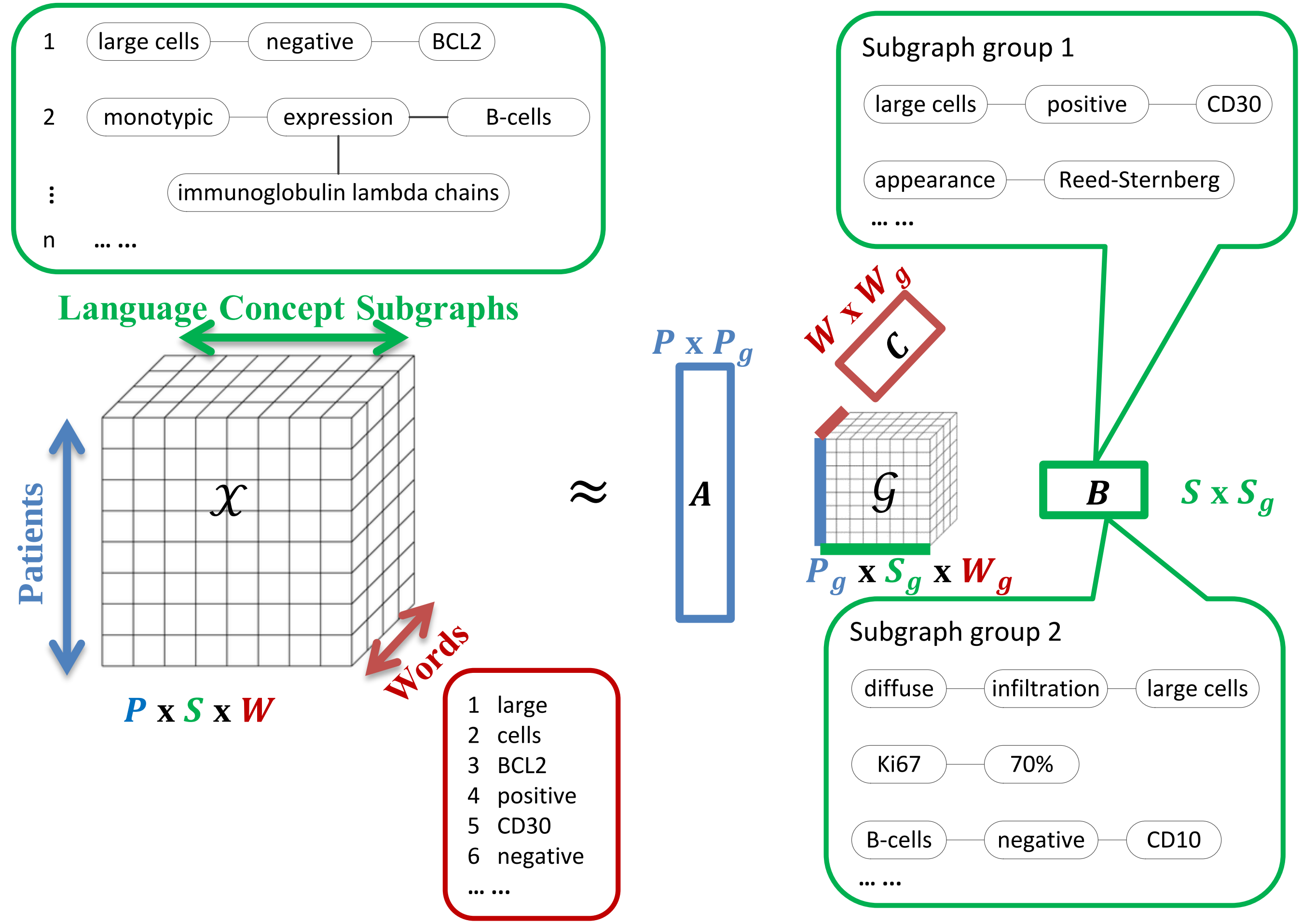
FSM: frequent subgraph mining

Persistent Storage – an extended Common Data Model (CDM)



https://github.com/yuanluo/lapnlp

Computational Phenotyping of Lymphoma

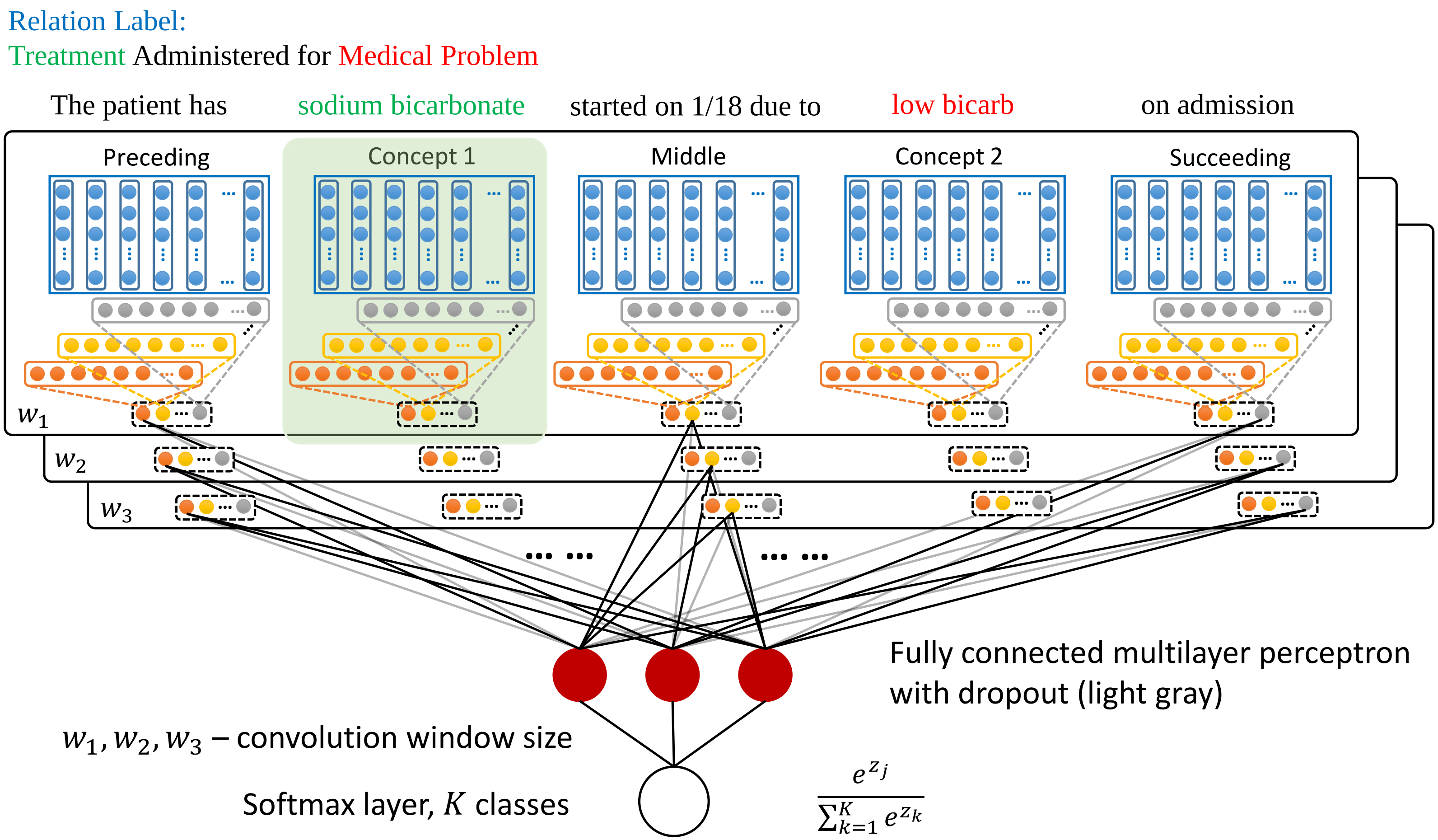


Y Luo, A Sohani, E Hochberg and P Szolovits. Automatic Lymphoma Classification with Sentence Subgraph Mining from Pathology Reports. *JAMIA 2014 21(5):824-832.*

Y Luo, Y Xin, E Hochberg, R Joshi, O Uzuner, P Szolovits. Subgraph Augmented Non-Negative Tensor Factorization (SANTF) for Modeling Clinical Text. *JAMIA 2015 22(5): 1009-1019.*

Semantic Relation Extraction

https://github.com/yuanluo/seg_cnn



Y Luo, Y Cheng, Ö Uzuner, P Szolovits, J Starren. Segment convolutional neural networks (Seg-CNNs) for classifying relations in clinical notes. *JAMIA 2017 Aug 31;25(1):93-8.*

Conclusion

- We develop a portable phenotyping system that is capable of integrating both rule-based and statistical machine learning based phenotyping approaches
- Our system can mine and store both standard UMLS features and the key features of rule-based systems from the unstructured text
- Our system can thus enable the development of new standard UMLS feature based NLP systems as well as the reuse, adaptation and extension of many existing rule-based clinical NLP systems
- We propose extensions to OMOP CDM NOTE and NOTE_NLP tables, especially with enhancement for relation extraction and graph mining

Thank you

- Collaboration welcome
- yuan.luo@northwestern.edu
- @yuanhypnosluo

Ongoing projects

- Mapping of Note Types to LOINC/standard vocabulary – Karthik Natarajan, Ruth Reeves, and Jon Duke
- Landscape Analysis of section identifier systems and proposal of a standard terminology for use – Hua Xu and Karthik Natarajan
- Mapping of CUIs to standard terminology – Juan Banda
- Standardization of term_modifiers and values – Hua Xu
- Evaluate and revise textual CDM tables by sharing practical issues and lessons learnt during ETL for processing textual data into CDM – Ruth Reeves, others?
- Develop tools (within Atlas) to facilitate uses of NLP data for cohort building/phenotyping : Collaborate with eMERGE consortium
- Conduct cross-site studies that use textual data
- Continue developing other NLP resources

Other issues

- Presentation scheduling
  - April 10th – Jon Duke – ClarityNLP
  - May 8th – Juan Banda - CUI mapping, ongoing work – Juan, Stephan Meyestre – tool to evaluate NLP systems
  - June 12th
  - July 10th
- Please let us know if you can present your related work at any of the above meetings.