COMP4901K/Math4824B
Machine Learning for Natural Language Processing

Knowledge Graphs and Its Applications in Healthcare

Instructor: Yangqiu Song

Outline

- Introduction
  - General knowledge graphs
  - Knowledge graphs in healthcare

- Knowledge Graph Construction

What's a Knowledge Graph?

- A knowledge graph has many names in the history
  - Semantic networks, knowledge base, ontology, …

- In 2012, Google released its project "Google Knowledge Graph"
  - A graph-based knowledge representation connecting real-world entities to support search
    - Landmarks, celebrities, cities, sports teams, buildings, geographical features, movies, celestial objects, works of art and more
  - Get information instantly relevant to a query

https://googleblog.blogspot.com/2012/05/introducing-knowledge-graph-things-not.html

anthony fauci

All News Images Videos Books More Settings Tools

About 46,500,000 results (0.52 seconds)

Top stories

Dr. Anthony Fauci undergoes surgery for vocal cord polyp
ABC News · 15 hours ago

White House coronavirus advisor Dr. Anthony Fauci recovering from vocal-cord surgery
CNBC.com · 17 hours ago

More for anthony fauci

www.niaid.nih.gov › about › director

Anthony S. Fauci, M.D., NIAID Director | NIH: National Institute ...

The NIAID budget for fiscal year 2020 is an estimated $5.9 billion. Dr. **Fauci** has advised six Presidents on HIV/AIDS and many other domestic and global health ...
Anthony S. Fauci, MD · Dr. Fauci in the News · Contact Us · Publications and Articles

www.niaid.nih.gov › about › anthony-s-fauci-md-bio

Anthony S. Fauci, M.D. | NIH: National Institute of Allergy and ...

Dr. **Fauci** was appointed director of NIAID in 1984. He oversees an extensive portfolio of basic and applied research to prevent, diagnose, and treat established ...
Dr. Fauci in the News · Profiles, Awards and Honors · Publications

en.wikipedia.org › wiki › Anthony_Fauci

Anthony Fauci - Wikipedia

Anthony Stephen **Fauci** is an American physician and immunologist who has served as the director of the National Institute of Allergy and Infectious Diseases ...

Institutions: National Institutes of Health, Nat... **Children**: 3
Education: College of the Holy Cross (BA); C... **Born**: Anthony Stephen Fauci; December 2...

Christine Grady · Humanism · White House Coronavirus ... · Deborah Birx

www.npr.org › coronavirus-live-updates › 2020/08/20

Anthony Fauci Has Surgery To Remove Polyp From Vocal ...

Anthony Fauci

American physician

Anthony Stephen Fauci is an American physician and immunologist who has served as the director of the National Institute of Allergy and Infectious Diseases since 1984. Wikipedia

Born: December 24, 1940 (age 79 years), Brooklyn, New York, United States

Spouse: Christine Grady (m. 1985)

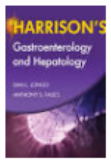
Education: College of the Holy Cross, Weill Cornell Medical College, Regis High School

Awards: Presidential Medal of Freedom, MORE

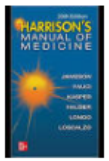
Children: Megan Fauci, Jennifer Fauci, Alison Fauci

Books

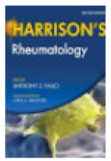
View 20+ more



Harrison's Gastroen... and Hep...
2010



Harrisons Manual of Medicin...
2019



Harrison's Rheumat...
2006

Harrison's Infectious Disease...
2013


Current Therapy in Allergy, I...
1992

People also search for

View 10+ more


Christine Grady


Deborah Birx


Donald Trump


Judy Mikovits


Joe Biden

COVID-19

All News Videos Images Shopping More Settings Tools

About 5,790,000,000 results (0.76 seconds)

Top stories

Hong Kong Free Press

Covid-19: New infections hit six-week low in Hong Kong, more public services to resume

4 hours ago

Mondaq News Alerts

Meeting Lockdown Challenges With Communication, Patience And Resolve...

17 hours ago

South China Morning Post

Expert warns virus transmission rate on rise as Hong Kong extends rules

4 days ago

Health information

Symptoms | Prevention | Treatments

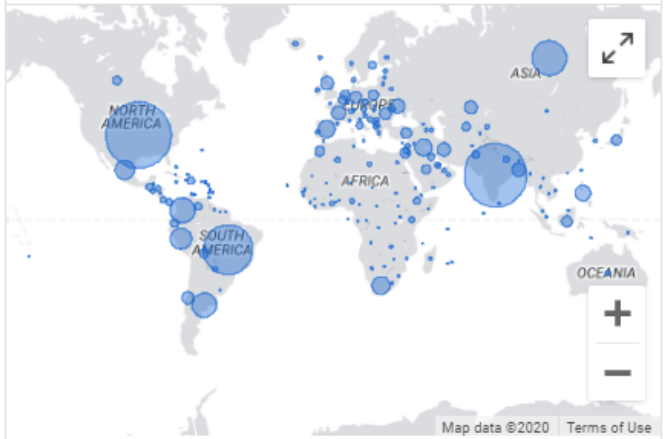
COVID-19 affects different people in different ways. Most infected people will develop mild to moderate illness and recover without hospitalization.

Most common symptoms:
- fever
- dry cough
- tiredness

Less common symptoms:
- aches and pains
- sore throat
- diarrhoea
- conjunctivitis
- headache
- loss of taste or smell
- a rash on skin, or discolouration of fingers or toes

Learn more on who.int

Map of cases (last 14 days)



Sources: Wikipedia and The New York Times. · About this data

Cases overview

Worldwide

Total cases	Recovered	Deaths
22.6M	14.5M	792K

More locations and statistics

'+' shows new cases reported yesterday · Updated less than 4 hours ago · Source: Wikipedia · About this data

Coronavirus disease (COVID-19) is an infectious disease caused by a newly discovered coronavirus.

Most people who fall sick with COVID-19 will experience mild to

State-of-the-art Enterprise-level KGs

	Data Model	Size of Nodes	Size of Edges	Development Stage
Google	Strongly typed entities, relations with domain and range inference	~1 Billion	~70 Billions	Actively used in products
Microsoft	The types of entities, relations, and attributes in the graph are defined in an ontology	~2 Billions	~55 Billions	Actively used in products
Facebook	All of the attributes and relations are structured and strongly typed, and optionally indexed to enable efficient retrieval, search, and traversal.	~50 Millions	~500 Millions	Actively used in products
eBay	Entities and relation, well-structured and strongly typed	~100 Million	~1 Billion	Early stages of development and deployment
IBM	Entities and relations with evidence information associated with them	~100 Millions	~5 Billions	Actively used in products and by clients

Natalya Fridman Noy, Yuqing Gao, Anshu Jain, Anant Narayanan, Alan Patterson, Jamie Taylor: Industry-scale knowledge graphs: lessons and challenges. Commun. ACM 62(8): 36-43 (2019)

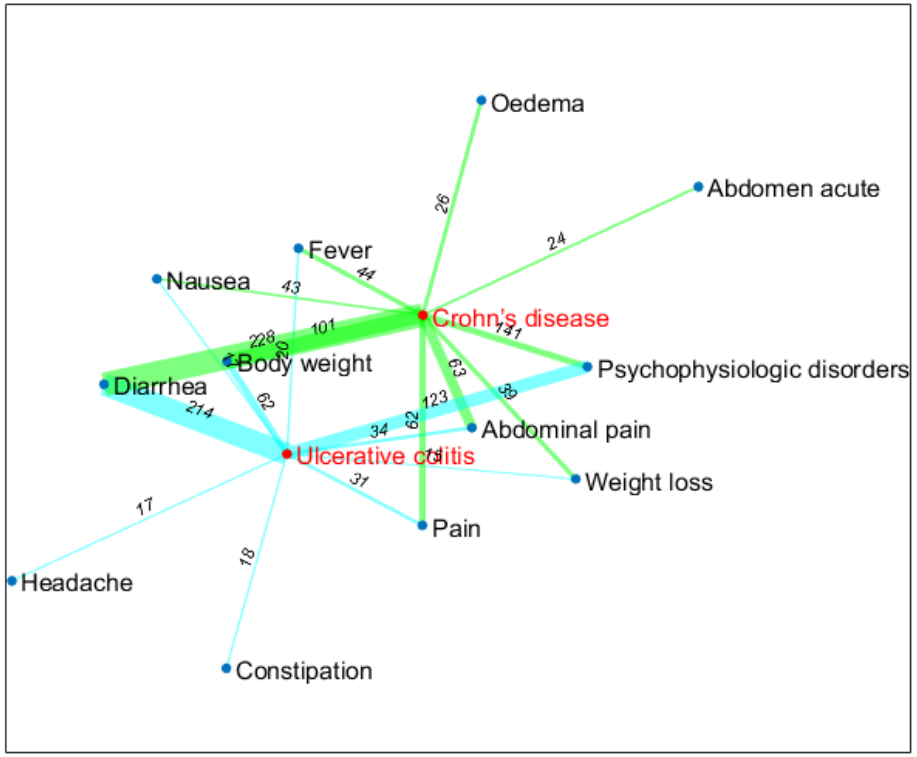
Google Knowledge Graph in Healthcare

- Knowledge graph is useful for clinical decision support systems and self-diagnostic symptom checkers
  - For example, major symptoms of a heart attack are pain or discomfort chest, arms or shoulder, jaw, neck, or back, feeling weak, lightheaded or faint and shortness of breath
- Answers to common medical questions for searches related to health conditions based on
  - Aggregated information from search results
  - Multi-faceted medical facts
    - Google worked with a “team of doctors” to “carefully compile, curate, and review this information”
    - “All of the gathered facts represent real-life clinical knowledge from these doctors and high-quality medical sources across the web, and the information has been checked by medical doctors at Google and the Mayo Clinic for accuracy.”

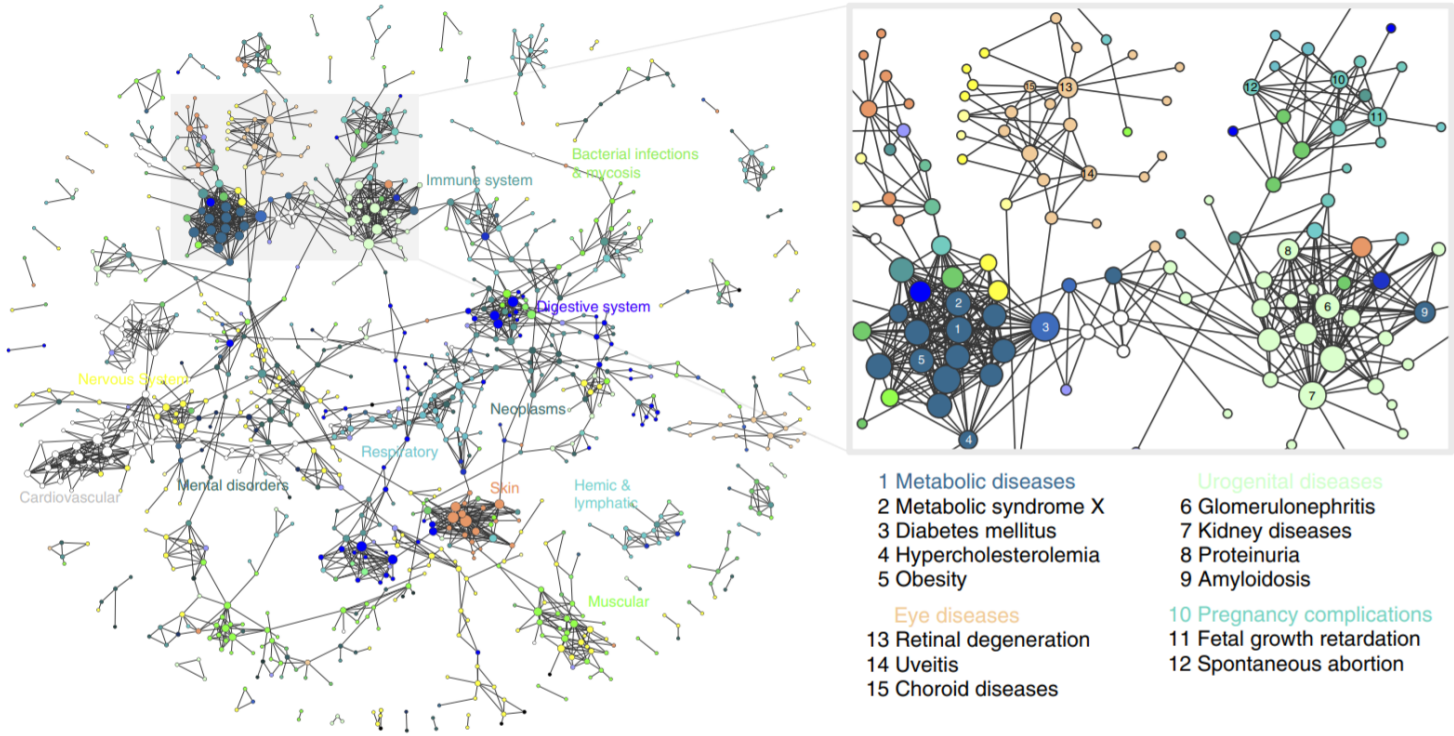
https://techcrunch.com/2015/03/11/google-knowledge-graph-meets-healthcare/
Figure from: https://googleblog.blogspot.com/2015/02/health-info-knowledge-graph.html

Example of Knowledge Graphs in Healthcare

Symptom and disease linking from biomedical literature database: PubMed



A symptom-disease network extracted from PubMed

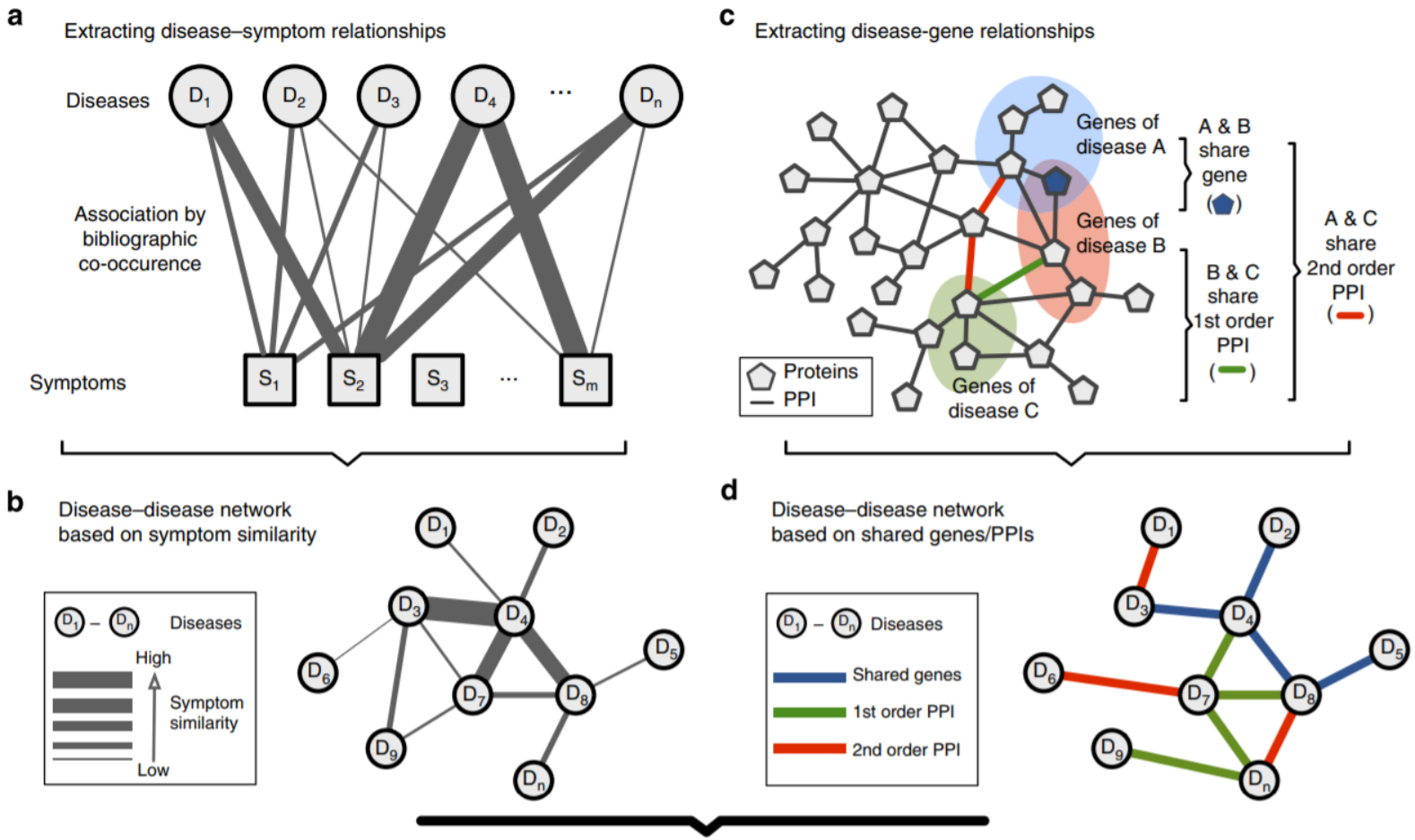


A disease network shows similarities based on symptoms reveals disease clusters

Data and figure from: XueZhong Zhou, Jörg Menche, Albert-László Barabási, Amitabh Sharma. Human symptoms–disease network. Nature Communications. 2014

Example of Knowledge Graphs in Healthcare

- The disease-symptom network can be further integrated to analyze disease similarities
  - A disease network where link weight between two diseases quantifies the similarity of their respective symptoms
  - A disease network where disease–gene association and protein–protein interaction (PPI) are used



- Shared symptoms indicate shared genes between diseases
- Shared symptoms indicate shared protein interactions.

Figure from: XueZhong Zhou, Jörg Menche, Albert-László Barabási, Amitabh Sharma. Human symptoms–disease network. Nature Communications. 2014

Example of Knowledge Graphs in Healthcare

Symptom and disease linking from large-scale EHR data

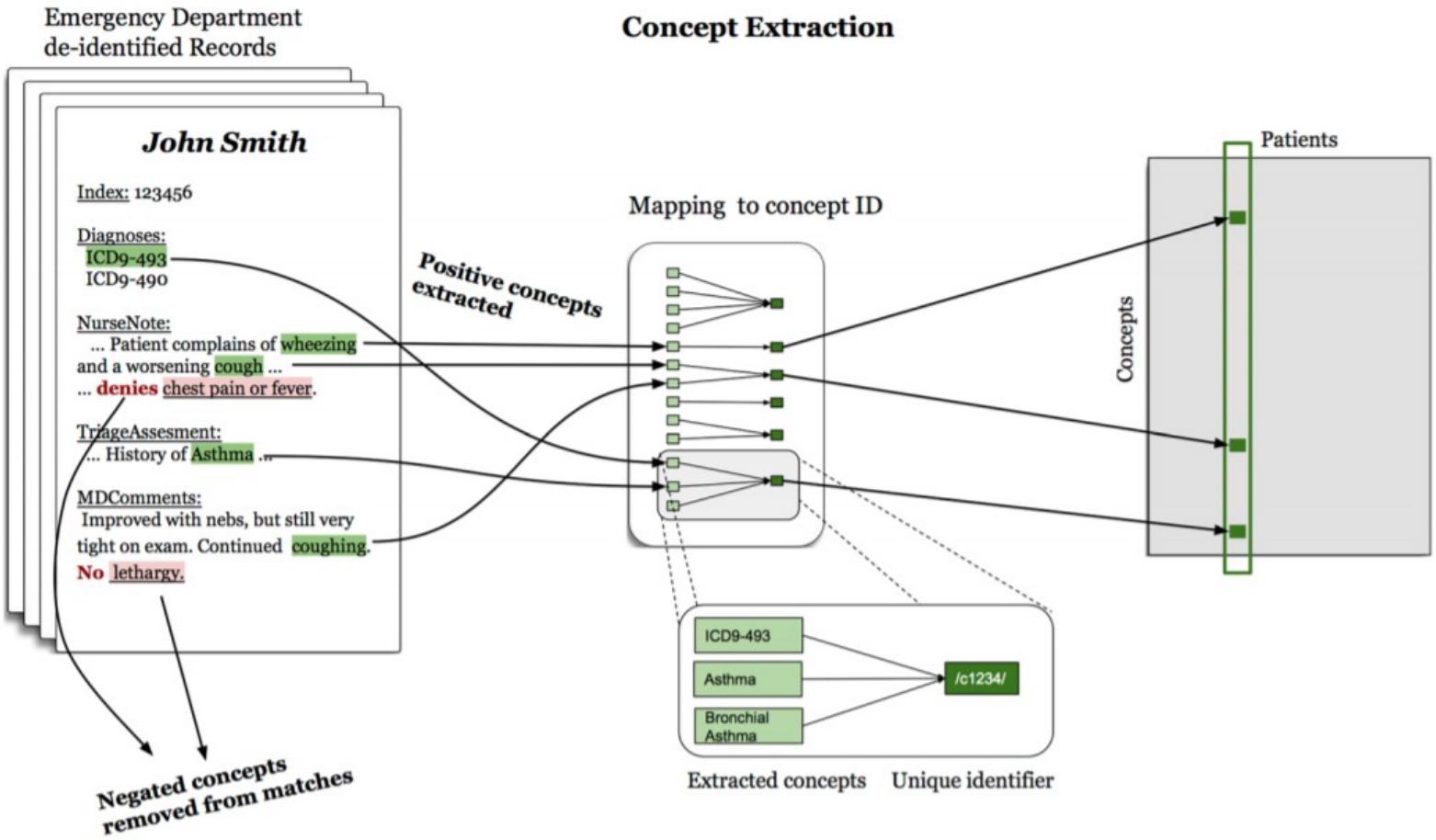
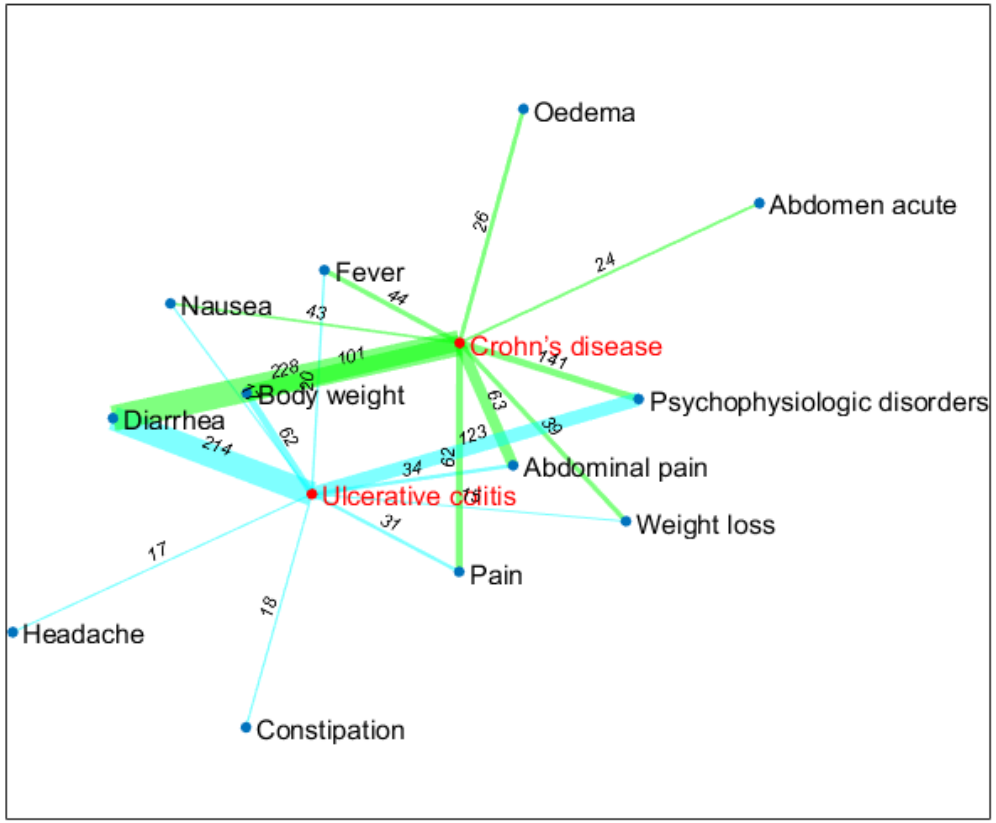


Figure from: Maya Rotmensch, Yoni Halpern, Abdulhakim Tlimat, Steven Horng & David Sontag. Learning a Health Knowledge Graph from Electronic Medical Records. Scientific Reports. 2017

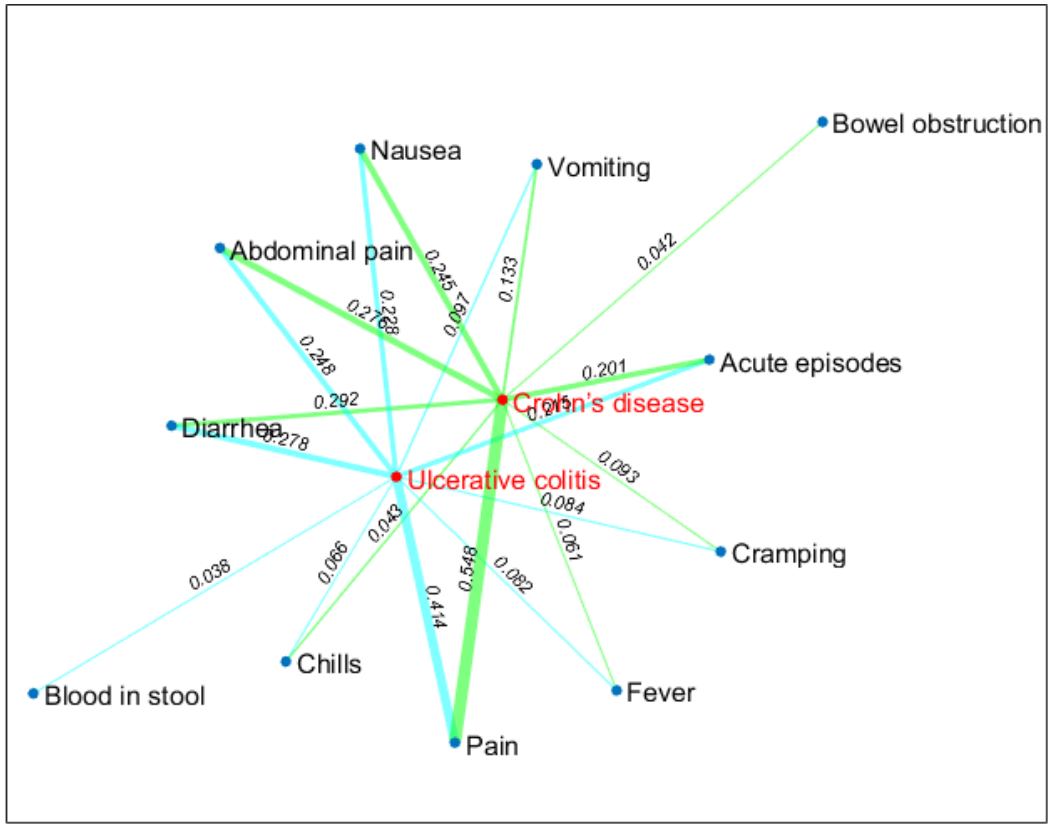
Comparison Between PubMed and EHR

PubMed: more declarative, more formal language and regular texts, more typical symptoms to promote learning



Data from: XueZhong Zhou, Jörg Menche, Albert-László Barabási, Amitabh Sharma. Human symptoms–disease network. Nature Communications. 2014

EHR: more statistical, noisier text but recording the practical medicine use



Data from: Maya Rotmensch, Yoni Halpern, Abdulhakim Tlimat, Steven Horng & David Sontag. Learning a Health Knowledge Graph from Electronic Medical Records. Scientific Reports. 2017

Quality of Such Kind of Knowledge Graphs

- Precision-recall curve rated according to physicians' expert opinion
- Google Health Knowledge Graph has two tags "always" and "frequent"

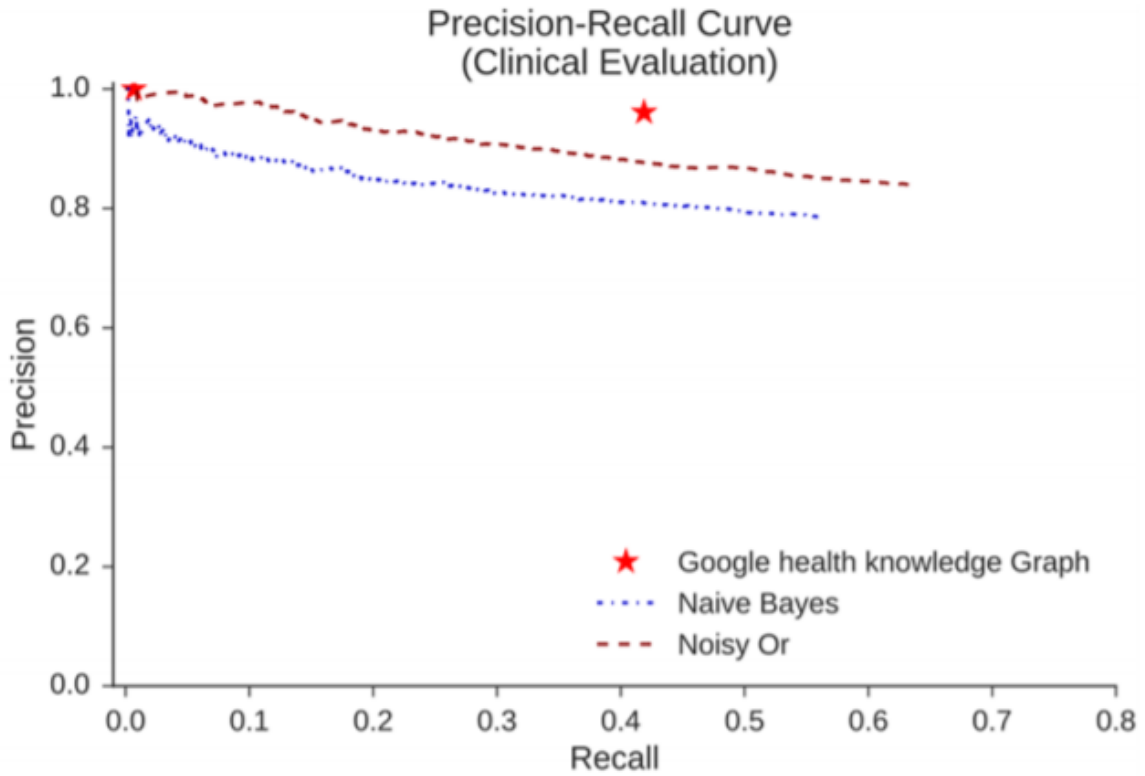


Figure from: Maya Rotmensch, Yoni Halpern, Abdulhakim Tlimat, Steven Horng & David Sontag. Learning a Health Knowledge Graph from Electronic Medical Records. Scientific Reports. 2017

Information Extraction and Text Mining

- Extracting information from PubMed/HER is useful, however, there is a lot of variability and ambiguity of language and terminology use. For example,
  - Amyotrophic lateral sclerosis, motor neurone disease, and Lou Gehrig's Disease refer to the same disease
  - According to Medical Subject Headings (MeSH), obesity belongs to
    - Nutritional and Metabolic Diseases
    - Diagnosis
    - Physiological Phenomena
    - Pathological Conditions, Signs and Symptoms

- There are many (~200) existing expert annotated knowledge bases in Unified Medical Language System (UMLS)
  - 127 semantic types organized as a hierarchy
  - 3.2 million unique concepts
    - A primary name and a set of aliases
    - 8% are linked to more than one types

Olivier Bodenreider. The Unified Medical Language System (UMLS): integrating biomedical terminology. 32 (Database issue):D267–D270, 2004.

Why Knowledge Graphs in Healthcare?

- Medical doctors are overwhelmed by the huge amount of data, e.g.,
  - Electronic Health Records (EHRs)
  - Research articles from PubMed
  - Providing complementary information to existing knowledge bases

- We need an efficient tool to “connect the dots”
  - Humans can only process a few objects/variables at a time
  - Human’s summarization of concepts can be vague
  - Types of concepts are heterogeneous, e.g., patient, disease, symptom, gene, chemical, etc.
  - Demands to reason multi-hop relations
  - Traditional logic inferences tools may not be able to handle
    - Large amount of heterogeneous knowledge sources
    - Ambiguity of entities and relations
    - Processes (activities), states, events, and their relations etc.

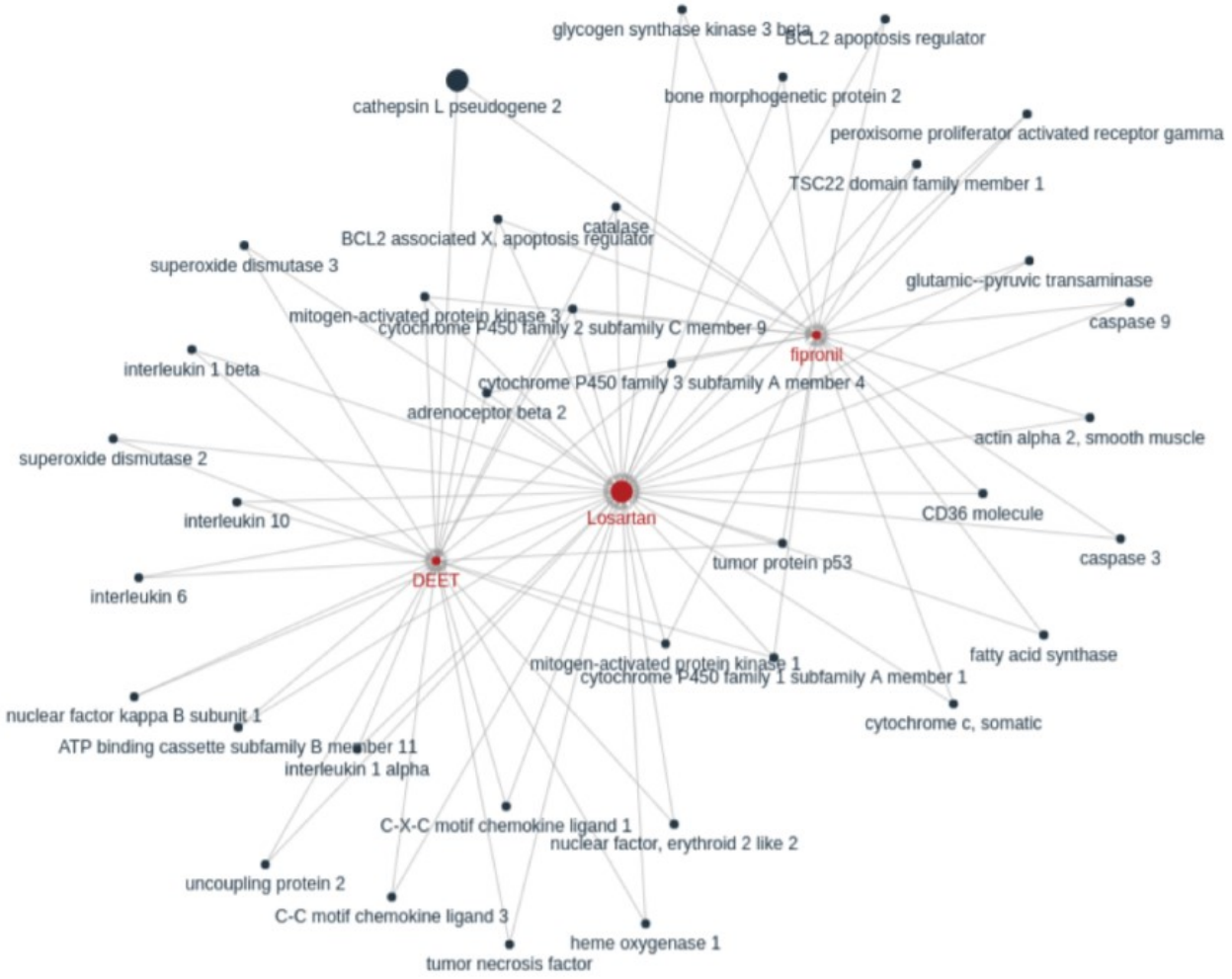
Learning a Health Knowledge Graph from Electronic Medical Records. Rotmensch, M.; Halpern, Y.; Tlimat, A.; Horng, S.; and Sontag, D. Nature Scientific Reports, 7(1): 5994. 2017.

UIUC COVID-19 Literature Knowledge Graph

- http://blender.cs.illinois.edu/covid19/

- Extract entities, relations and events from tex
  - 50,752 Gene nodes
  - 10,781 Disease nodes
  - 5,738 Chemical nodes
  - 535 Organism nodes
  - 133 relation types
  - 13 Event types

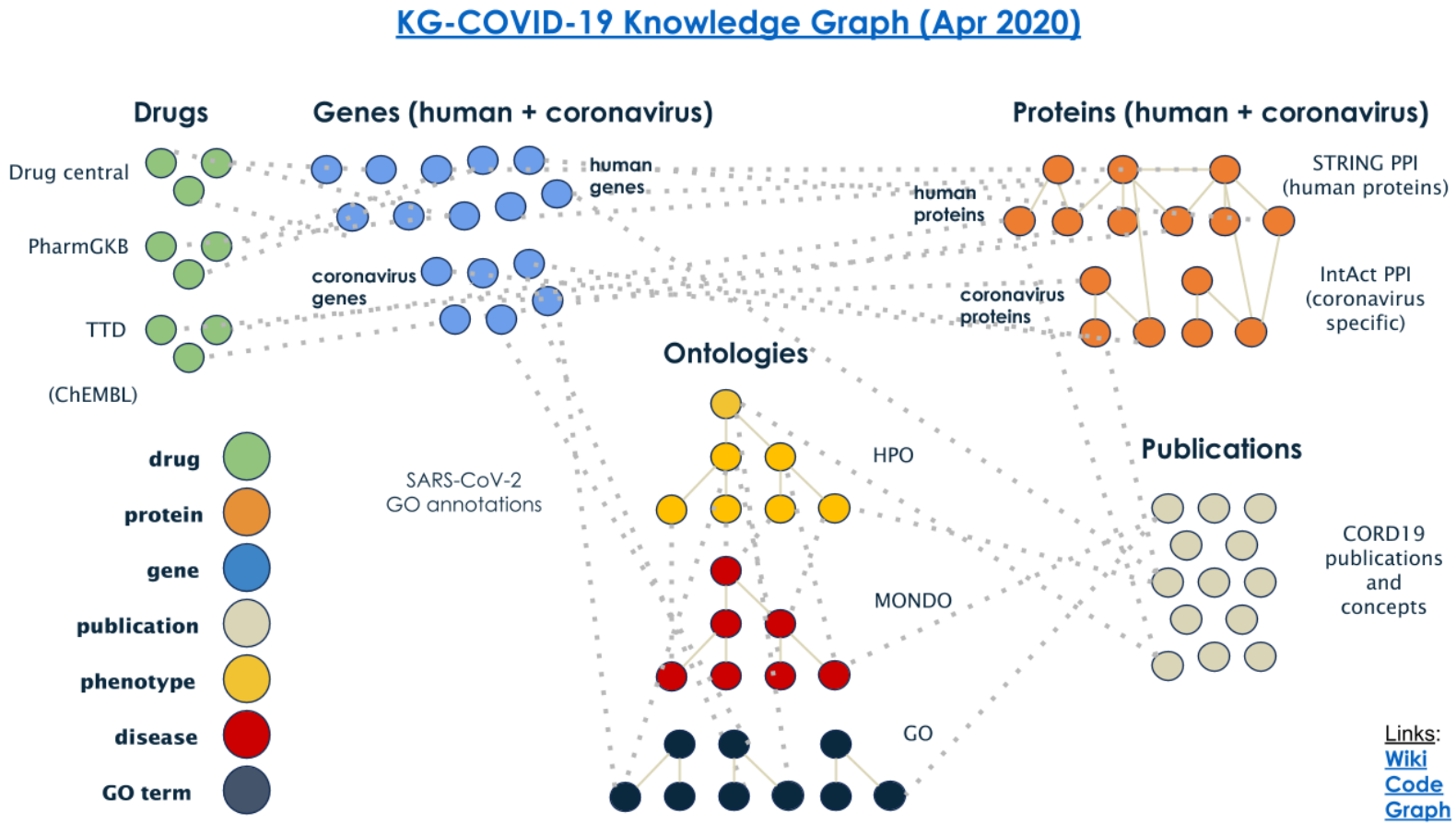
- Knowledge extraction from images, and do cross-media fusion and inference with entitie and events



Q. Wang et al., COVID-19 Literature Knowledge Graph Construction and Drug Repurposing Report Generation. Arxiv, 2020

Berkeley Lab COVID-19 Knowledge Graph

32,000 drugs, 21,000 human and 272 viral proteins plus roughly the same number of genes, and more than 50,000 scientific studies and clinical trials.

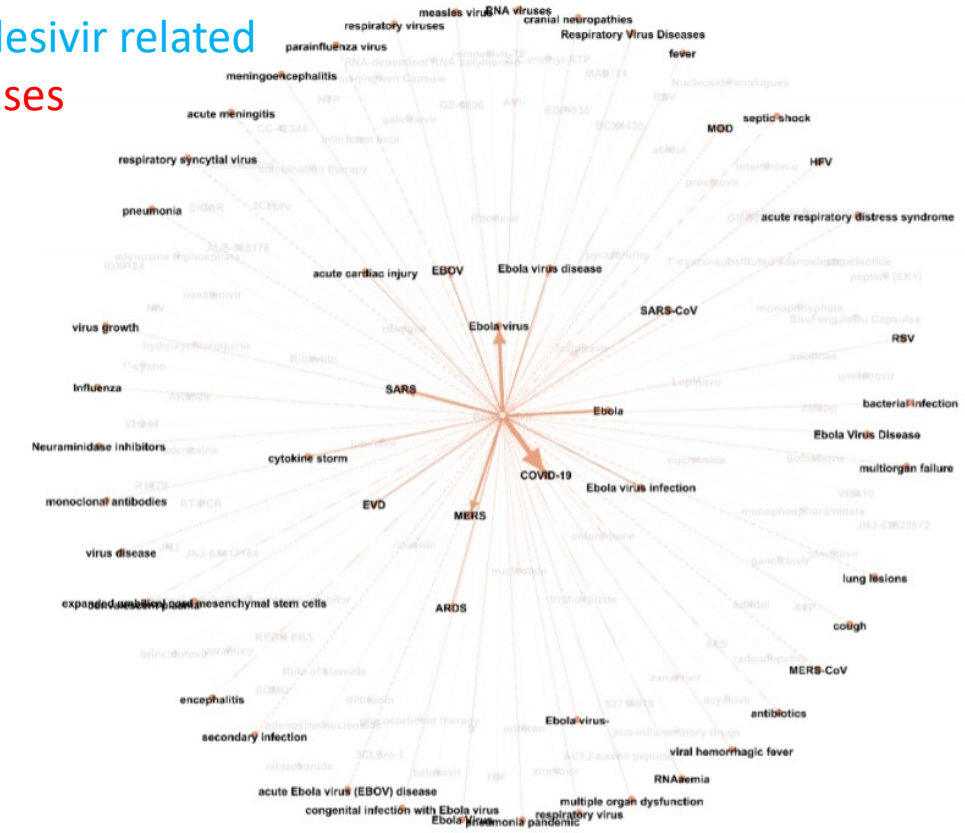


https://federallabs.org/news/berkeley-lab-creates-knowledge-graph-to-make-covid-19-drug-predictions
Image from: https://github.com/Knowledge-Graph-Hub/kg-covid-19/wiki

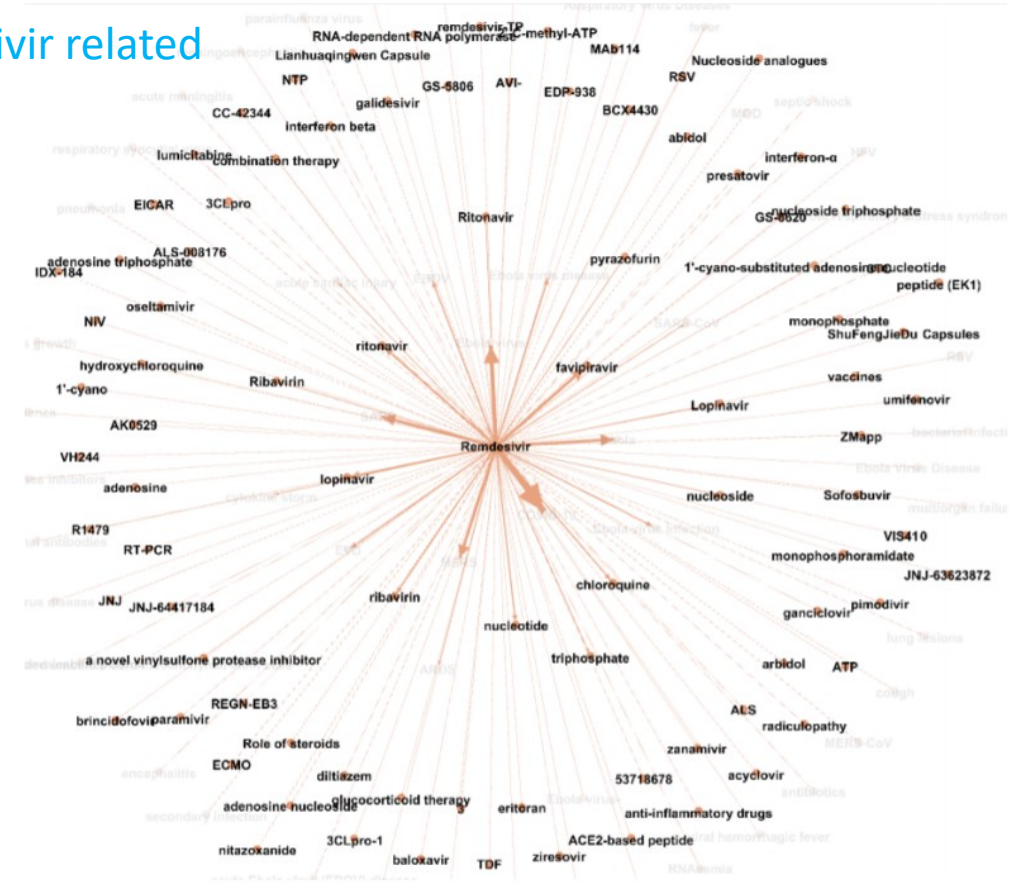
UT Austin COVID-19 Knowledge Graph

- 53,523 Drugs, 12,077 Diseases, 15,519 Species, 18,678 Genes, Gene mutations extracted from CORD-19 dataset

remdesivir related diseases



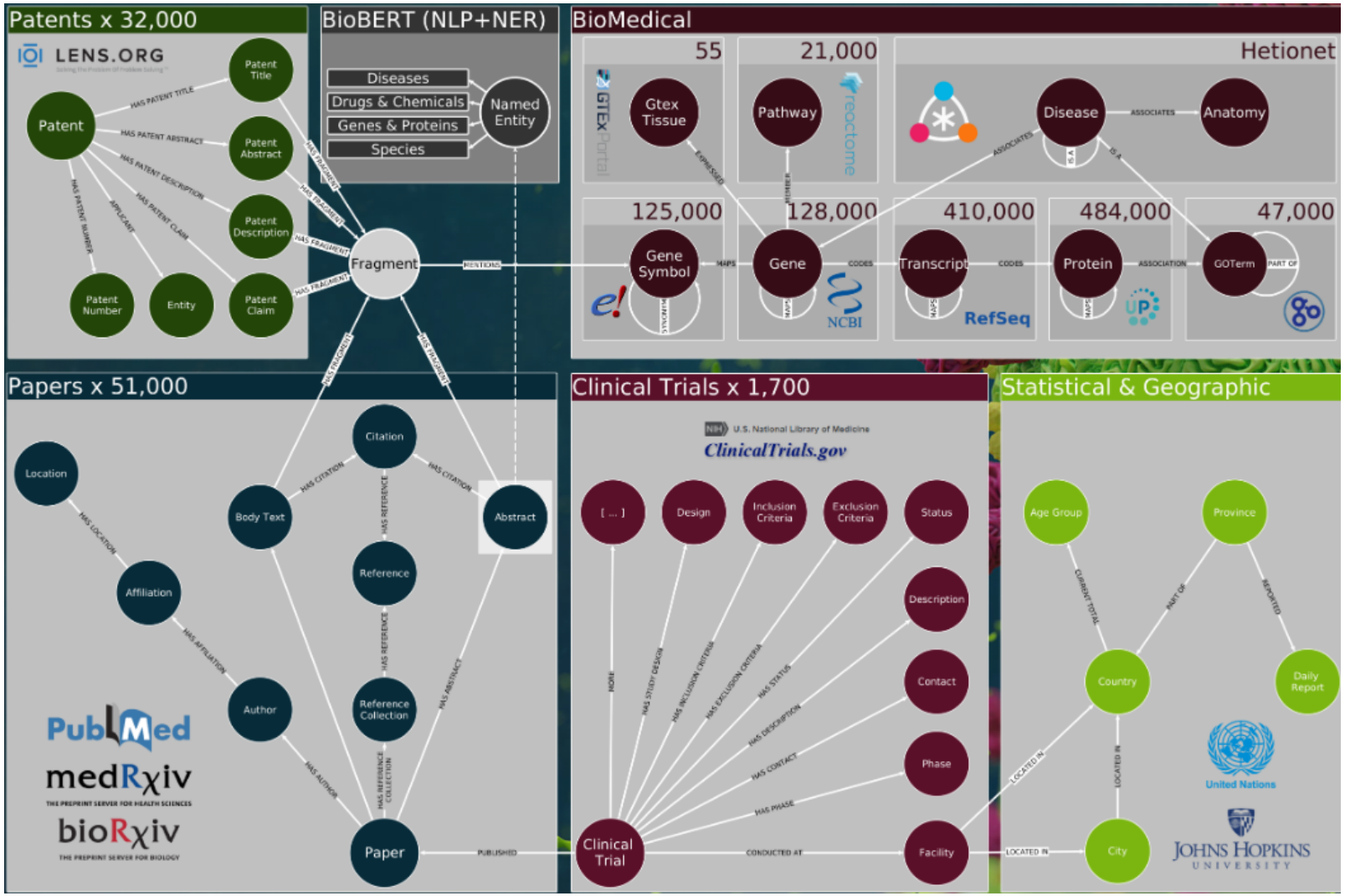
remdesivir related drugs



Chen, C., Ebeid, I.A., Bu, Y., & Ding, Y. (2020). Coronavirus knowledge graph: A case study. KDD Workshop on Knowledge Graph, 2020.
https://www.semanticscholar.org/cord19

COVIDGraph: https://covidgraph.org/

- "CovidGraph is a non-profit collaboration of researchers, software developers, data scientists and medical professionals."

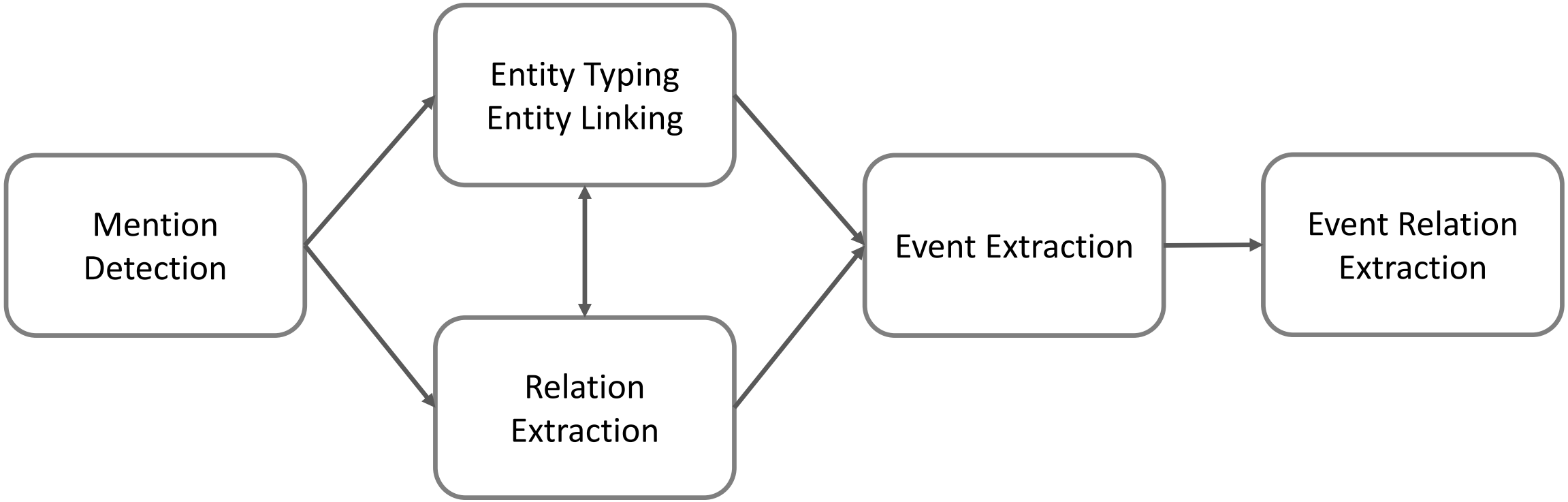


Outline

- Introduction

- Knowledge Graph Construction
  - Entities, typing and linking
  - Entity Relations
  - Events
  - Event relations

General Procedure of Information Extraction for Knowledge Graph Construction



Information Extraction

- Information extraction can be pattern based or learning based

- Pattern based detection
  - Reflecting human's knowledge of language
  - Perform well when there are less annotated data
  - Especially good for open information extraction (Open-IE)
    - No pre-defined domains and relation (predicate) types among mentions

- Learning based detection
  - Better performance when there are many annotated training examples

Pattern Based Information Extraction

- Patterns can be extracted from dependency parsing results (Wang et al., 2018)

Pattern	Type	Example	Derived Clauses
		Basic Patterns	
SV_i	SV	Colectomy works.	(Colectomy, works)
SV_eA	SVA	Pulmonary toxicity is in lungs.	(Pulmonary toxicity, is, in lungs)
SV_cC	SVC	Pulmonary toxicity is vital.	(Pulmonary toxicity, is, vital)
$SV_{mt}O$	SVO	Nitrofurantoin causes pulmonary toxicity.	(Nitrofurantoin, causes, pulmonary toxicity)
$SV_{dt}O_iO$	SVOO	Colectomy gives the patient a chance.	(Colectomy, gives, the patient, a chance)
$SV_{ct}OA$	SVOA	Colectomy removes the colon away.	(Colectomy, removes, the colon, away)
$SV_{ct}OC$	SVOC	Pulmonary toxicity causes rats to die.	(Pulmonary toxicity, causes, rats, to die)

  - Extended pattern: Pulmonary toxicity often appears in lungs in rats. (SVAA)

- But as there is no restricted relations/predicates, relations should be disambiguated

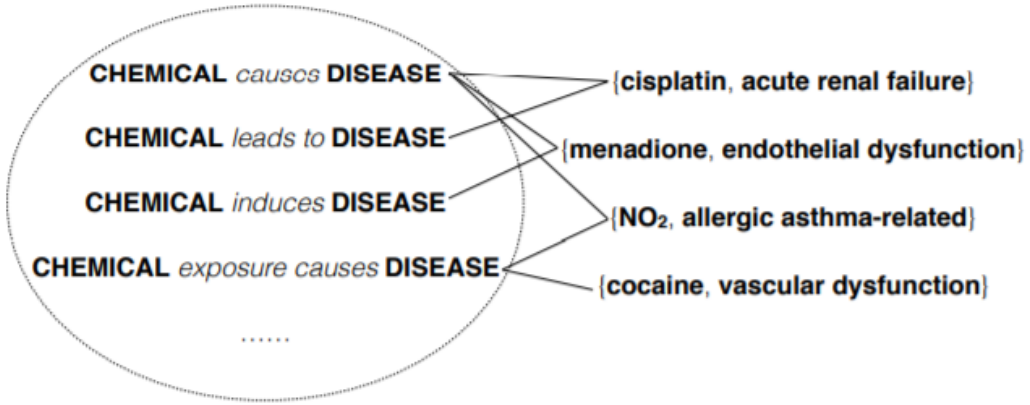


Table and Figure from: Xuan Wang, Yu Zhang, Qi Li, Yinyin Chen, Jiawei Han: Open Information Extraction with Meta-pattern Discovery in Biomedical Literature. BCB 2018

Learning Based Information Extraction

- New trend: pre-training and fine-tuning
- Pre-training with large amount of data, e.g.,

Model	Data	Method
BioBERT	PubMed abstracts/full texts	continual pre-training from BERT
Clinical-BERT	MINIC clinical notes	continual pre-training from BERT
SciBERT	Scientific papers from Semantic Scholar	from scratch
BlueBERT:	PubMed abstracts + MIMIC clinical notes	continual pre-training from BERT
PubMedBERT	PubMed abstracts/full texts	from scratch

- Fine-tuning: tasks specific datasets, e.g.,
  - Named entity recognition
  - Relation extraction

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, Jaewoo Kang: BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 2020
Emily Alsentzer, John R. Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, Matthew B. A. McDermott: Publicly Available Clinical BERT Embeddings. CoRR abs/1904.03323 (2019)
Iz Beltagy, Kyle Lo, Arman Cohan: SciBERT: A Pretrained Language Model for Scientific Text. EMNLP/IJCNLP (1) 2019
Yifan Peng, Qingyu Chen, Zhiyong Lu: An Empirical Study of Multi-Task Learning on BERT for Biomedical Text Mining. BioNLP 2020: 205-214
Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, Hoifung Poon: Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. CoRR abs/2007.15779 (2020)

BioBERT

- Shows 0.62% Average F1 score improvement on 9 biomedical named entity recognition datasets and 2.80% Average F1 score improvement on 3 biomedical relation extraction datasets

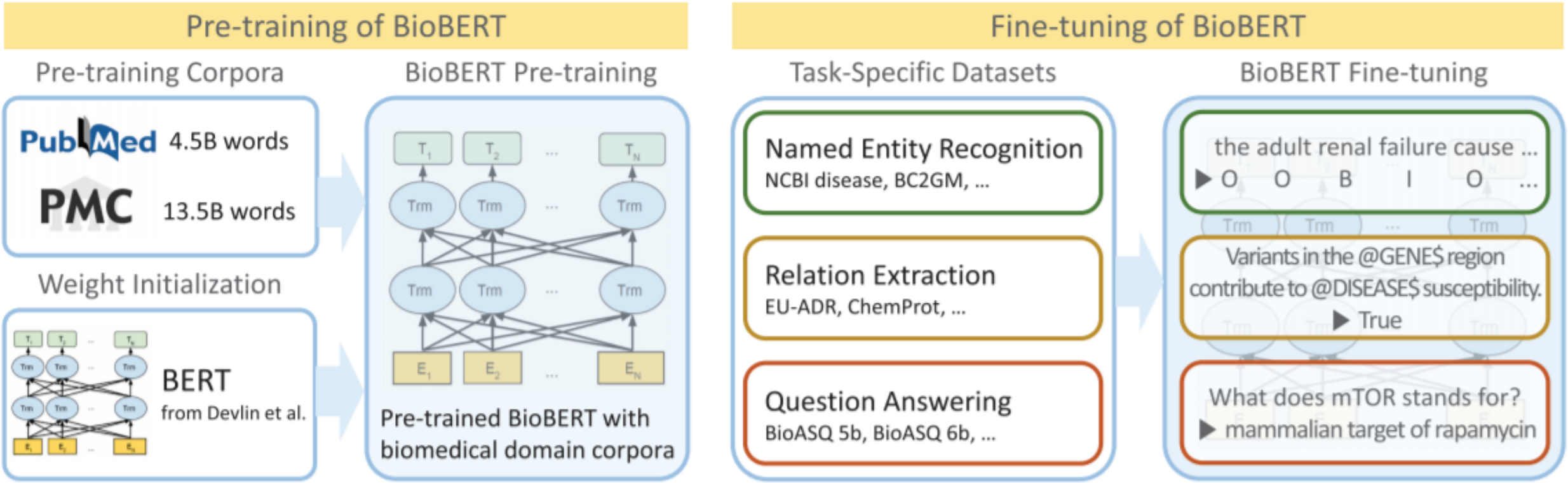
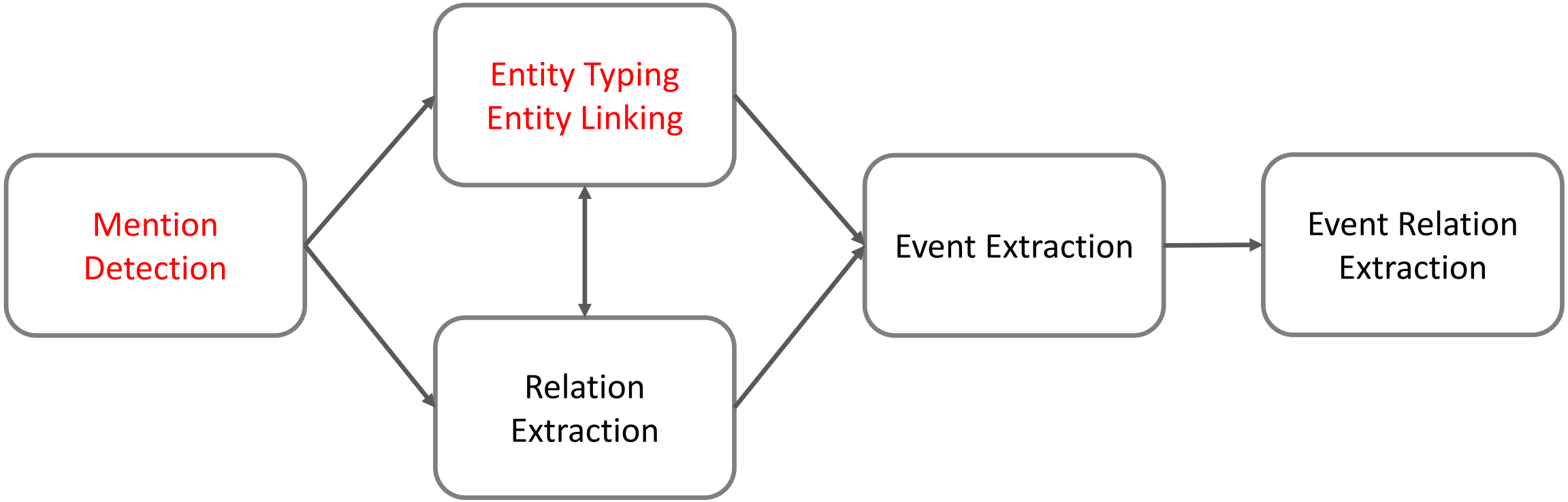


Figure from: Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, Jaewoo Kang: BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020

General Procedure of Information Extraction for Knowledge Graph Construction



Entity Mention Detection and Typing

- Treat mention detection and typing as a word classification problem

- A typical label encoding is the BIO encoding
  - B-NP: beginning of a named entity chunk
  - I-NP: inside of a named entity chunk
  - O: outside of a named entity chunk

The most common fatal bacterial diseases are respiratory infections .

O O O O O O O B-Disease I-Disease O

- Then we can build a multi-class classifier or CRF model over certain features
  - E.g., representations provided by deep models

Entity Mention Detection and Typing

- Can also treat detection and typing as two separate tasks

- First, we use a tagger to extraction all mentions of interests
  - In this case, we can collect more training examples for each label

The most common fatal bacterial diseases are respiratory infections .												
O	O	O		O	O		O		O	B	I	O

- Second, we apply a multi-class (sometimes multi-label) classifier over the mentions

Entity Linking can Improve Entity Typing

- Especially for fine-grained entity typing
  - Most of the tasks are based weak/distant supervision
    - E.g., generating training data using the anchor links in Wikipedia
  - Entity linking to existing knowledge bases provide useful information

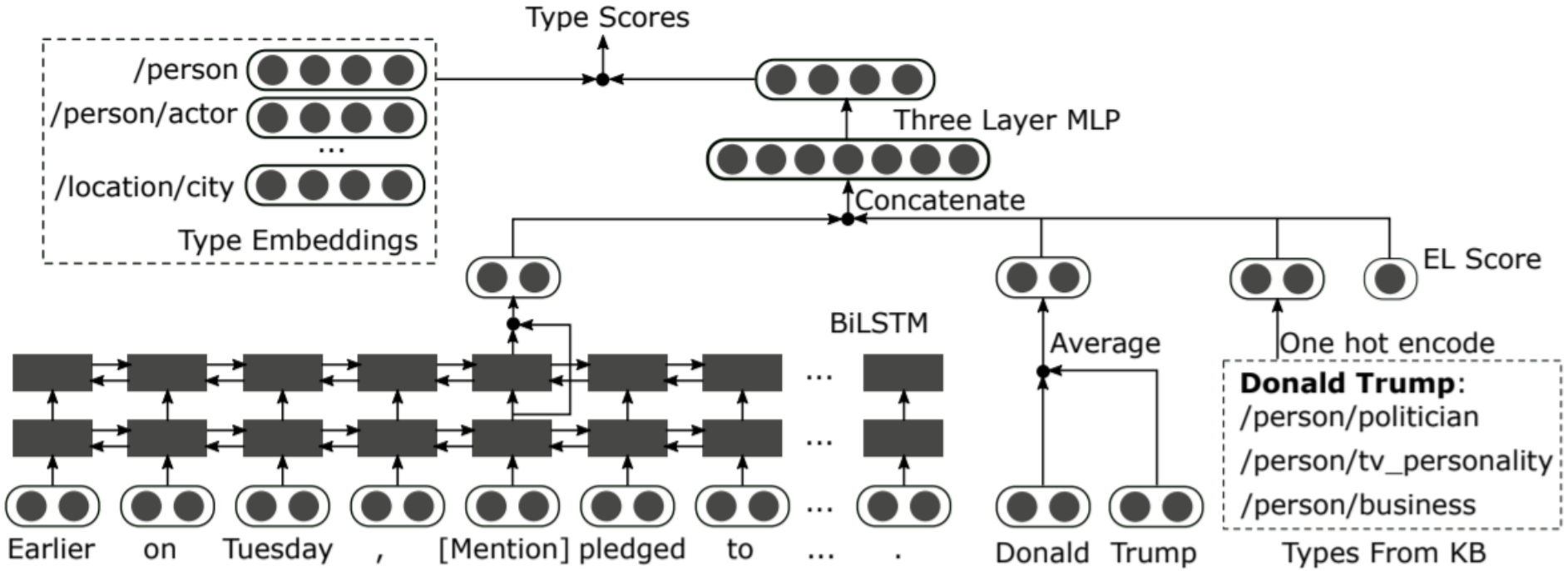


Figure from: Hongliang Dai, Donghong Du, Xin Li, Yangqiu Song: Improving Fine-grained Entity Typing with Entity Linking. EMNLP/IJCNLP (1) 2019
Shikhar Vashishth, Rishabh Joshi, Ritam Dutt, Denis Newman-Griffis, Carolyn Penstein Rosé: MedType: Improving Medical Entity Linking with Semantic Type Prediction. Arxiv, 2020

An Example of COVID-19 NER

- 75 entity types including
  - Common biomedical entity types (e.g., genes, chemicals, and diseases),
  - New types related to COVID-19 (e.g., coronaviruses, viral proteins, evolution, materials, substrates and immune responses)

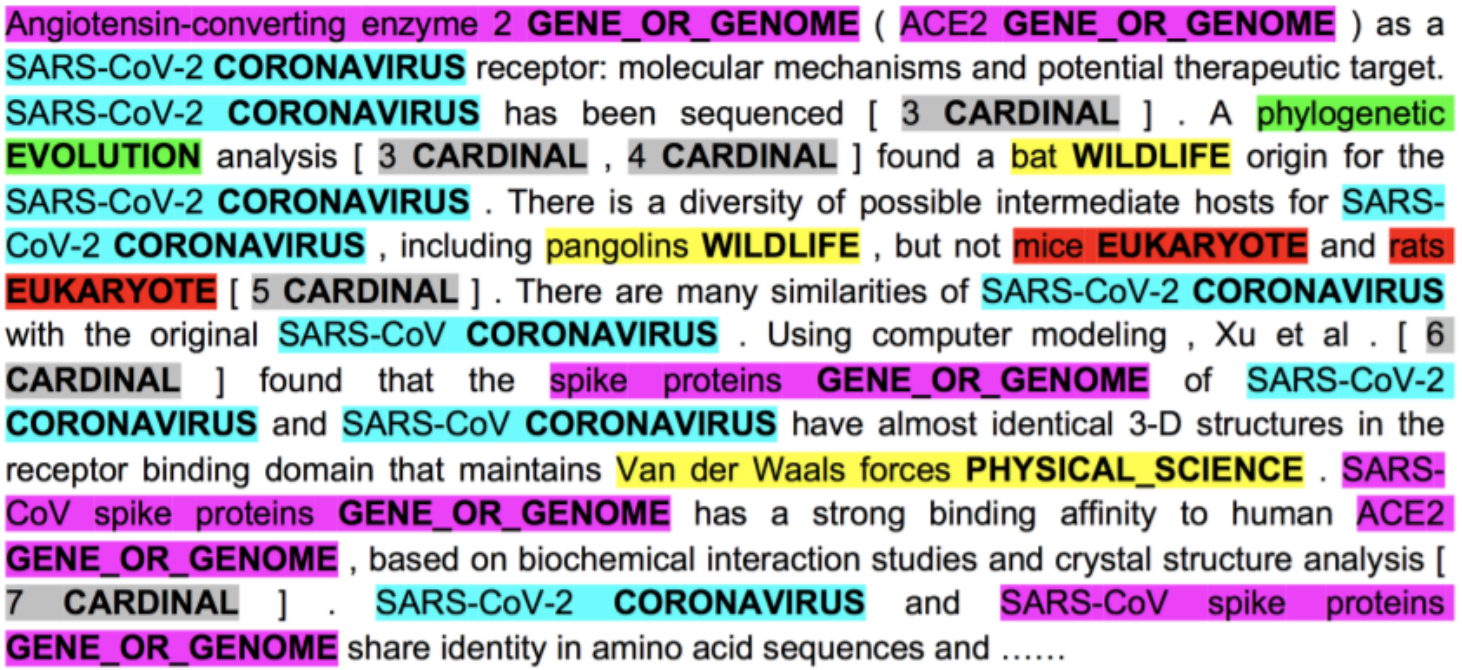


Figure from: Xuan Wang, Xiangchen Song, Yingjun Guan, Bangzheng Li, Jiawei Han: Comprehensive Named Entity Recognition on CORD-19 with Distant or Weak Supervision. CoRR abs/2003.12218 (2020)

Entity Linking

- Link a mention to an existing knowledge graph
- Can be done with a two-step approach
  - Mention detection
  - Disambiguation
- Disambiguation can be helped with relatively coarse-grained entity typing
  - E.g., aggregated 24 super types from 127 fine-grained UMLS types

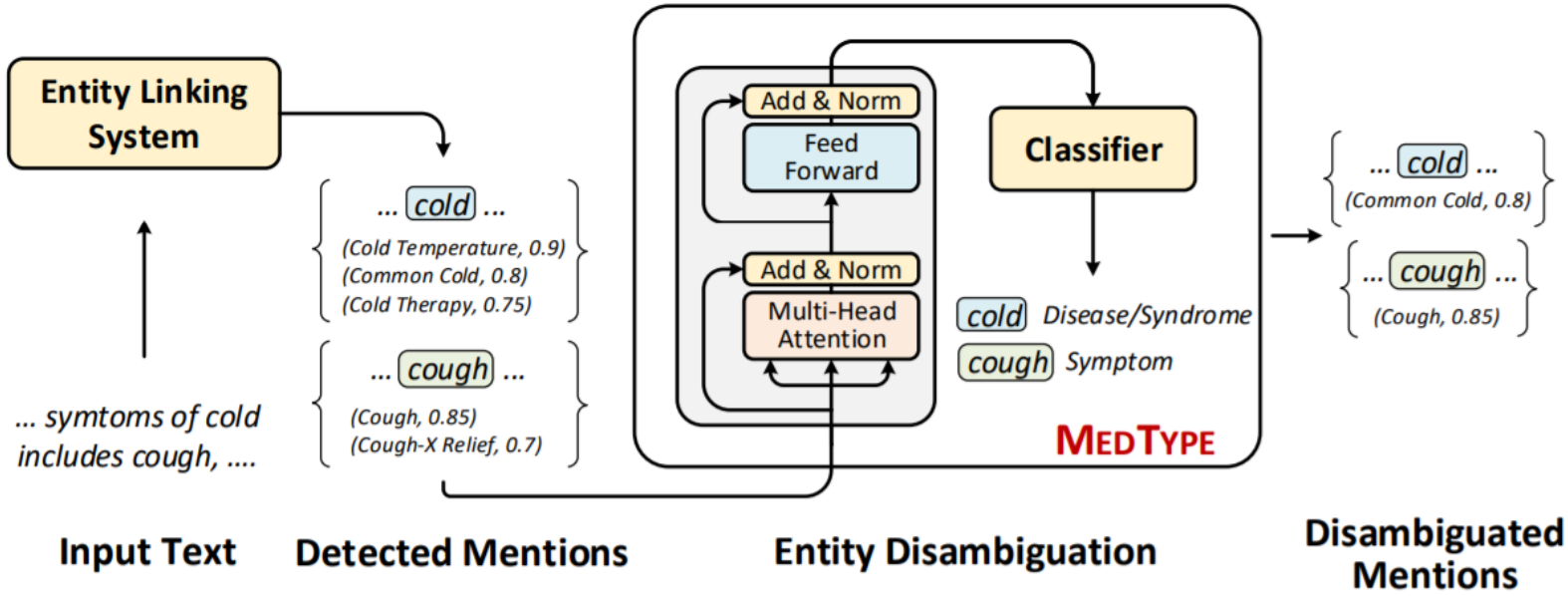


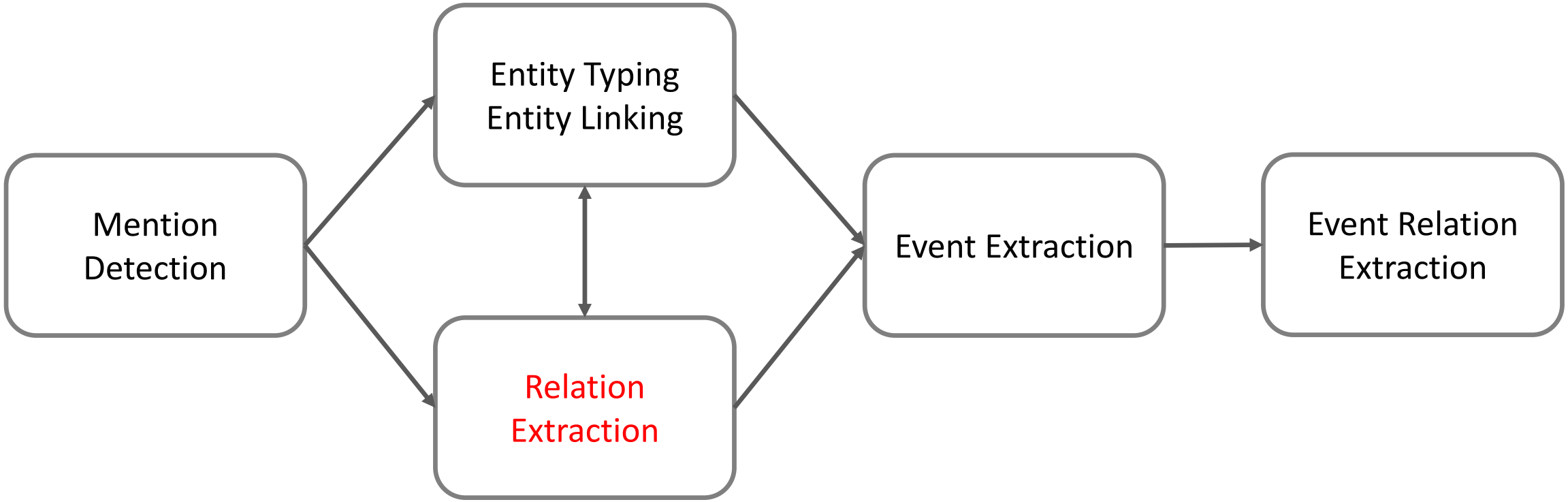
Figure from: Shikhar Vashishth, Rishabh Joshi, Ritam Dutt, Denis Newman-Griffis, Carolyn Penstein Rosé: MedType: Improving Medical Entity Linking with Semantic Type Prediction, Arxiv, 2020

Mention Detection, Typing, and Linking

- Three tasks can be similar in designing models
  - With different annotation scheme and loss functions
- When typing is more difficult
  - E.g., there is lack of annotation
  - Adding linking results into features is helpful
- When linking is more difficult
  - E.g., lots of out-of-sample examples in test set
  - Adding typing can improve the results
- Can also perform joint learning when both types of annotation are available (Leaman and Lu, 2016; Zhao et al., 2019; Mohan and Li, 2019)

Robert Leaman, Zhiyong Lu: TaggerOne: joint named entity recognition and normalization with semi-Markov Models. Bioinform. 32(18): 2839-2846 (2016)
Sunil Mohan and Donghui Li: MedMentions: A Large Biomedical Corpus Annotated with UMLS Concepts. AKBC, 2019
Sendong Zhao, Ting Liu, Sicheng Zhao, Fei Wang: A Neural Multi-Task Learning Framework to Jointly Model Medical Named Entity Recognition and Normalization. AAAI 2019

General Procedure of Information Extraction for Knowledge Graph Construction



Relation Extraction

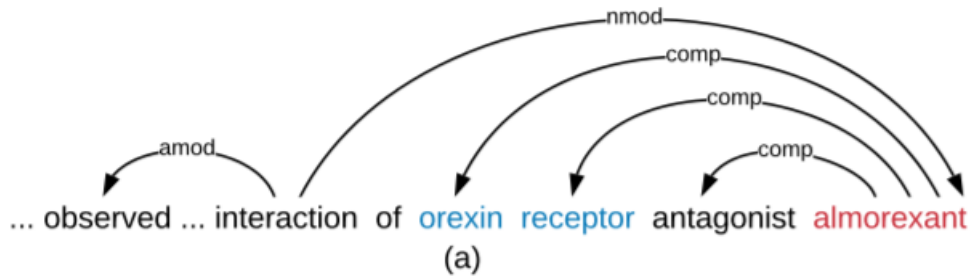
- Many related tasks for relation extraction in biomedical domain
  - Chemical (drug) induced diseases [Jiao Li et al., 2016]
  - Chemical (drug)-protein (gene) interaction [Martin Krallinger et al. 2017]
  - Phenotype (e.g., disease)-gene relations [Diana Sousa et al., 2019]
  - …
- For example, “observed ... interaction of orexin receptor antagonist almorexant”
  - Identity entities, e.g., protein and chemical
  - Classify relations

Group	*Eval.*	*CHEMPROT relations belonging to this group*
CPR:1	N	PART_OF
CPR:2	N	REGULATOR|DIRECT_REGULATOR|INDIRECT_REGULATOR
CPR:3	Y	UPREGULATOR|ACTIVATOR|INDIRECT_UPREGULATOR
CPR:4	Y	DOWNREGULATOR|INHIBITOR|INDIRECT_DOWNREGULATOR
CPR:5	Y	AGONIST|AGONIST-ACTIVATOR|AGONIST-INHIBITOR
CPR:6	Y	ANTAGONIST
CPR:7	N	MODULATOR|MODULATOR-ACTIVATOR|MODULATOR-INHIBITOR
CPR:8	N	COFACTOR
CPR:9	Y	SUBSTRATE|PRODUCT_OF|SUBSTRATE_PRODUCT_OF
CPR:10	N	NOT

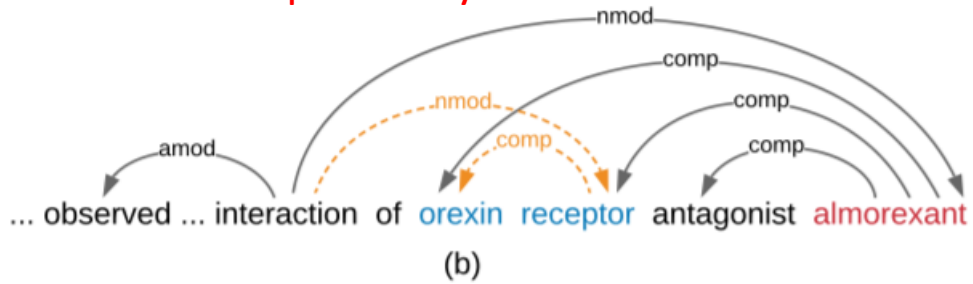
Jiao Li et al., BioCreative V CDR task corpus: a resource for chemical disease relation extraction, Database. 2016
Figure from: Martin Krallinger et al., Overview of the BioCreative VI chemical-protein interaction Track, BioCreative Challenge Evaluation Workshop. 2017
Diana Sousa, Andre Lamurias, Francisco M. Couto: A Silver Standard Corpus of Human Phenotype-Gene Relations. NAACL-HLT (1) 2019

Learning for Relation Extraction

- A recent work show that using dependency parse forest than the 1-best tree with a GNN can improve the performance



Dependency Parse Tree



Dependency Parse Forest

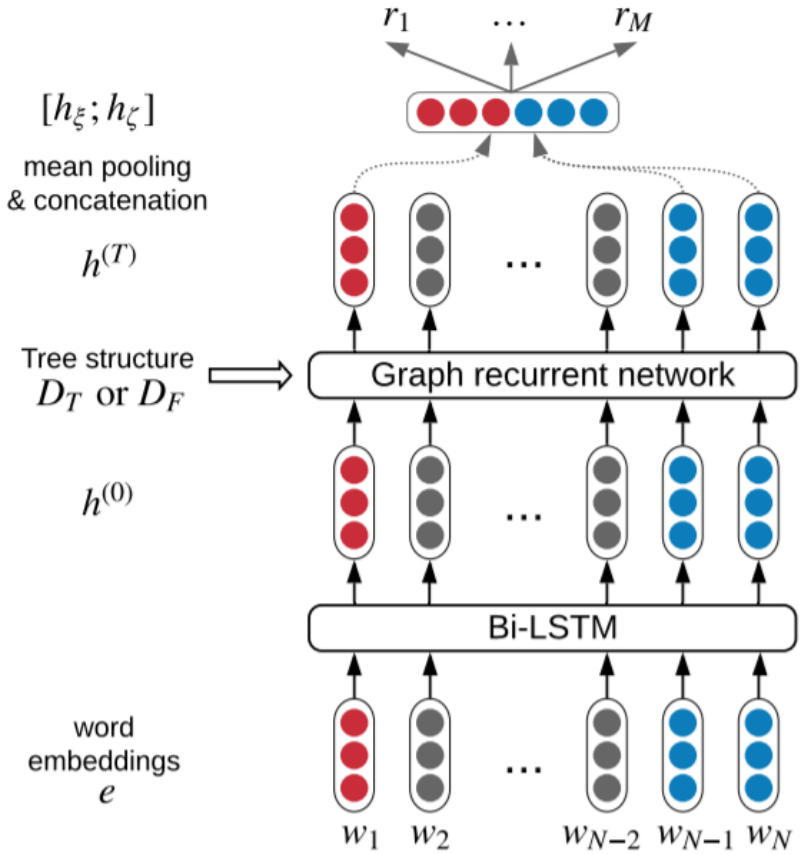


Figure from: Linfeng Song, Yue Zhang, Daniel Gildea, Mo Yu, Zhiguo Wang, Jinsong Su: Leveraging Dependency Forest for Neural Medical Relation Extraction. EMNLP/IJCNLP (1) 2019

N-Ary Relation Extraction

- Many relations (facts) involve more than two arguments, not always possible to decompose them into binary facts without losing information

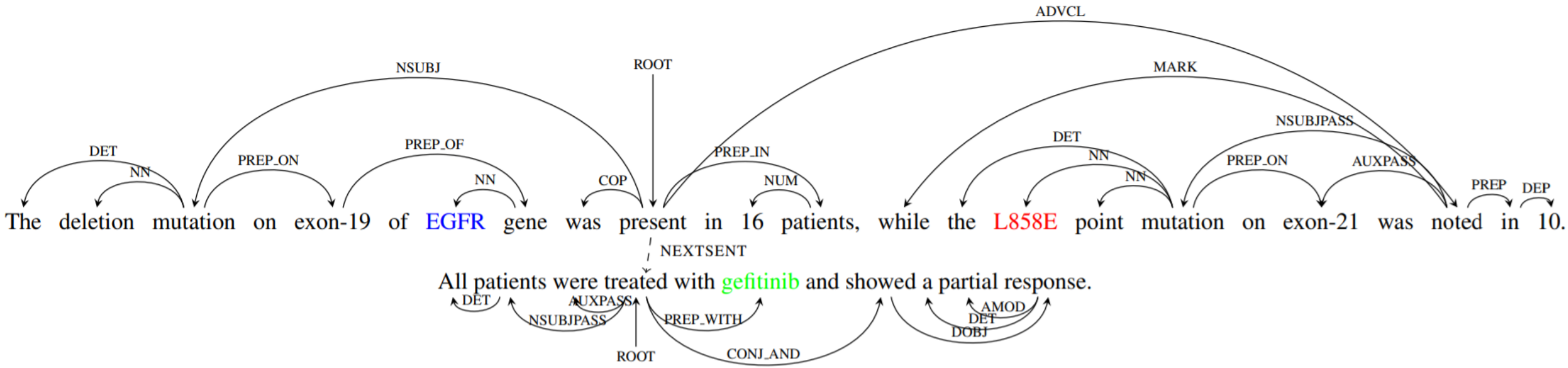
- 2.5 mg Albuterol may be used to treat acute exacerbations, particularly in children.
- Salmonella infection is a common cause of bacteremia in Africa.

Relation	Arity	Signature
Treats	5	$Drug \times Disease \times Dosage \times DosageForm \times Targetgroup$
ReducesRisk	4	$(Drug \cup Behavior \cup Ecofactor) \times Disease \times Targetgroup \times Condition$
Causes	4	$Disease \times Disease \times Targetgroup \times Condition$
Diagnoses	3	$DiagnosticProcedure \times Disease \times (BodyPart \cup Organ)$

Table from: Patrick Ernst, Amy Siu, Gerhard Weikum: HighLife: Higher-arity Fact Harvesting. WWW 2018

Learning for N-Ary Relation Extraction

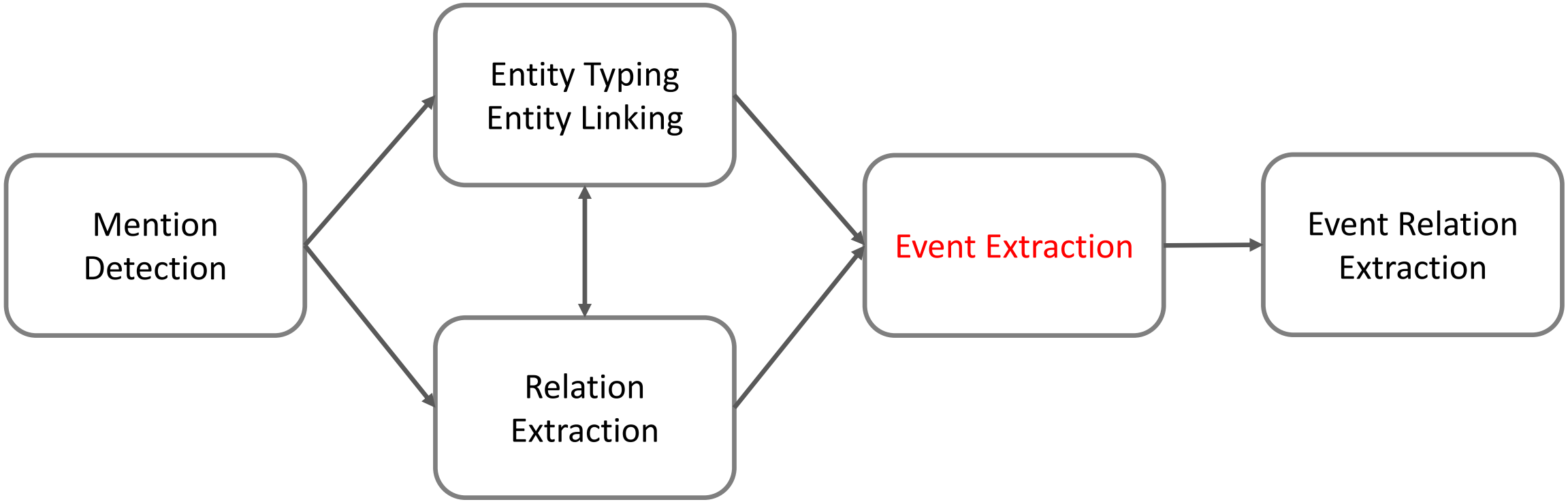
- N-Ary relation extraction with graph LSTM
  - Compared to event extraction, there is no simplification based on Davidsonian semantics (trigger-argument relations)
- Cross-sentence extraction
- Assisted with many distant supervision examples



Fact: tumors with L858E mutation in EGFR gene can be treated with gefitinib.

Figure from: Nanyun Peng, Hoifung Poon, Chris Quirk, Kristina Toutanova, Wen-tau Yih: Cross-Sentence N-ary Relation Extraction with Graph LSTMs. TACL, 2017

General Procedure of Information Extraction for Knowledge Graph Construction



Event Extraction

- According to BioNLP Shared Task 2013, there are several domain-dependent event extraction tasks
  - Genia Event Extraction
  - Cancer Genetics
  - Pathway Curation
  - Corpus Annotation with Gene Regulation Ontology
  - Gene Regulation Network in Bacteria
  - Bacteria Biotopes
- They share similar annotation types
  - Text-bound annotation (entity/event trigger)
  - Equiv: entity aliases
  - E: event
  - M: event modification
  - R: relation
  - N: normalization (external reference)

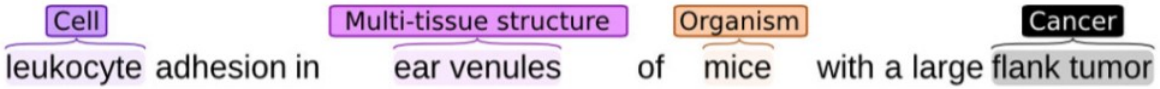
Claire Nédellec, Robert Bossy, Jin-Dong Kim, Jung-Jae Kim, Tomoko Ohta, Sampo Pyysalo, Pierre Zweigenbaum: Overview of BioNLP Shared Task 2013. BioNLP@ACL (Shared Task) 2013

Event Extraction Example

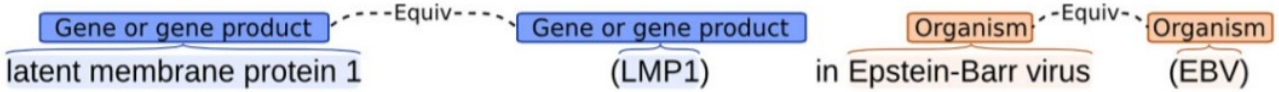
- Cancer genetics
  - 40 event types
  - 18 types of entities
  - 600 PubMed abstracts
  - 17,000 events

- Example Roles:
  - Theme: Entity or event that undergoes the primary effects of the event.
  - Cause: Entity or event that is causally active in the event.

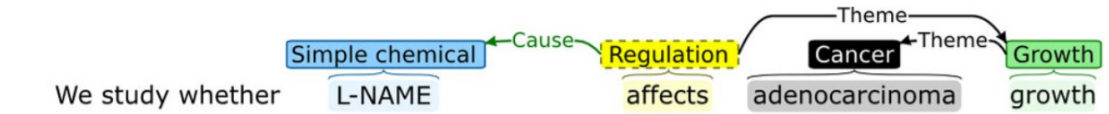
Entity



Relation



Event



Text

```
We study whether L-NAME affects adenocarcinoma growth.
[...]
```

	ID	annotation type	offset	marked text	
Given	T1	Simple_chemical	17 23	L-NAME	entities
	T2	Cancer	32 46	adenocarcinoma	
Targets	T3	Regulation	24 31	affects	triggers
	T4	Growth	47 53	growth	
	E1	Regulation:T3	Cause:T1 Theme:E2		events
	E2	Growth:T4	Theme:T2		
	A1	Speculation	E1		modifications

event/modif. arguments

Figure from: Sampo Pyysalo, Tomoko Ohta, Rafal Rak, Andrew Rowley, Hong-Woo Chun, Sung-Jae Jung, Sung-Pil Choi, Jun'ichi Tsujii, Sophia Ananiadou. BMC Bioinformatics, 2015.

Learning for Event Extraction

- Usually done with a pipeline based approach
  - Step 1: Trigger identification and classification
  - Step 2: Argument identification and role classification
    - Sometimes need to consider entity type constraints

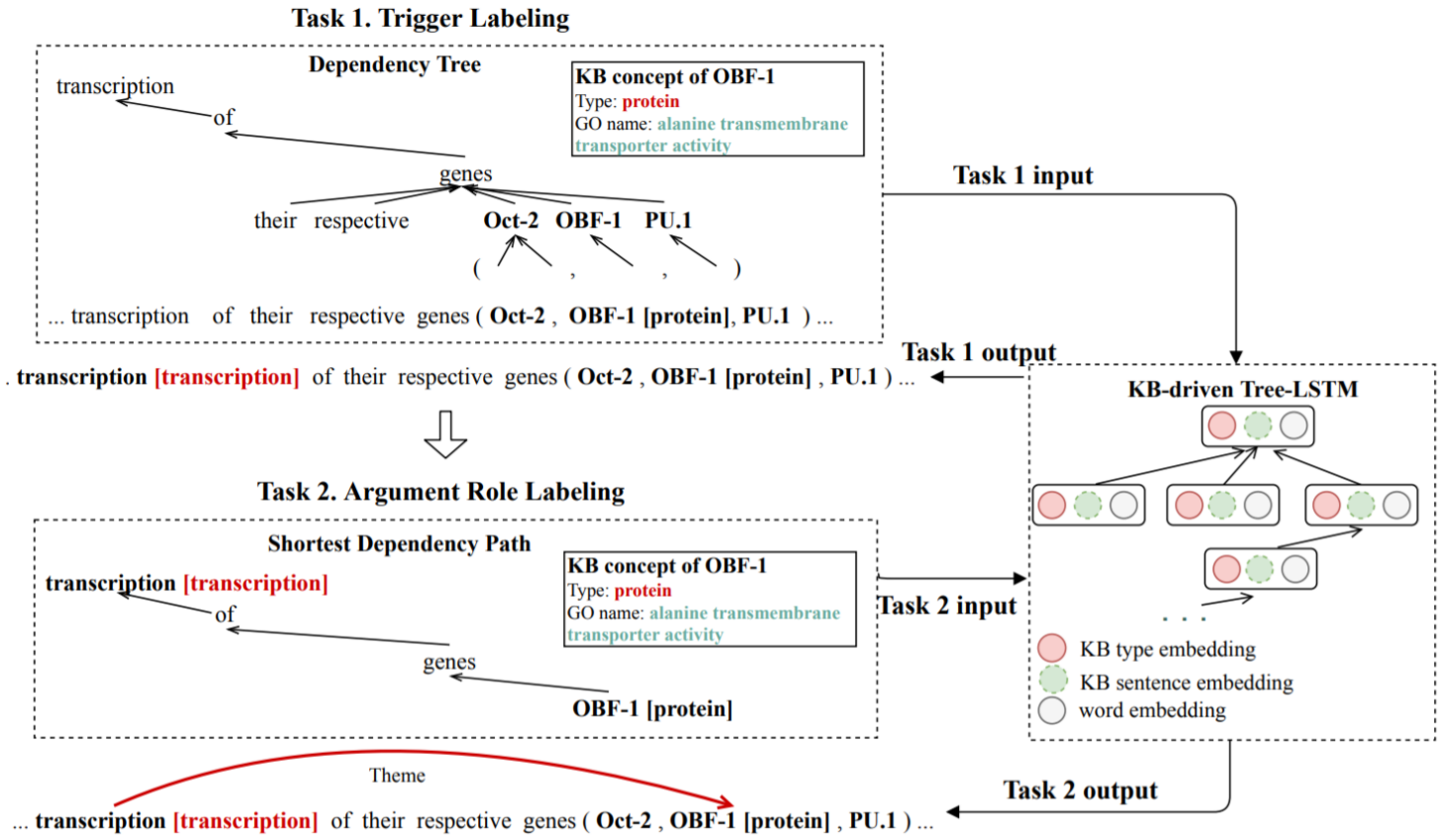
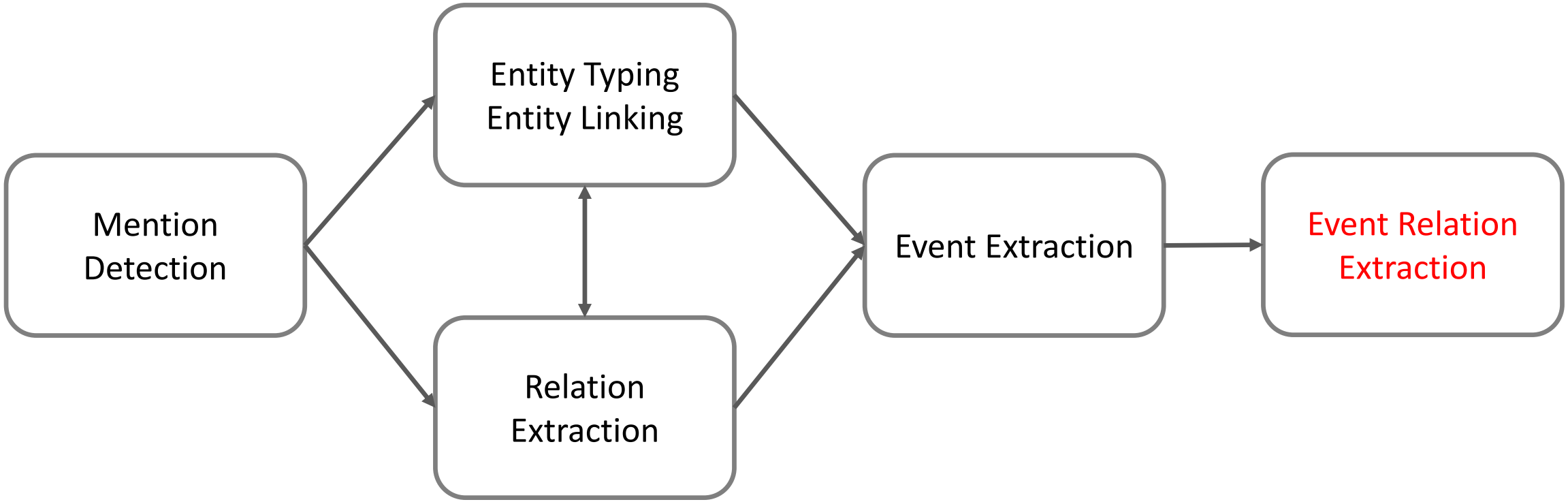


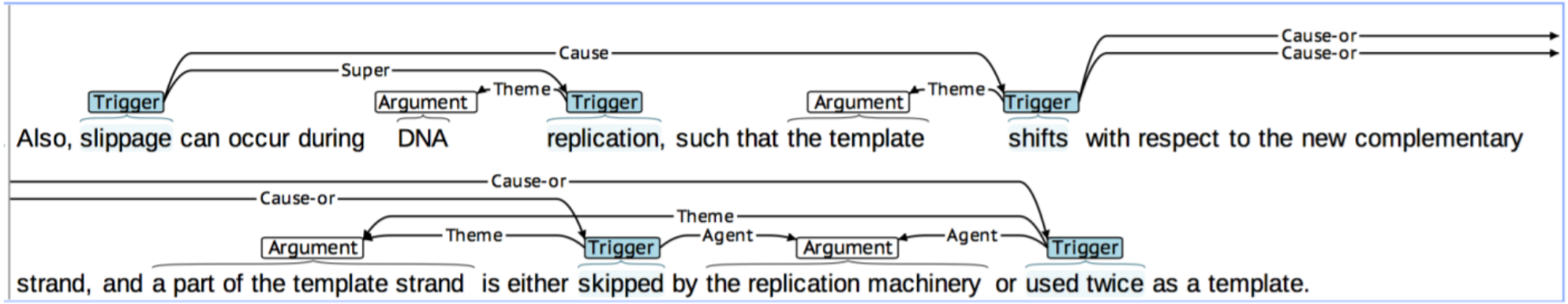
Figure from: Diya Li, Lifu Huang, Heng Ji, Jiawei Han: Biomedical Event Extraction based on Knowledge-driven Tree-LSTM. NAACL-HLT (1) 2019

General Procedure of Information Extraction for Knowledge Graph Construction



Event Relations

- Usually appear in temporal relation extraction, narrative schema, discourse analysis in NLP
- Annotation Scheme
  - Triggers and arguments
  - Semantic roles
  - Event-event relations
    - Cause, Enable, Prevent, Super



Aju Thalappillil Scaria, Jonathan Berant, Mengqiu Wang, Peter Clark, Justin Lewis, Brittany Harding, Christopher D. Manning: Learning Biological Processes with Global Constraints. EMNLP 2013
Figure from: Jonathan Berant, Vivek Srikumar, Pei-Chun Chen, Abby Vander Linden, Brittany Harding, Brad Huang, Peter Clark, Christopher D. Manning: Modeling Biological Processes for Reading Comprehension. EMNLP 2014

Event Relations

- Answering questions

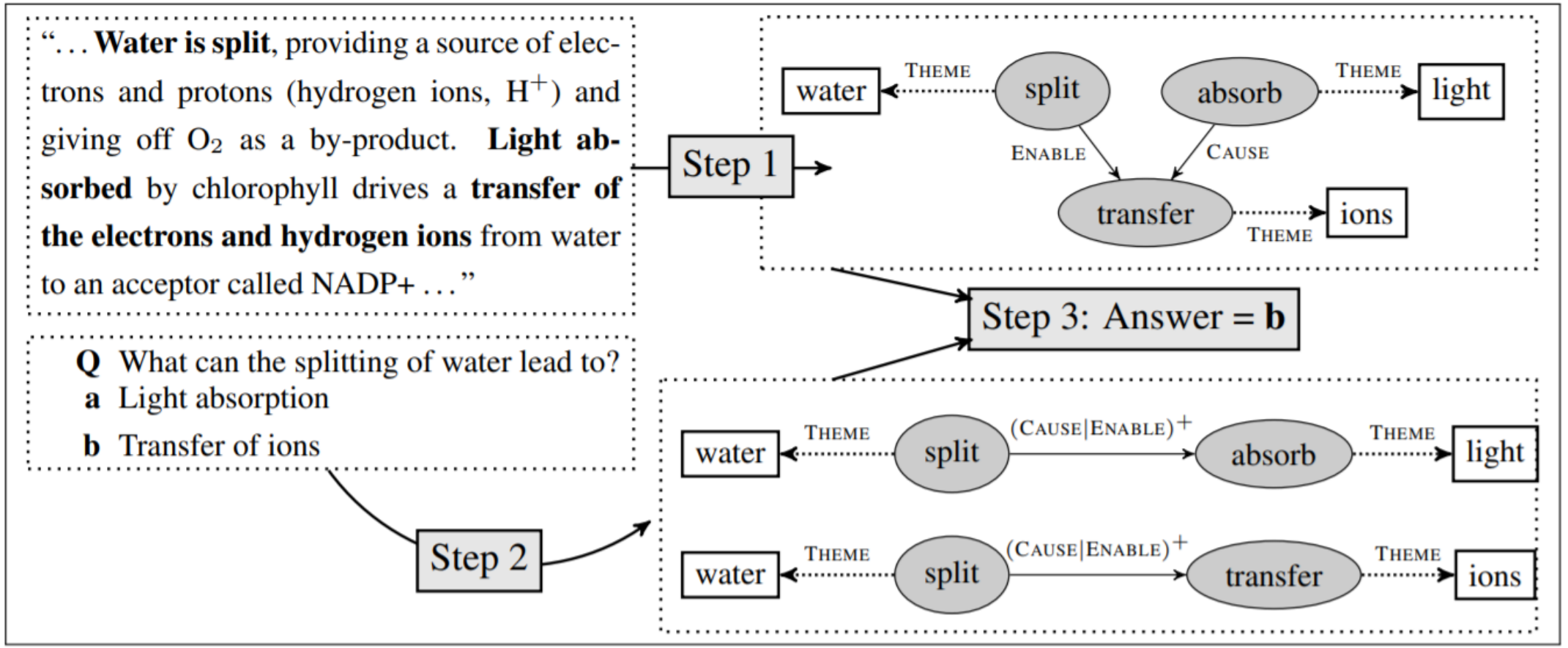


Figure from: Jonathan Berant, Vivek Srikumar, Pei-Chun Chen, Abby Vander Linden, Brittany Harding, Brad Huang, Peter Clark, Christopher D. Manning: Modeling Biological Processes for Reading Comprehension. EMNLP 2014

Conclusions

- In AI for healthcare, many new knowledge graph based technologies have been developed for
  - Natural language processing
    - Generation
    - Dialogue
    - Personalization
  - Recommendation
  - Drug discovery
  - …