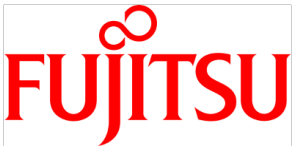
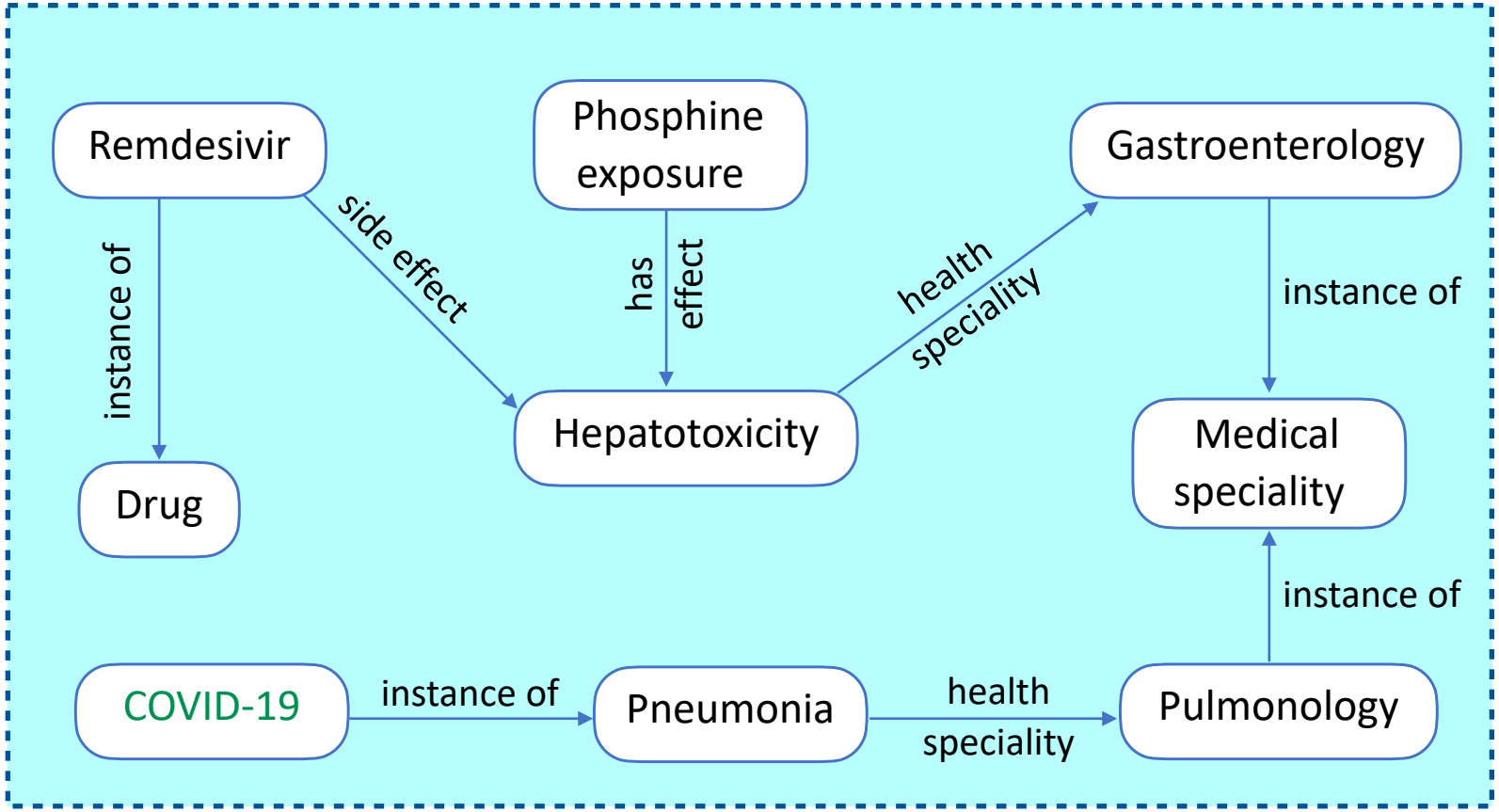
A Systematic Investigation of KB-Text Embedding Alignment at Scale

Vardaan Pahuja Yu Gu Wenhu Chen Mehdi Bahrami
Lei Liu Wei-Peng Chen Yu Su



KBs and text contain complementary knowledge



Knowledge Base

- KBs contain structured knowledge.
- Most KBs are incomplete.

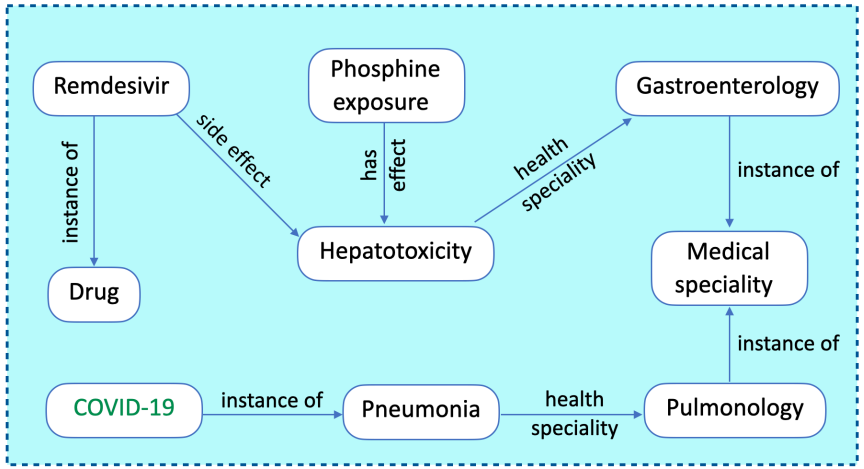
KBs and text contain complementary knowledge

Coronavirus disease 2019 (COVID-19) is a contagious disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). Symptoms of COVID-19 are variable, but often include fever, cough, headache, fatigue, breathing difficulties, and loss of smell and taste. The U.S. FDA has approved the antiviral drug Remdesivir for the treatment of patients with COVID-19.

Article on COVID-19

- Contains encyclopedic knowledge in the form of unstructured texts.
- More timely updated knowledge.

KBs and text contain complementary knowledge



Knowledge Base

Information about:

- COVID-19 symptoms ✗
- COVID-19 disease type ✓
- Side effects of FDA-approved COVID-19 drug ✓

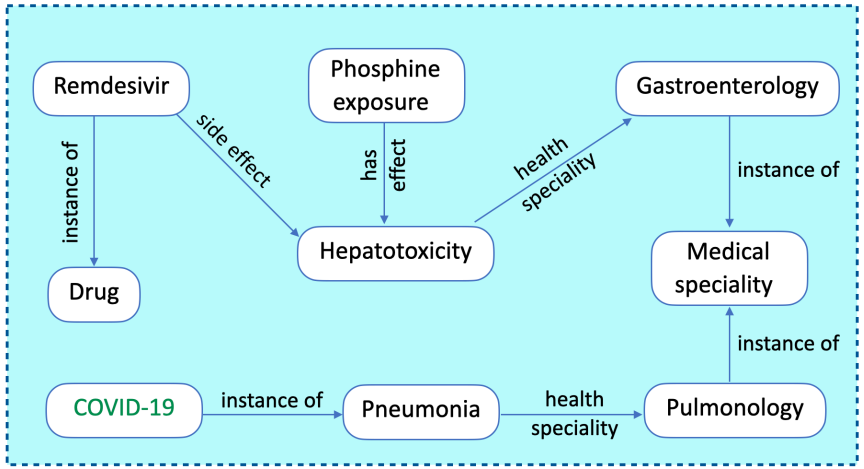
Coronavirus disease 2019 (COVID-19) is a contagious disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). Symptoms of COVID-19 are variable, but often include fever, cough, headache, fatigue, breathing difficulties, and loss of smell and taste. The U.S. FDA has approved the antiviral drug Remdesivir for the treatment of patients with COVID-19.

Article on COVID-19

Information about:

- COVID-19 symptoms ✓
- COVID-19 disease type ✗
- Side effects of FDA-approved COVID-19 drug ✗

KBs and text contain complementary knowledge



Knowledge Base

+

Coronavirus disease 2019 (COVID-19) is a contagious disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). Symptoms of COVID-19 are variable, but often include fever, cough, headache, fatigue, breathing difficulties, and loss of smell and taste. The U.S. FDA has approved the antiviral drug Remdesivir for the treatment of patients with COVID-19.

Article on COVID-19

Information about:

- COVID-19 symptoms ✓
- COVID-19 disease type ✓
- Side effects of FDA-approved COVID-19 drug ✓

Motivation

- Methods to separately embed KBs and text into a vector space(s).

Motivation

- Methods to separately embed KBs and text into a vector space(s).
- Will *aligning* the KB and text vector spaces be an effective way to inject KB information into text embedding and vice versa?

Motivation

- Methods to separately embed KBs and text into a vector space(s).
- Will *aligning* the KB and text vector spaces be an effective way to inject KB information into text embedding and vice versa?
- If so, what is the best alignment method?

Main Contributions

- First systematic investigation on KB-text embedding alignment at scale
  - Wikidata: 14.6M entities, 1.2K relations, 261M facts
  - Wikipedia: 8.2M articles, 2.1M words, 12.3M entities

Main Contributions

- First systematic investigation on KB-text embedding alignment at scale
  - Wikidata: 14.6M entities, 1.2K relations, 261M facts
  - Wikipedia: 8.2M articles, 2.1M words, 12.3M entities
- Evaluation framework with two tasks:
  - Few-shot link prediction: text to KB
  - Analogical reasoning: KB to text

Main Contributions

- First systematic investigation on KB-text embedding alignment at scale.
  - Wikidata: 14.6M entities, 1.2K relations, 261M facts
  - Wikipedia: 8.2M articles, 2.1M words, 12.3M entities
- Evaluation framework with two tasks:
  - Few-shot link prediction: text to KB
  - Analogical reasoning: KB to text
- Release joint KB-text embeddings trained on the largest-scale data to date.

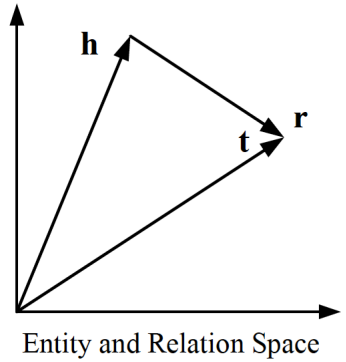
Outline

- Alignment Methods
- Evaluation Tasks
- Results
- Case Study on COVID-19

[Bordes et al. 2013]

Alignment Methods

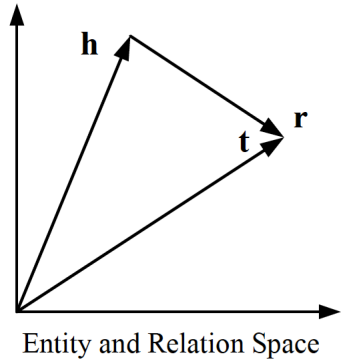
- TransE as the KB embedding model



[Yamada et al. 2016]

Alignment Methods

- TransE as the KB embedding model



- Skip-gram as text embedding model (text = words + entities)

Article: **United Nations Secretariat**

Headquartered in New York, the Secretariat functions through duty stations in Addis Ababa, Bangkok, Beirut, Geneva, Nairobi, Santiago and Vienna, in addition to offices all over the world

Word-Word co-occurrences: (Headquartered, in)
Word-Entity co-occurrences: (Headquartered, New York)
Entity-Entity co-occurrences: (United Nations Secretariat, New York)

→ Skip-gram model → Word and Entity Embeddings

- Both of these have a *linear structure* in their embedding space

Alignment Methods

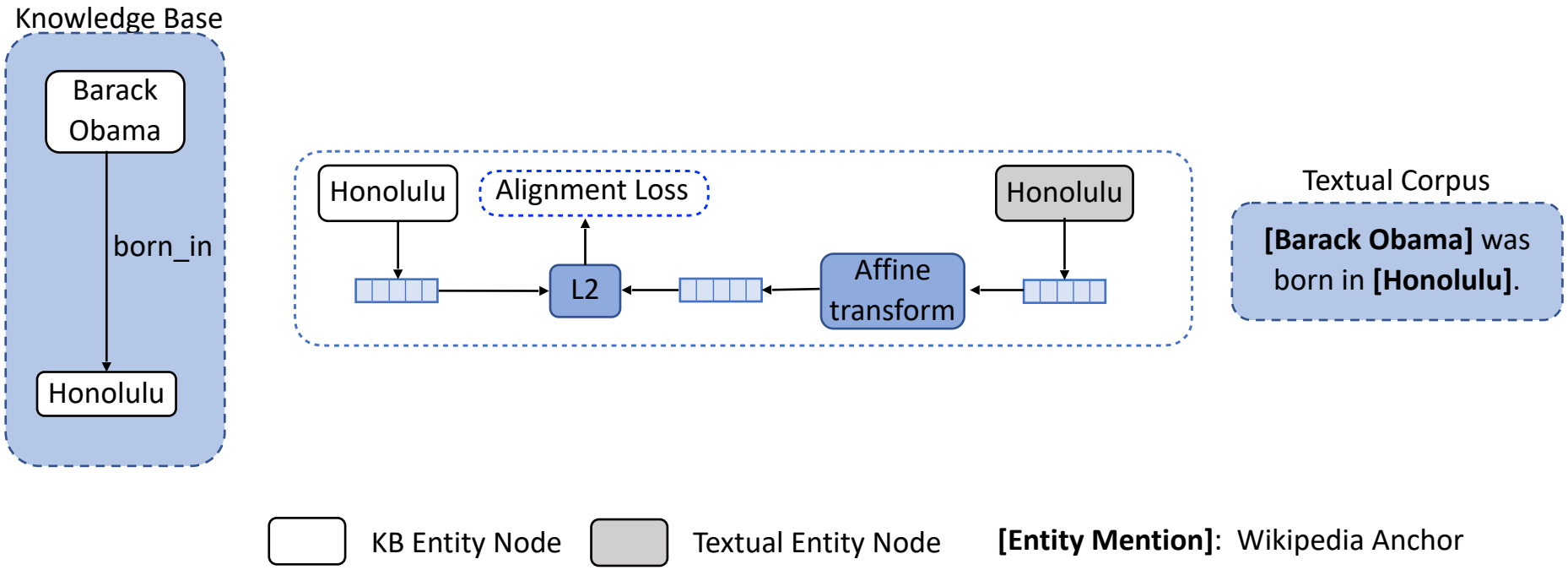
- Alignment using Projection
- Alignment using Entity Names
- Alignment using Same Embedding
- Alignment using Wikipedia Anchors

Training Objective $\mathcal{L} = \mathcal{L}_{KB} + \mathcal{L}_{SG} + \lambda \mathcal{L}_{align}$

λ = Balance parameter

Alignment using Projection

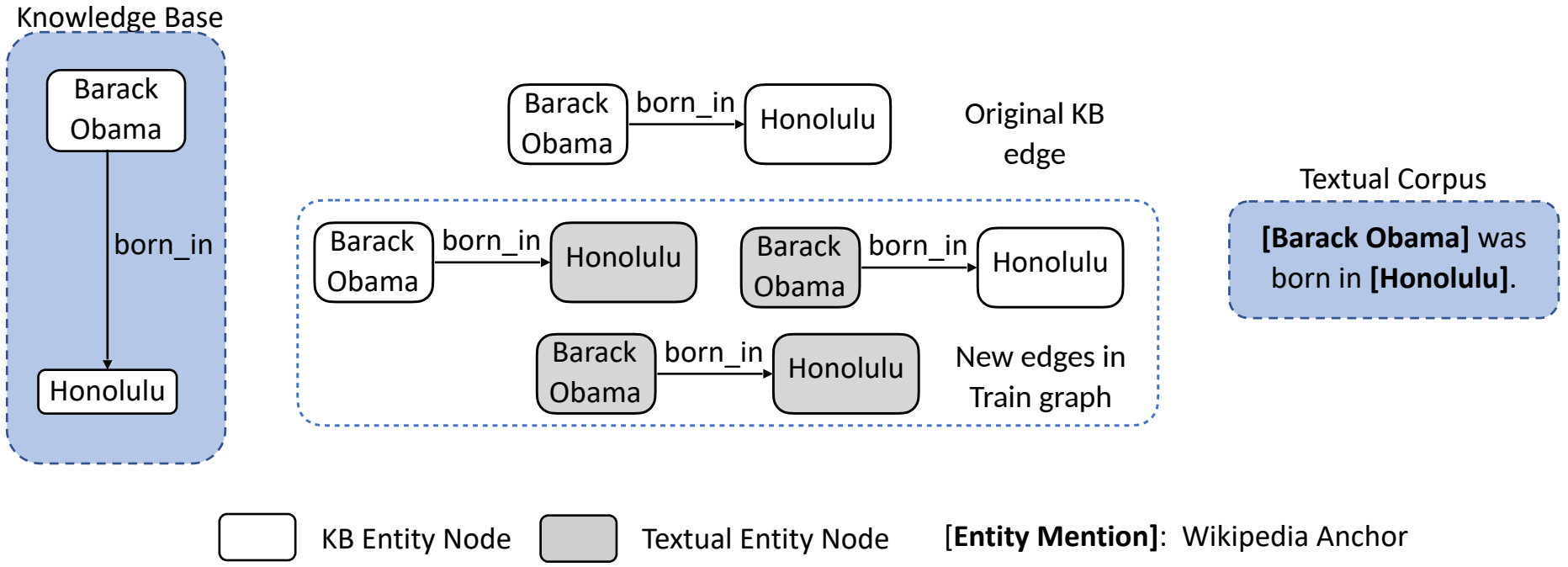
Key idea: Use an affine transformation between embeddings for alignment.



[Wang et al. 2014]

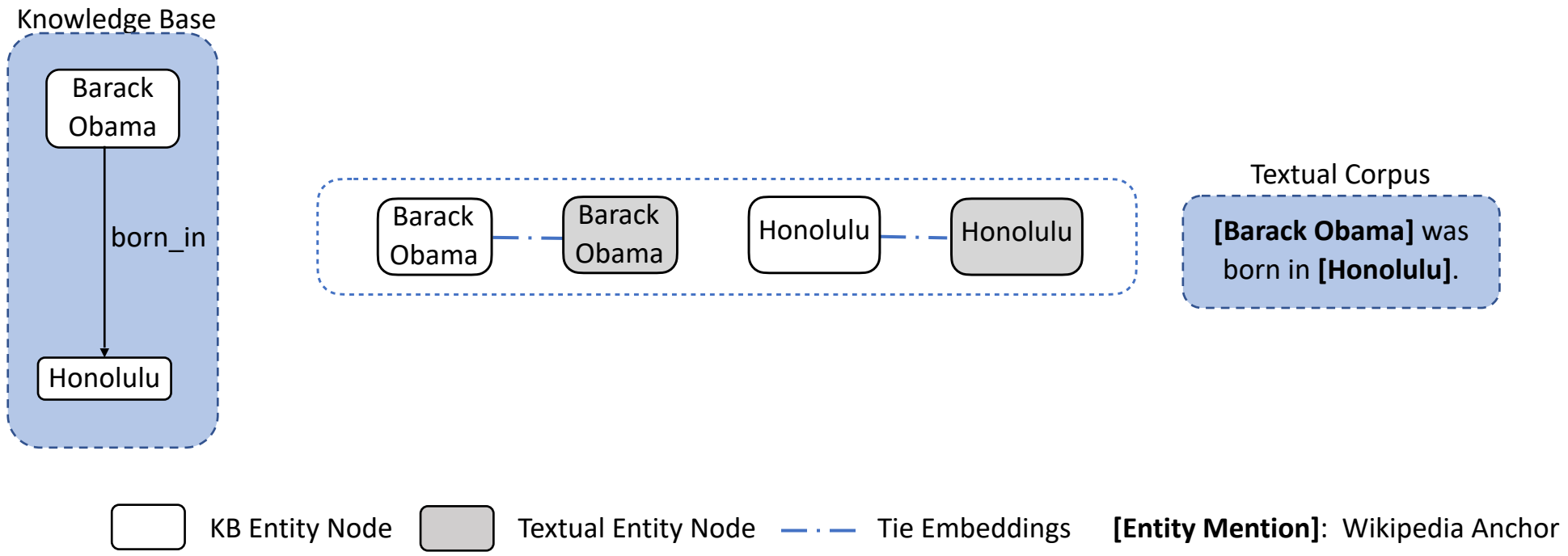
Alignment using Entity Names

Key idea: Introduce new edges in the KB in cases when a KB entity has an equivalent in the text corpus.



Alignment using Same Embedding

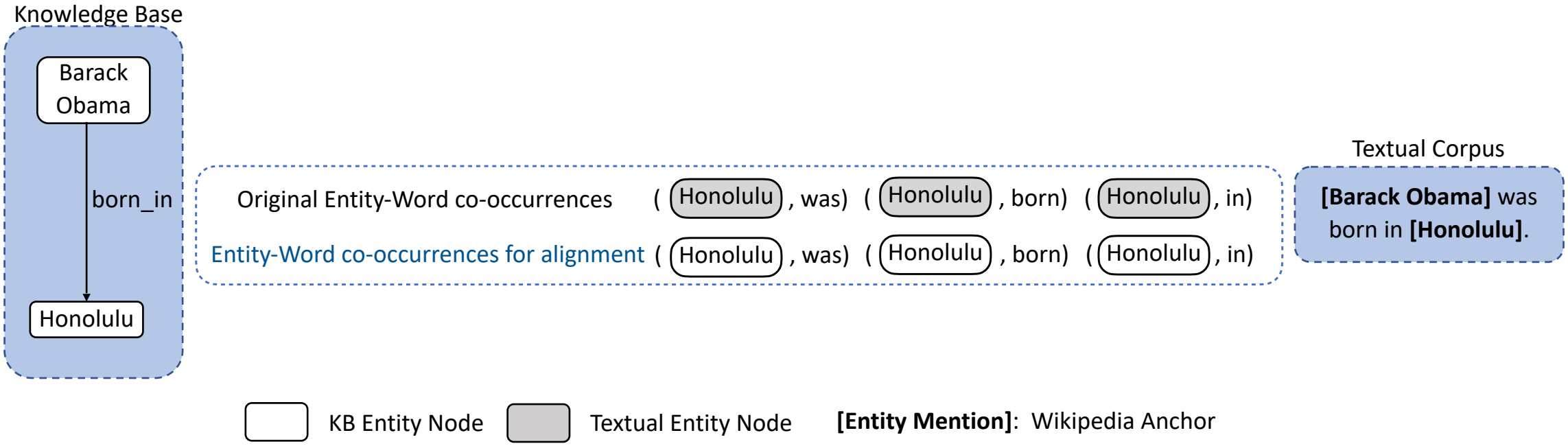
Key idea: Use same entity representations for KB and textual entities.



[Wang et al. 2014]

Alignment using Wikipedia Anchors

Key idea: Substitute the textual entity embedding by its KB counterpart in the skip-gram objective.

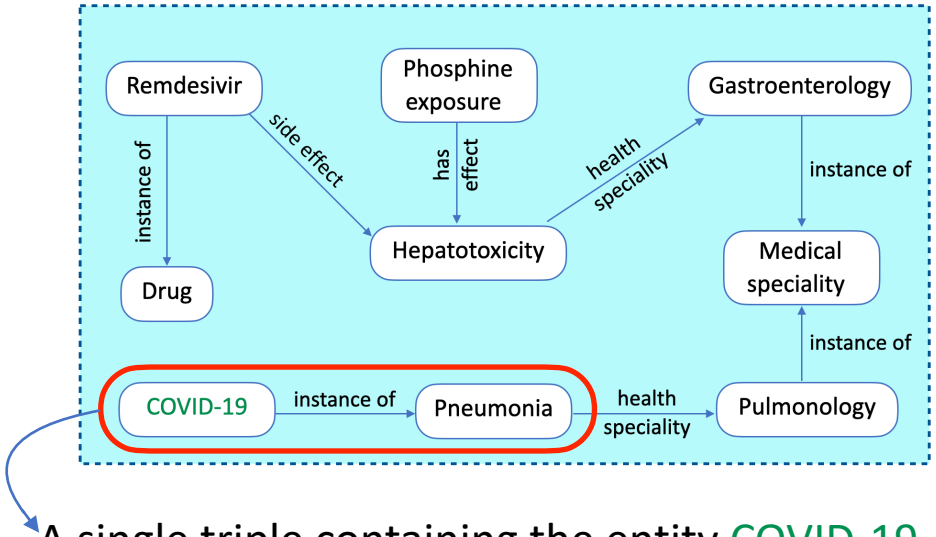


Outline

- Model
- Evaluation Tasks
- Results
- Case Study on COVID-19

Few-Shot Link Prediction

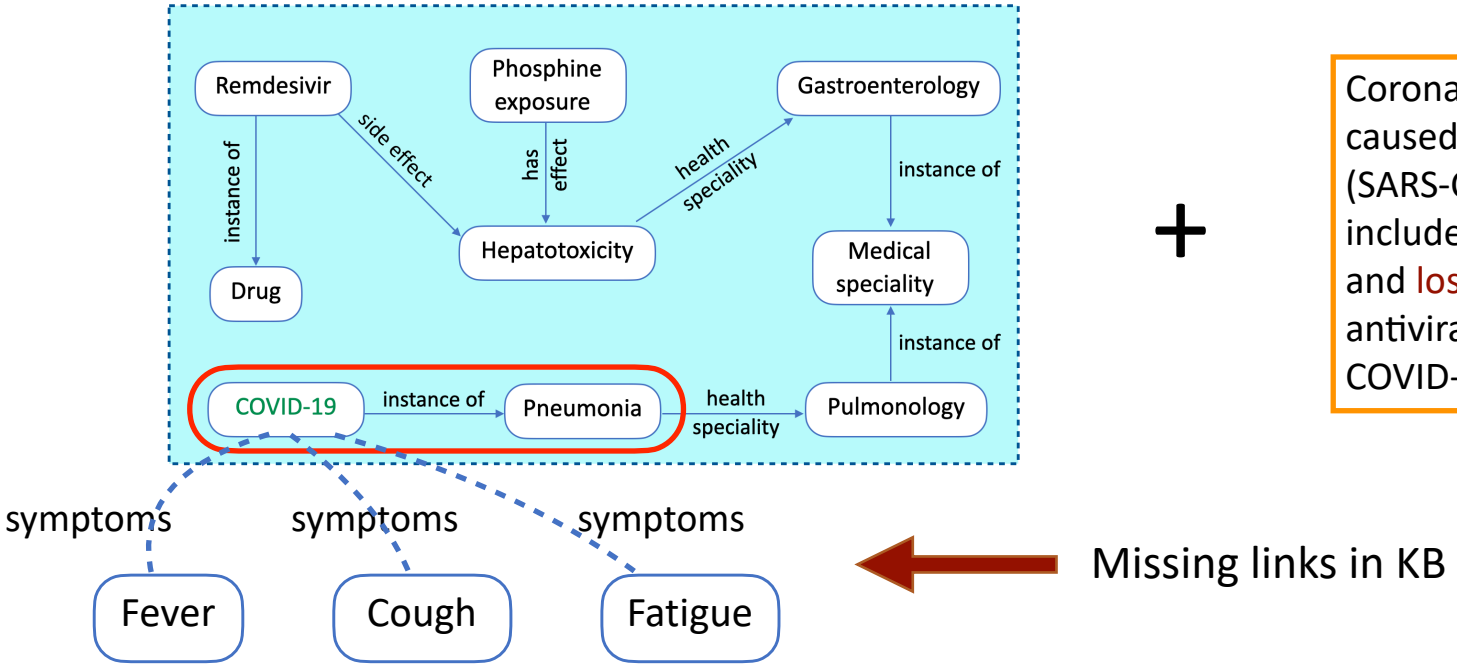
- Key idea: Do link prediction for entities occurring rarely in the training set.



A single triple containing the entity COVID-19.

Few-Shot Link Prediction

- Key idea: Do link prediction for entities occurring rarely in the training set.



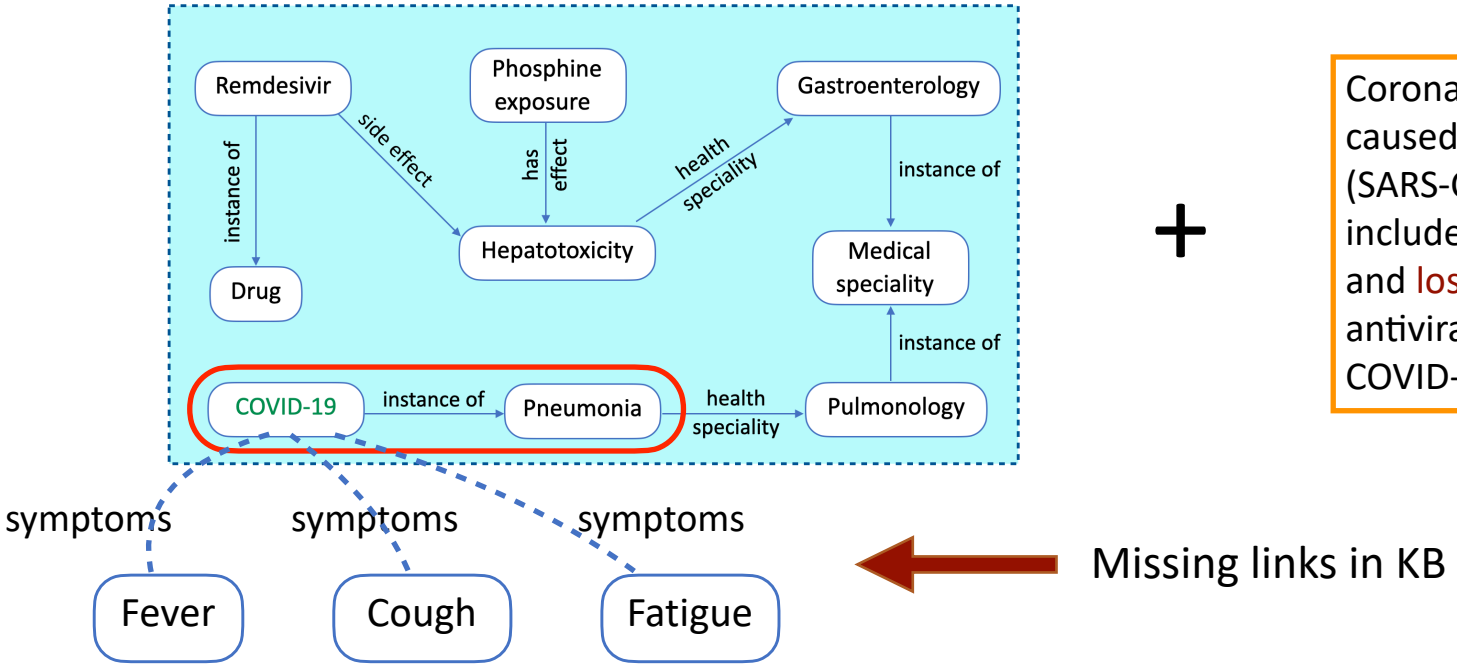
+

Coronavirus disease 2019 (COVID-19) is a contagious disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). Symptoms of COVID-19 are variable, but often include fever, cough, headache, fatigue, breathing difficulties, and loss of smell and taste. The U.S. FDA has approved the antiviral drug Remdesivir for the treatment of patients with COVID-19.

Missing links in KB

Few-Shot Link Prediction

- Key idea: Do link prediction for entities occurring rarely in the training set.



+

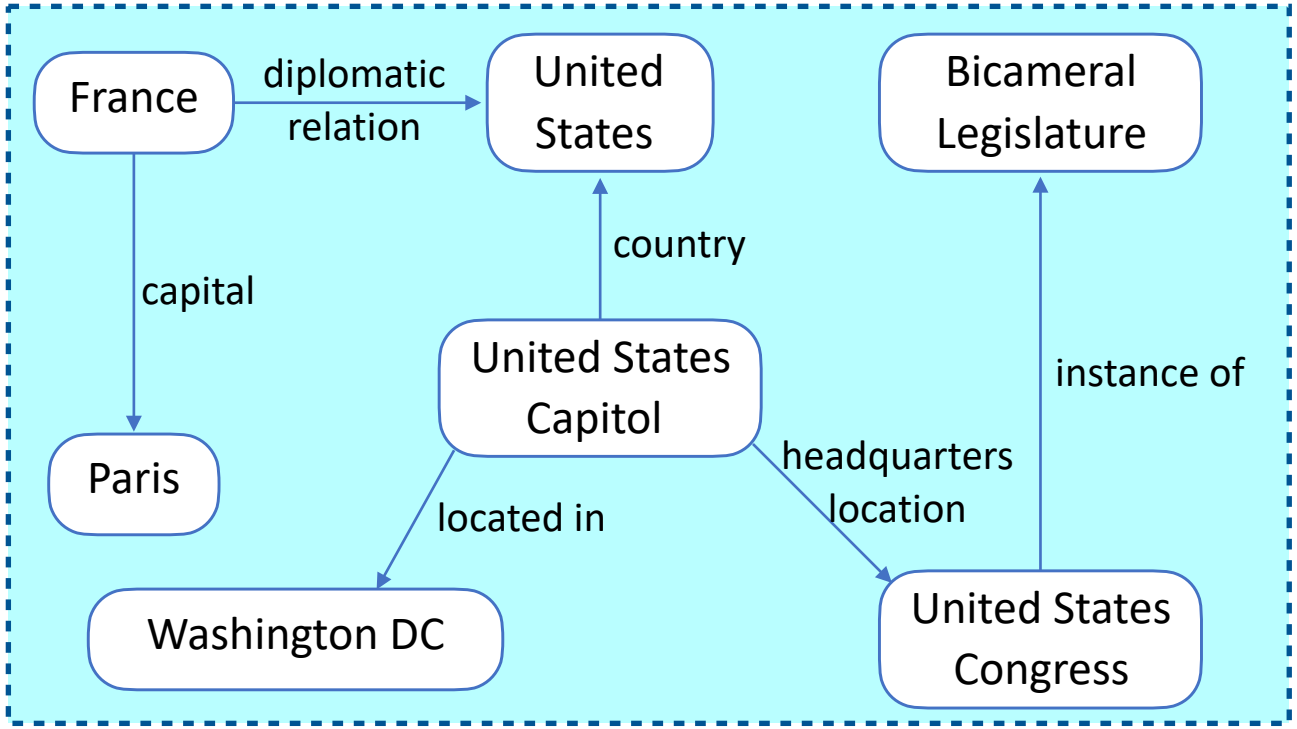
Coronavirus disease 2019 (COVID-19) is a contagious disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). Symptoms of COVID-19 are variable, but often include fever, cough, headache, fatigue, breathing difficulties, and loss of smell and taste. The U.S. FDA has approved the antiviral drug Remdesivir for the treatment of patients with COVID-19.

Missing links in KB

- Test set: Contains relational triples corresponding to a subset of KB entities.
- Train set: Contains just one triple for every test set entity (hence few-shot).

Analogical Reasoning

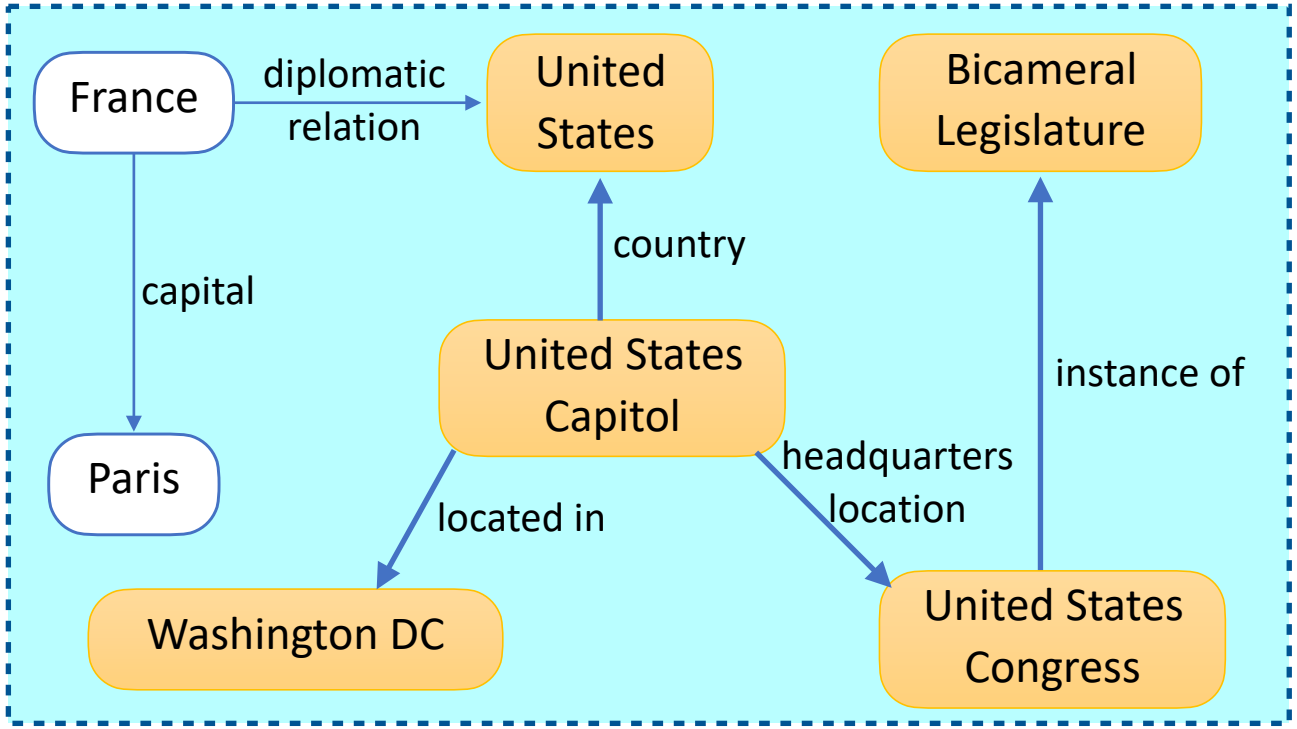
- (France: Paris :: United States: ?)



The **United States of America** is a country primarily located in North America. It consists of 50 states, a federal district, five major unincorporated territories, 326 Indian reservations, and some minor possessions. The most populous city is New York City.

Analogical Reasoning

- (France: Paris :: United States: ?) -> Washington DC ✓



The **United States of America** is a country primarily located in North America. It consists of 50 states, a federal district, five major unincorporated territories, 326 Indian reservations, and some minor possessions. The most populous city is New York City.

Analogical Reasoning

- (h1, t1) -> Training triples set and (h2, t2) -> Test triples set, connected by same relation.
- 50 relations (one-to-one and many-to-one)
  - spouse
  - country of citizenship
  - place of birth
  - capital of
- 1000 examples per relation

Outline

- Model
- Evaluation Tasks
- Results
- Case Study on COVID-19

Few-Shot Link Prediction

- Alignment methods significantly outperform the naive TransE baseline.

Model	MR	Hits@1	Hits@10
TransE	187	20.3	40.4
Projection	134	22.9	47.2
Same Embedding align.	**102**	**30.7**	**51.8**
Entity Name align.	116	23.1	46.7
Wikipedia Anchors align.	138	25.8	46.2

Analogical Reasoning

- Alignment methods significantly outperform the naive skip-gram baseline.

Model	MR	Hits@1	Hits@10
Skip-gram	25	50.6	78.0
Projection	12	65.9	89.0
Same Embedding align.	11	60.7	87.5
Entity Name align.	**8**	**66.5**	**91.0**
Wikipedia Anchors align.	14	56.1	84.8

Discussion

- Entity name alignment introduces new edges involving textual entity embeddings
- KB embedding objective incorporates relational information

Model	MR	Hits@1	Hits@10
Skip-gram	25	50.6	78.0
Projection	12	65.9	89.0
Same Embedding align.	11	60.7	87.5
Entity Name align.	**8**	**66.5**	**91.0**
Wikipedia Anchors align.	14	56.1	84.8

Outline

- Model
- Evaluation Tasks
- Results
- Case Study on COVID-19

Case Study on COVID-19

- Knowledge base completion on emerging events and entities.
- Relations of interest
  - Risk Factor
  - Symptoms
  - Medical Condition
  - Cause of Death

Case Study on COVID-19

- Knowledge base completion on emerging events and entities.
- Relations of interest
  - Risk Factor
  - Symptoms
  - Medical Condition
  - Cause of Death
- Use the March 2020 Wikidata and December 2020 Wikipedia to train the alignment models.
- Evaluate on the difference of COVID related triples between March 2020 and December 2020 snapshots of Wikidata.

Case Study on COVID-19

- Alignment methods outperform the TransE model in majority of cases.

Relation	TransE	Projection	Same Embed.
Risk factor	312	261	**153**
Symptoms	37	**36**	39
Medical cond.	371	**267**	330
Cause of death	314	**246**	299

Key Takeaways

- Joint reasoning through alignment enhances both KB and text entity representations.
- The inductive bias of a particular alignment method can affect its performance on a particular evaluation task.

Contact: pahuja.9@osu.edu

Code: https://github.com/dki-lab/joint-kb-text-embedding

Thank You!