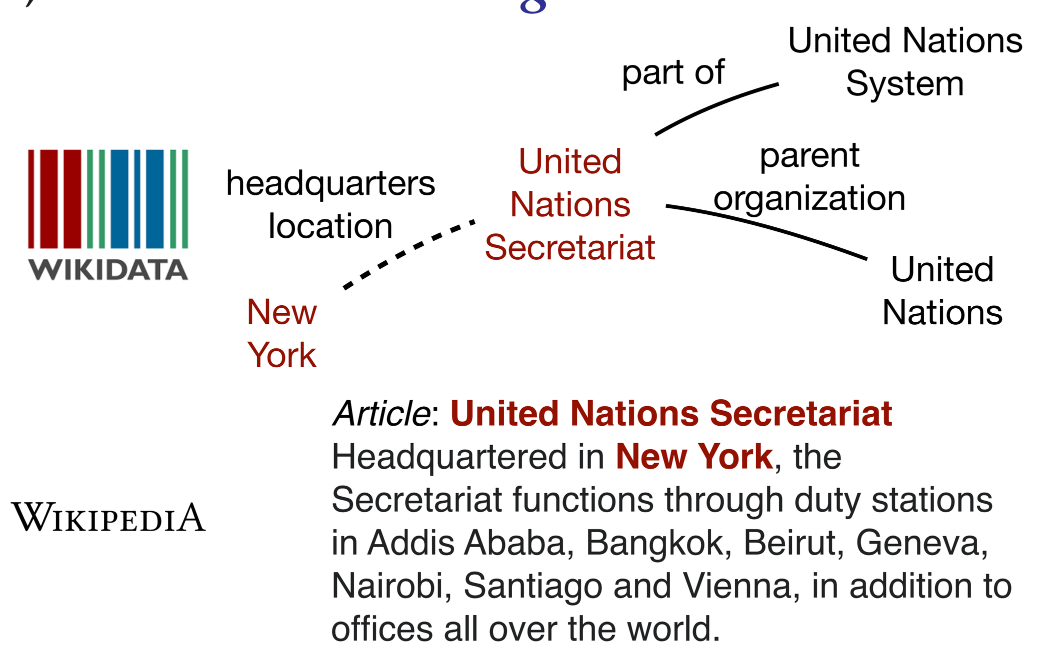
A Systematic Investigation of KB-Text Embedding Alignment at Scale

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

The Ohio State University · UCSB · Fujitsu

Motivation

- Methods to separately embed KBs and text into a vector space(s) have been well-studied.
- 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?***

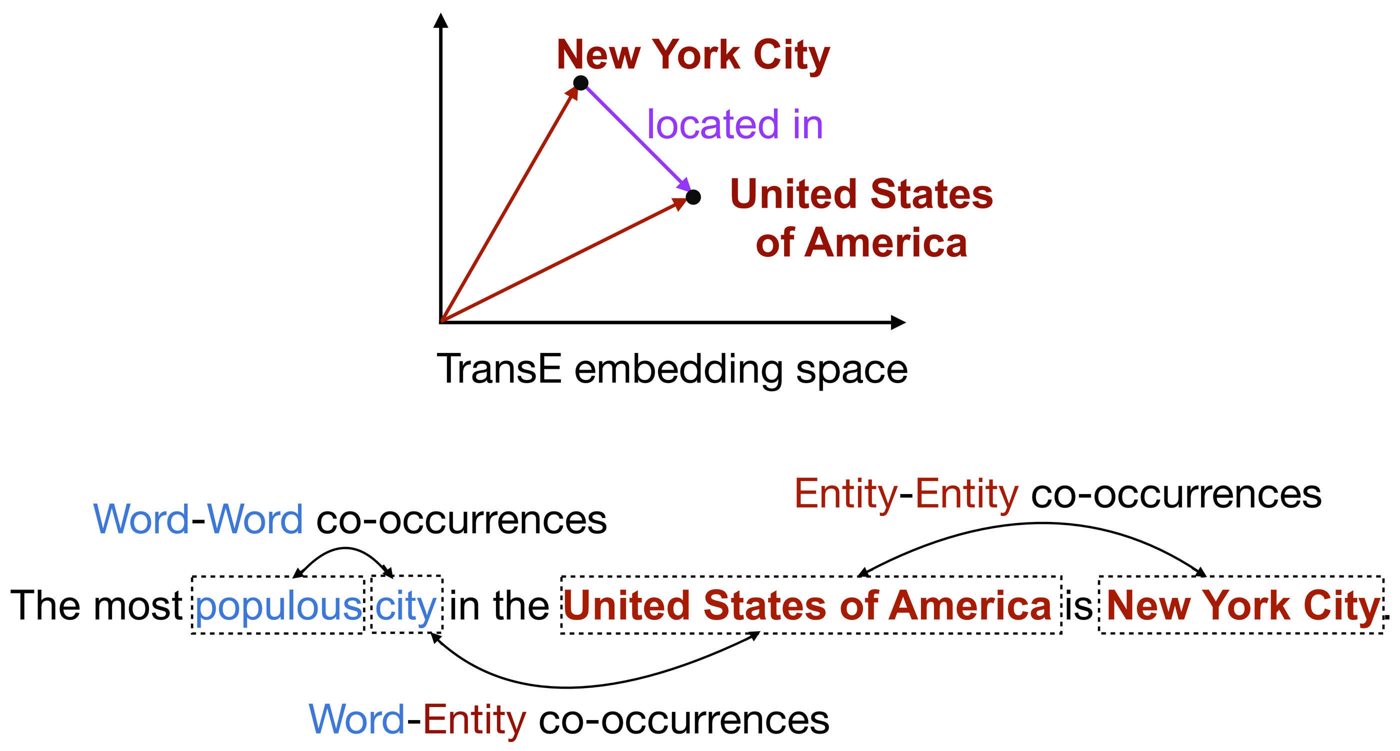


Key Contributions

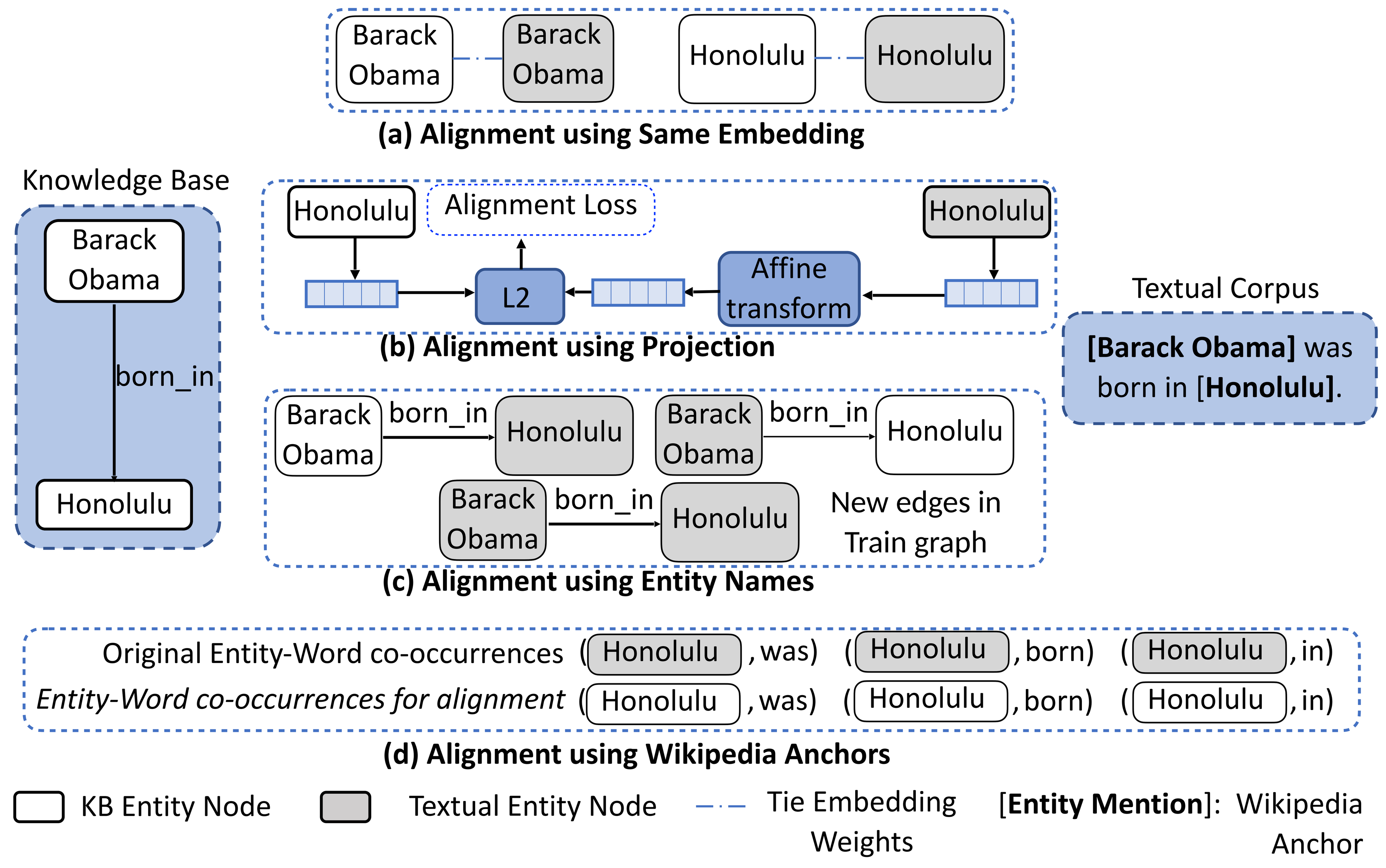
- First systematic investigation on KB-text embedding alignment at scale.
  - **Wikidata**: 14.6M entities, 1.2K relations, 261M facts
  - **Wikipedia**: 8.2M articles, 2.1M words, 12.3M entities
- Evaluation framework with two tasks:
  - **Few-shot link prediction**: text → KB
  - **Analogical reasoning**: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

Embedding Methods

Knowledge Base embedding model: TransE
Text embedding model: Skip-gram



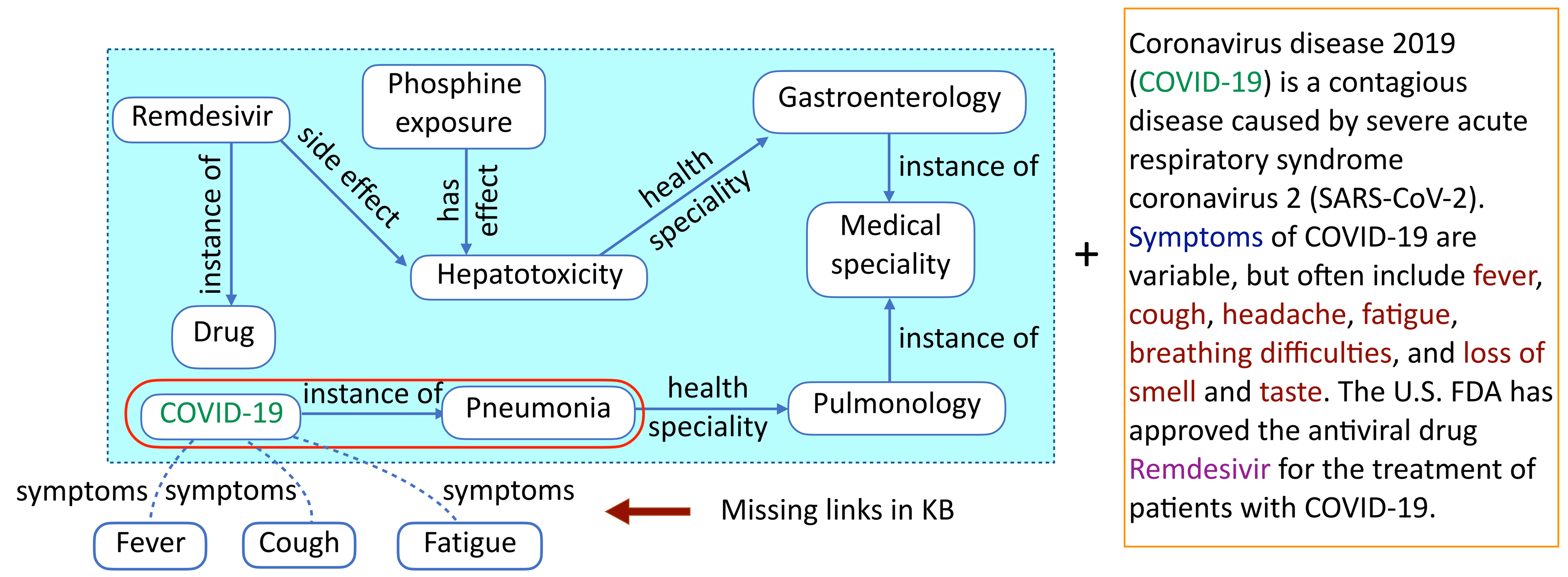
Alignment Methods



Evaluation Framework

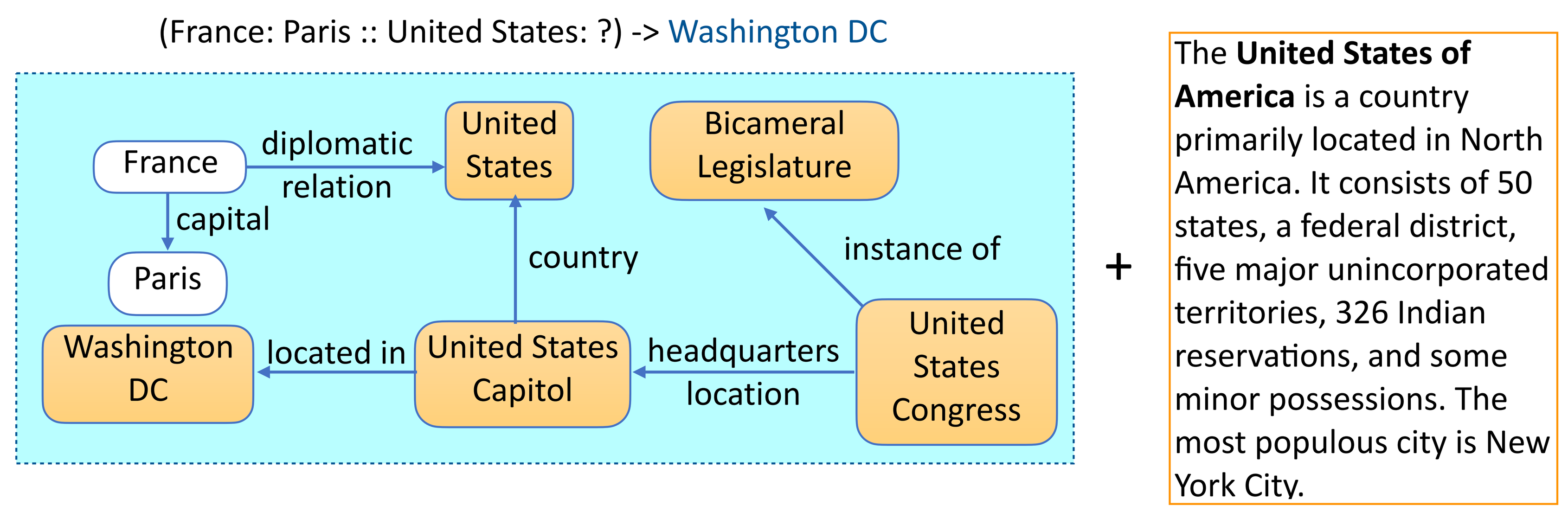
Few-shot link prediction

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



Analogical Reasoning

- Test the information flow from the knowledge-base embeddings to the skip-gram embeddings.
- $(h_1 : t_1) :: (h_2 : ?)$



Experiments

Overall Results

Model	Few-shot Link Prediction MR	Few-shot Link Prediction Hits@1	Few-shot Link Prediction Hits@10	Analogical Reasoning MR	Analogical Reasoning Hits@1	Analogical Reasoning Hits@10
TransE	187	20.3	40.4	–	–	–
Skip-gram	–	–	–	25	50.6	78.0
Projection	134	22.9	47.2	12	65.9	89.0
Same Embedding align.	**102**	**30.7**	**51.8**	11	60.7	87.5
Entity Name align.	116	23.1	46.7	**8**	**66.5**	**91.0**
Wikipedia Anchors align.	138	25.8	46.2	14	56.1	84.8

Table 1: Overall results for both evaluation tasks.

- Alignment methods significantly outperform the naive TransE and skip-gram baselines for few-shot link prediction and analogical reasoning respectively.
- Joint reasoning through alignment enhances both KB and text entity representations.
- The inductive bias of a particular alignment method can affect its performance on an evaluation task.

Case Study

- Knowledge base completion for COVID related relations using alignment models.
- 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.
- Alignment methods outperform the TransE baseline in a 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

Table 2: Link Prediction results for COVID-19 case study (Mean Rank).

Contact Information

Paper · Code

Author Contact: pahuja.9@osu.edu