CS11-747 Neural Networks for NLP

Learning From/For Knowledge Bases

Graham Neubig

Carnegie Mellon University
Language Technologies Institute

Site
https://phontron.com/class/nn4nlp2019/

Knowledge Bases

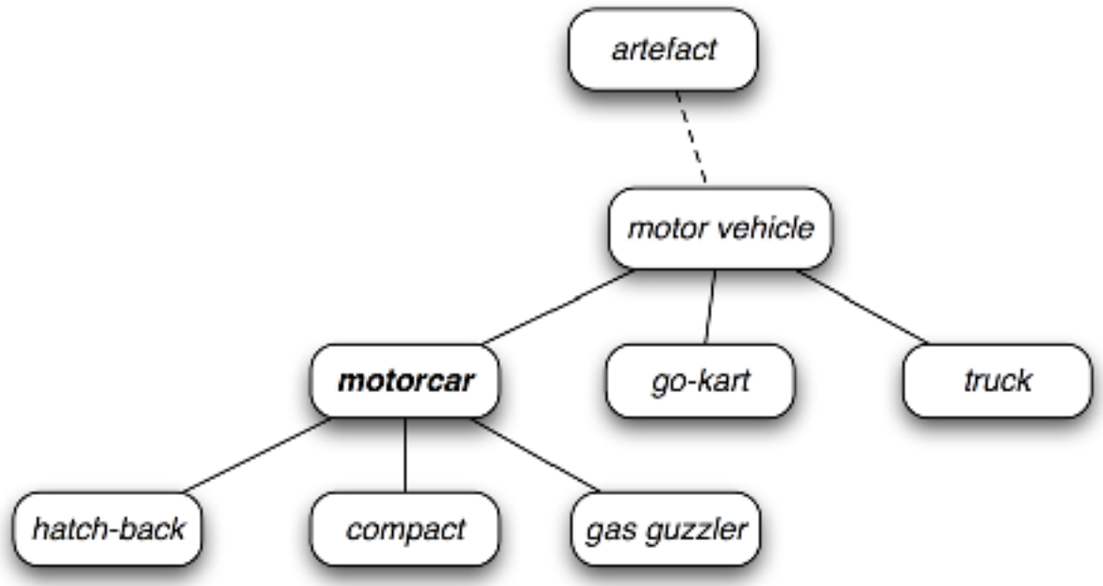
- Structured databases of knowledge usually containing
  - Entities (nodes in a graph)
  - Relations (edges between nodes)

- How can we **learn to create/expand knowledge bases** with neural networks?
- How can we **learn from the information in knowledge bases** to improve neural representations?
- How can we use structured knowledge to answer questions (see also semantic parsing class)

Types of Knowledge Bases

WordNet (Miller 1995)

- WordNet is a large database of words including parts of speech, semantic relations



- Nouns: is-a relation (hatch-back/car), part-of (wheel/car), type/instance distinction
- Verb relations: ordered by specificity (communicate -> talk -> whisper)
- Adjective relations: antonymy (wet/dry)

Image Credit: NLTK

Cyc (Lenant 1995)

- A manually curated database attempting to encode all common sense knowledge, 30 years in the making

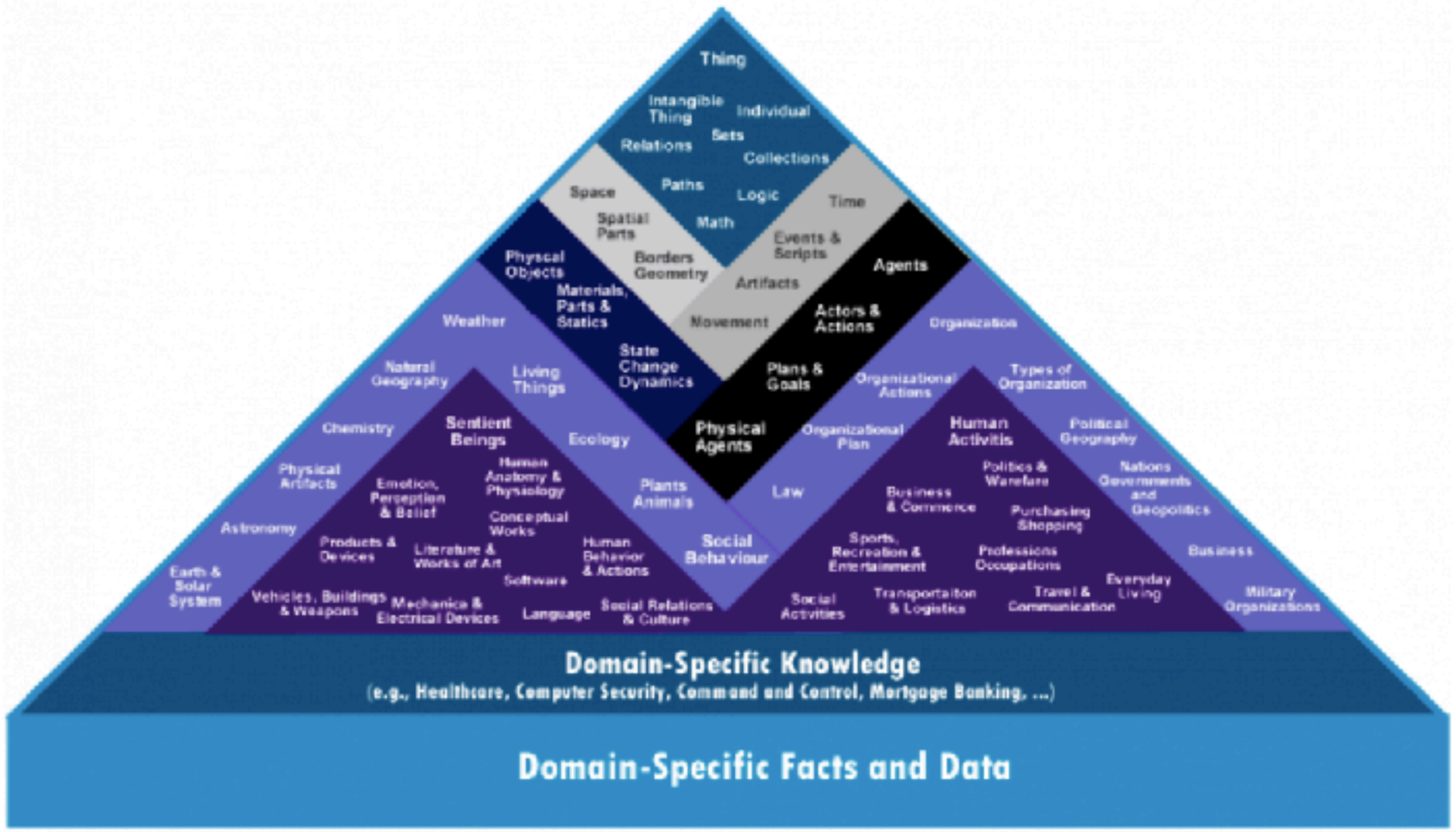
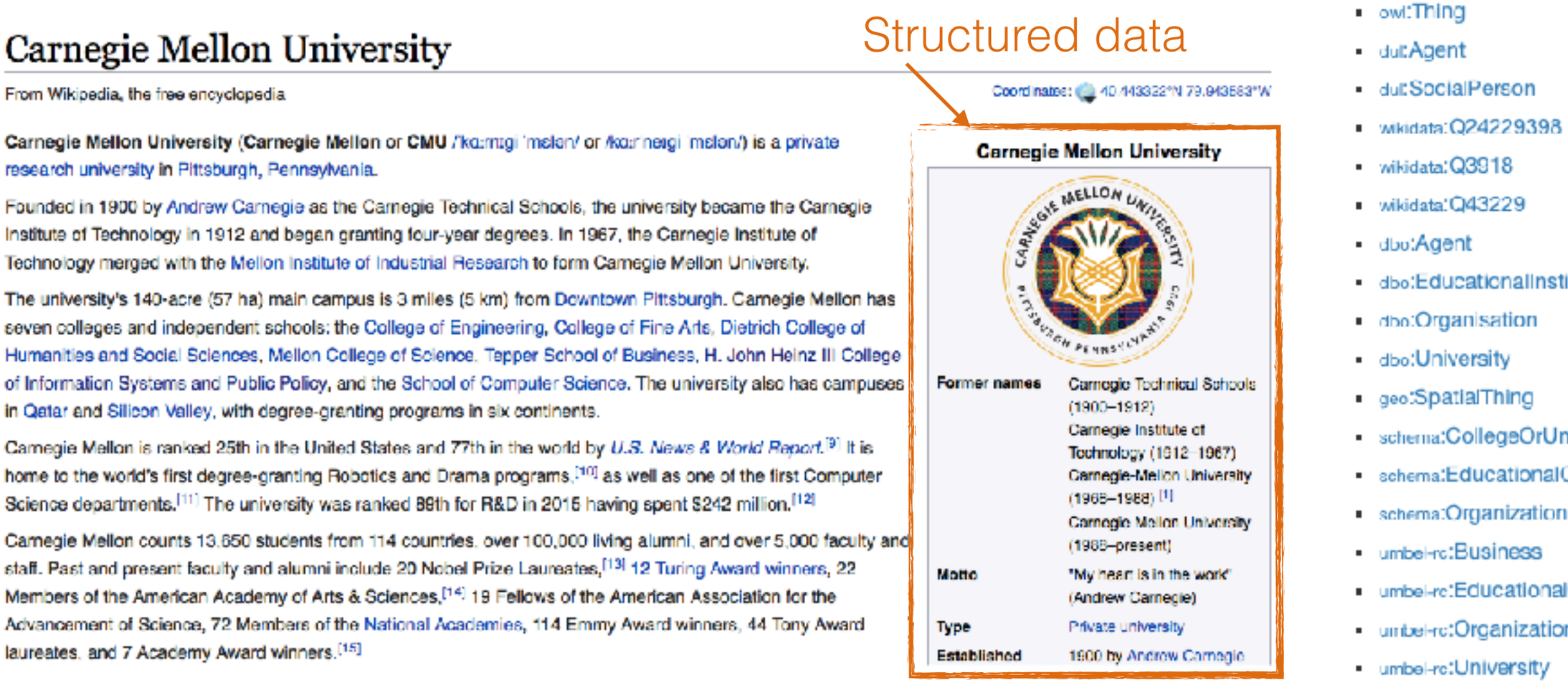


Image Credit: NLTK

DBPedia (Auer et al. 2007)

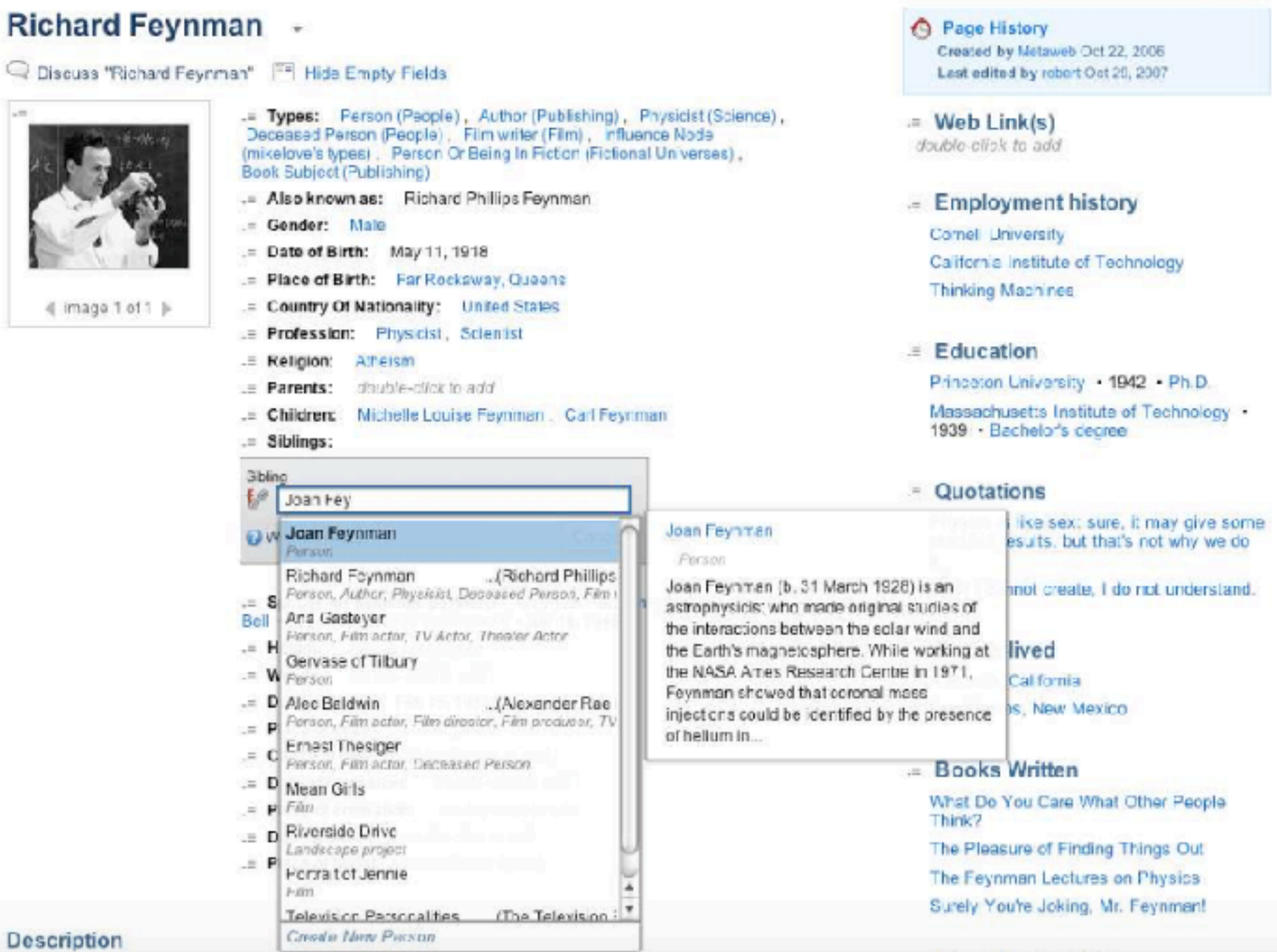
- Extraction of structured data from Wikipedia



Freebase/WikiData

(Bollacker et al. 2008)

- *Curated* database of entities, linked, and extremely large scale



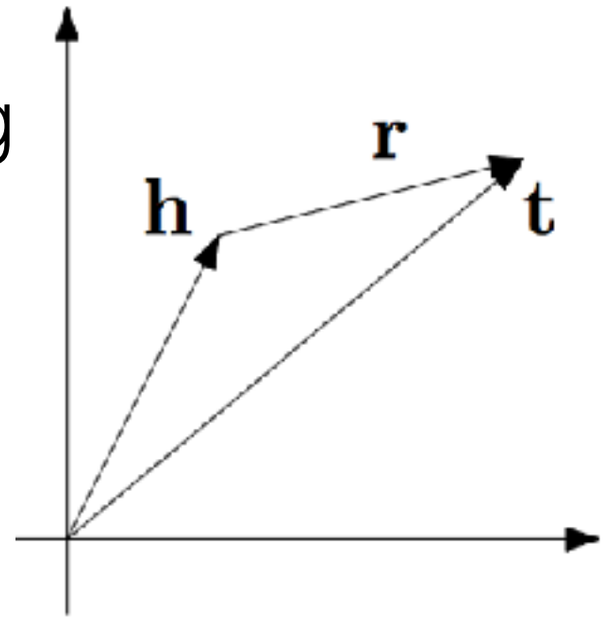
Learning Representations for Knowledge Bases

Learning Knowledge Graph Embeddings (Bordes et al. 2013)

- Motivation: express triples as additive transformation

- Method: minimize the distance of existing triples with a margin-based loss that

$$\sum_{(h,\ell,t)\in S} \sum_{(h',\ell,t')\in S'_{(h,\ell,t)}} \left[\gamma + d(\boldsymbol{h}+\boldsymbol{\ell},\boldsymbol{t}) - d(\boldsymbol{h}'+\boldsymbol{\ell},\boldsymbol{t}')\right]_+$$



(a) TransE

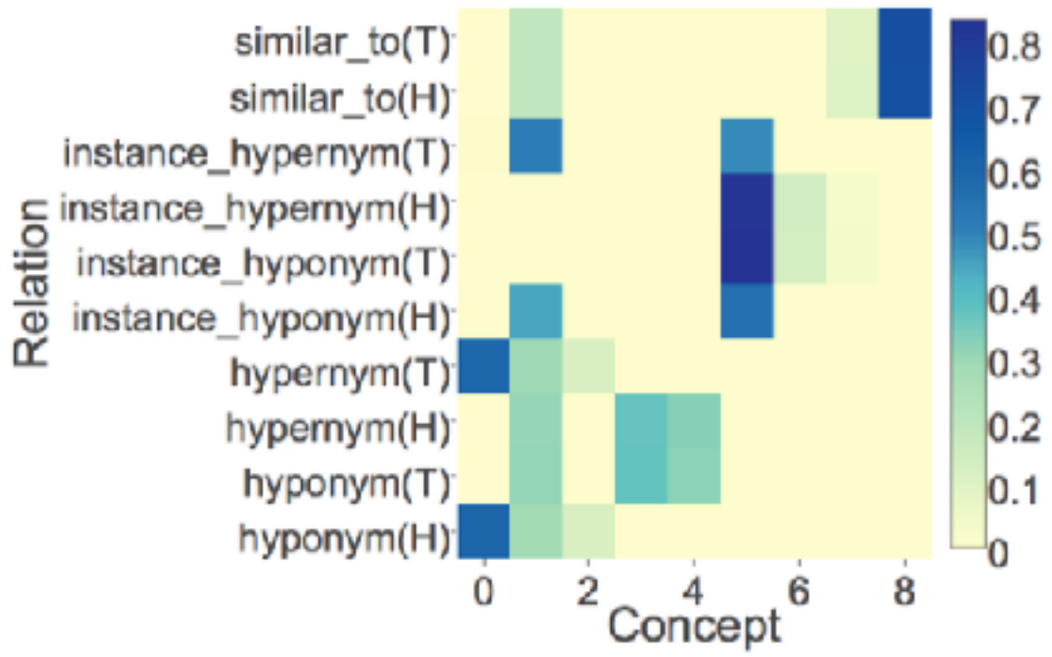
- Note: one vector for each relation, additive modification only, intentionally simpler than NTN

Decomposable Relation Model (Xie et al. 2017)

- Idea: There are many relations, but each can be represented by a limited number of “concepts”
- Method: Treat each relation map as a mixture of concepts, with sparse mixture vector α

$$f_r(h,t) = \|\boldsymbol{\alpha}_r^H \cdot \mathbf{D} \cdot \mathbf{h} + \mathbf{r} - \boldsymbol{\alpha}_r^T \cdot \mathbf{D} \cdot \mathbf{t}\|_\ell$$

- Better results, and also somewhat interpretable relations

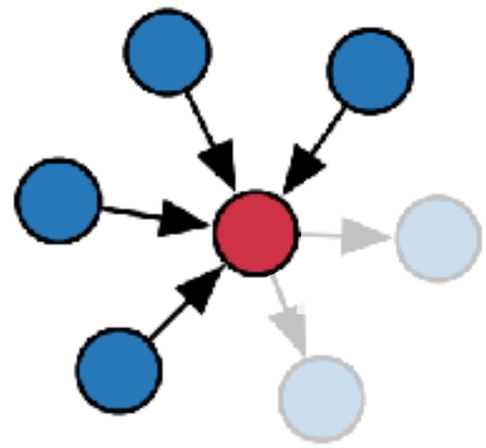


Multi-hop Relational Context w/ Graph Neural Networks (Schlichtkrull et al., 2017)

- Idea: consider all the local neighborhood entities instead of each triples

- Method: apply the message-passing framework using Graph Convolutional Network

$$h_i^{(l+1)} = \sigma\left(\sum_{r\in\mathcal{R}}\sum_{j\in\mathcal{N}_i^r}\frac{1}{c_{i,r}}W_r^{(l)}h_j^{(l)} + W_0^{(l)}h_i^{(l)}\right)$$



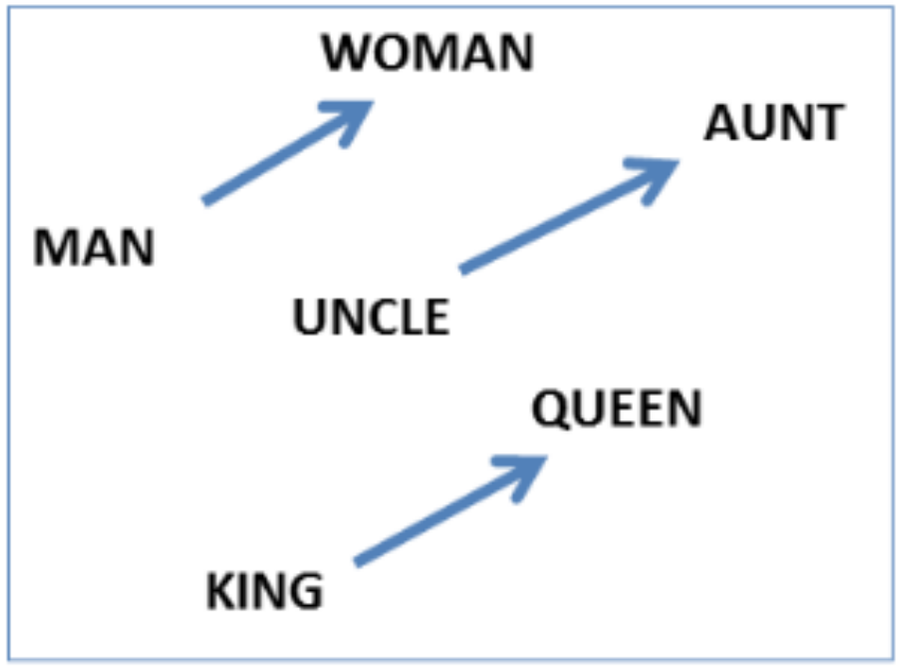
- Recurrent application will allow capturing K-hop neighbor nodes.

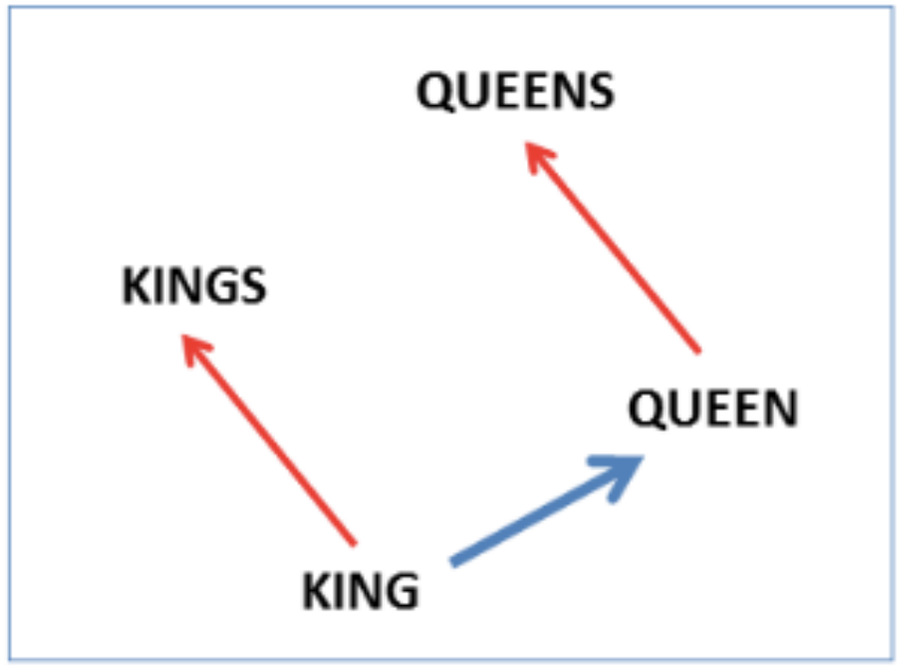
Knowledge Base Incompleteness

- Even w/ extremely large scale, knowledge bases are by nature incomplete
- e.g. in FreeBase 71% of humans were missing “date of birth” (West et al. 2014)
- Can we perform “relation extraction” to extract information for knowledge bases?

Remember: Consistency in Embeddings

e.g. king-man+woman = queen (Mikolov et al. 2013)





Relation Extraction w/ Neural Tensor Networks (Socher et al. 2013)

- A first attempt at predicting relations: a multi-layer perceptron that predicts whether a relation exists

$$u_R^T f\left(W_{R,1} e_1 + W_{R,2} e_2\right)$$

- Neural Tensor Network: Adds bi-linear feature extractors, equivalent to projections in space

$$g(e_1, R, e_2) = u_R^T f\left(e_1^T W_R^{[1:k]} e_2 + V_R \begin{bmatrix} e_1 \\ e_2 \end{bmatrix} + b_R\right)$$

- Powerful model, but perhaps overparameterized!

Learning from Text Directly

Distant Supervision for Relation Extraction (Mintz et al. 2009)

- Given an entity-relation-entity triple, extract all text that matches this and use it to train

[Steven Spielberg]'s film [Saving Private Ryan] is loosely based on the brothers' story.

Allison co-produced the Academy Award-winning [Saving Private Ryan], directed by [Steven Spielberg]...

- Creates a large corpus of (noisily) labeled text to train a system

Relation Classification w/ CNNs (Zeng et al. 2014)

- Extract features w/o syntax using CNN
  - Lexical features of the words themselves
  - Features of the whole span extracted using convolution

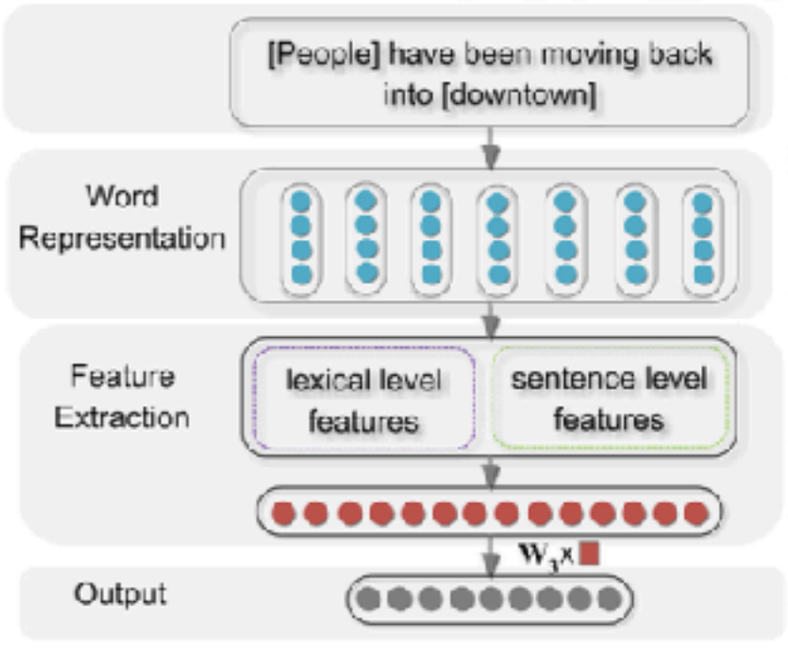


Figure 1: Architecture of the neural network used for relation classification.

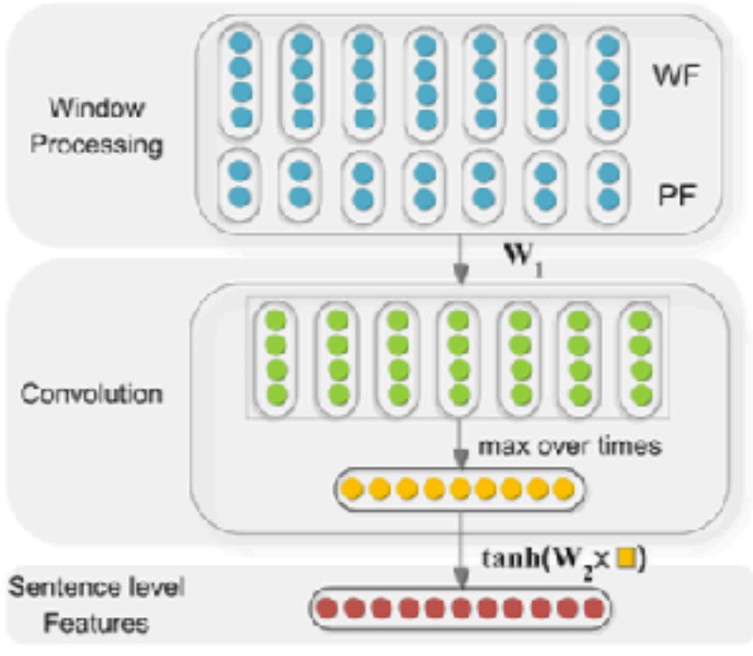


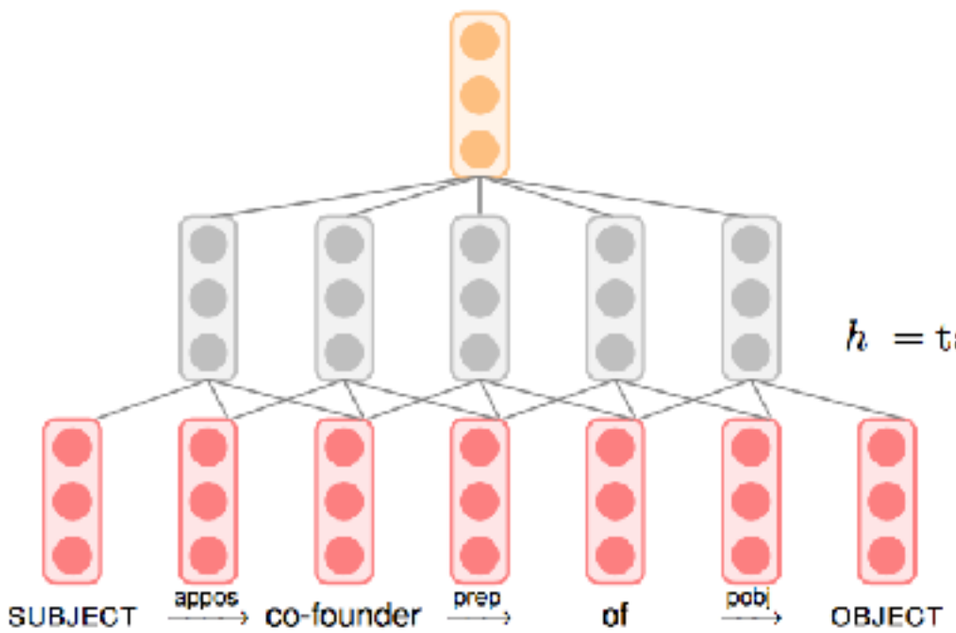
Figure 2: The framework used for extracting sentence level features.

Jointly Modeling KB Relations and Text (Toutanova et al. 2015)

- To model textual links between words w/ neural net: aggregate over multiple instances of links in dependency tree

Textual Pattern	Count
SUBJECT $\xrightarrow{\text{appos}}$ founder $\xrightarrow{\text{prep}}$ of $\xrightarrow{\text{pobj}}$ OBJECT	12
SUBJECT $\xleftarrow{\text{nsubj}}$ co-founded $\xrightarrow{\text{dobj}}$ OBJECT	3
SUBJECT $\xrightarrow{\text{appos}}$ co-founder $\xrightarrow{\text{prep}}$ of $\xrightarrow{\text{pobj}}$ OBJECT	3
SUBJECT $\xrightarrow{\text{conj}}$ co-founder $\xrightarrow{\text{prep}}$ of $\xrightarrow{\text{pobj}}$ OBJECT	3

- Model relations w/ CNN



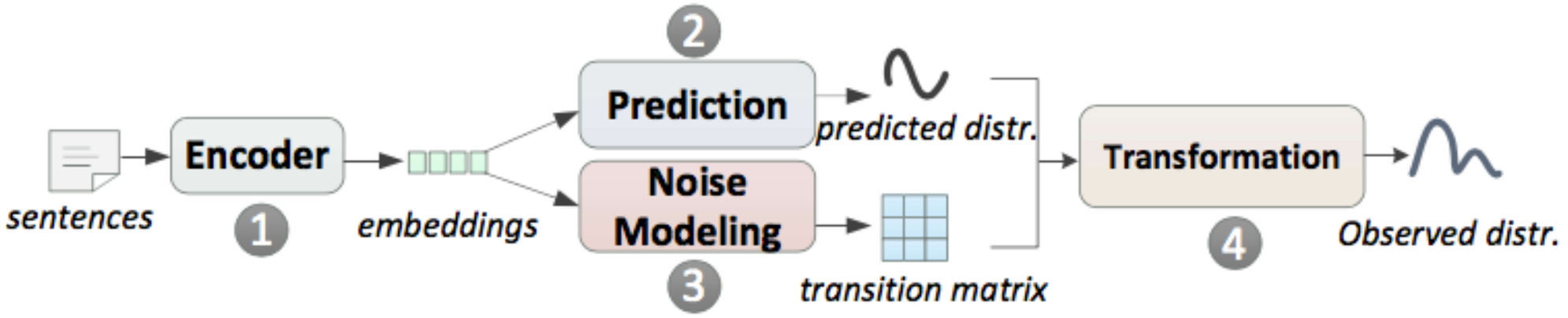
$$r = \max\{h\}$$

$$h = \tanh(W^{-1}v_{-1} + W^{0}v + W^{1}v_{+1} + b)$$

$$v = Ve$$

Modeling Distant Supervision Noise in Neural Models (Luo et al. 2017)

- Idea: there is noise in distant supervision labels, so we want to model it



- By controlling the “transition matrix”, we can adjust to the amount of noise expected in the data
  - Trace normalization to try to make matrix close to identity
  - Start training w/ no transition matrix on data expected to be clean, then phase in on full data

Using Knowledge Bases to Inform Neural Models

Retrofitting of Embeddings to Existing Lexicons (Faruqui et al. 2015)

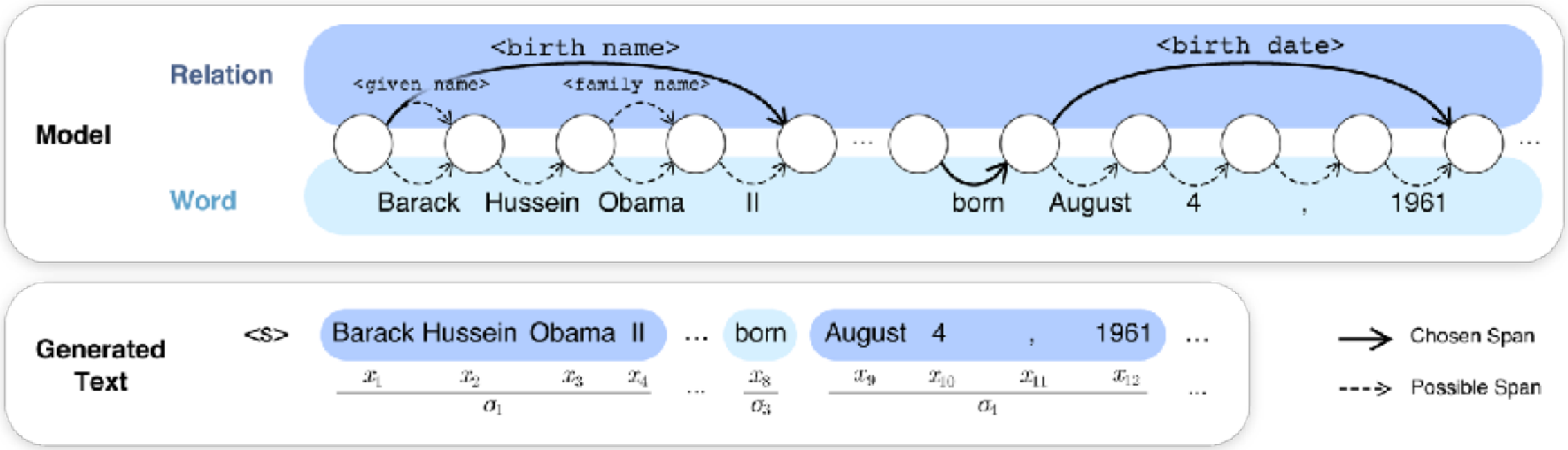
- Similar to joint learning, but done through post-hoc transformation of embeddings
    - Advantage of being usable with any pre-trained embeddings
- Double objective of making transformed embeddings close to neighbors, and close to original embedding

$$\Psi(Q) = \sum_{i=1}^{n} \left[\alpha_i \|q_i - \hat{q}_i\|^2 + \sum_{(i,j) \in E} \beta_{ij} \|q_i - q_j\|^2 \right]$$

- Can also force antonyms away from each-other (Mrksic et al. 2016)

Injecting Knowledge into Language Models (Hayashi et al. 2020)

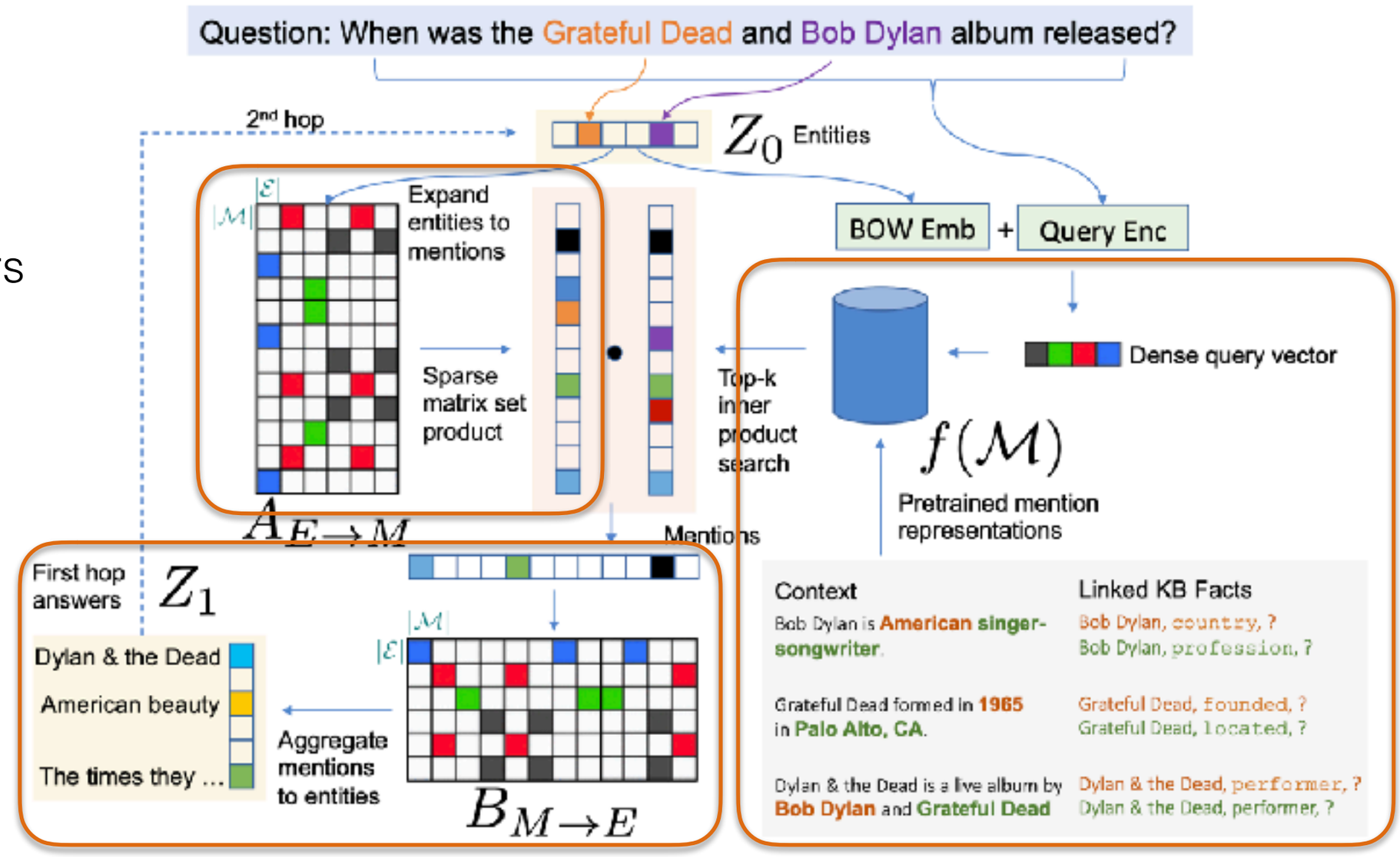
- Provide LMs with topical knowledge in the form of copiable graphs
    - Each (Wiki) text is given relevant KB taken from Wikidata
- Examine all possible decoding "paths" and maximize the marginal probability



Reasoning over Text Corpus as a Knowledge Base (Dhingra et al. 2020)

- Answering questions using text corpora as a traceable KB
- Relevance matching over **mentions**

1. Create mention vectors
2. Retrieve relevant mentions from pre-trained models
3. Aggregate scores



Schema-Free Extraction

Open Information Extraction
(Banko et al 2007)

- Basic idea: **the text is the relation**

- e.g. "United has a hub in Chicago, which is the headquarters of United Continental Holdings"
    - {United; has a hub in; Chicago}
    - {Chicago; is the headquarters of; United Continental Holdings}

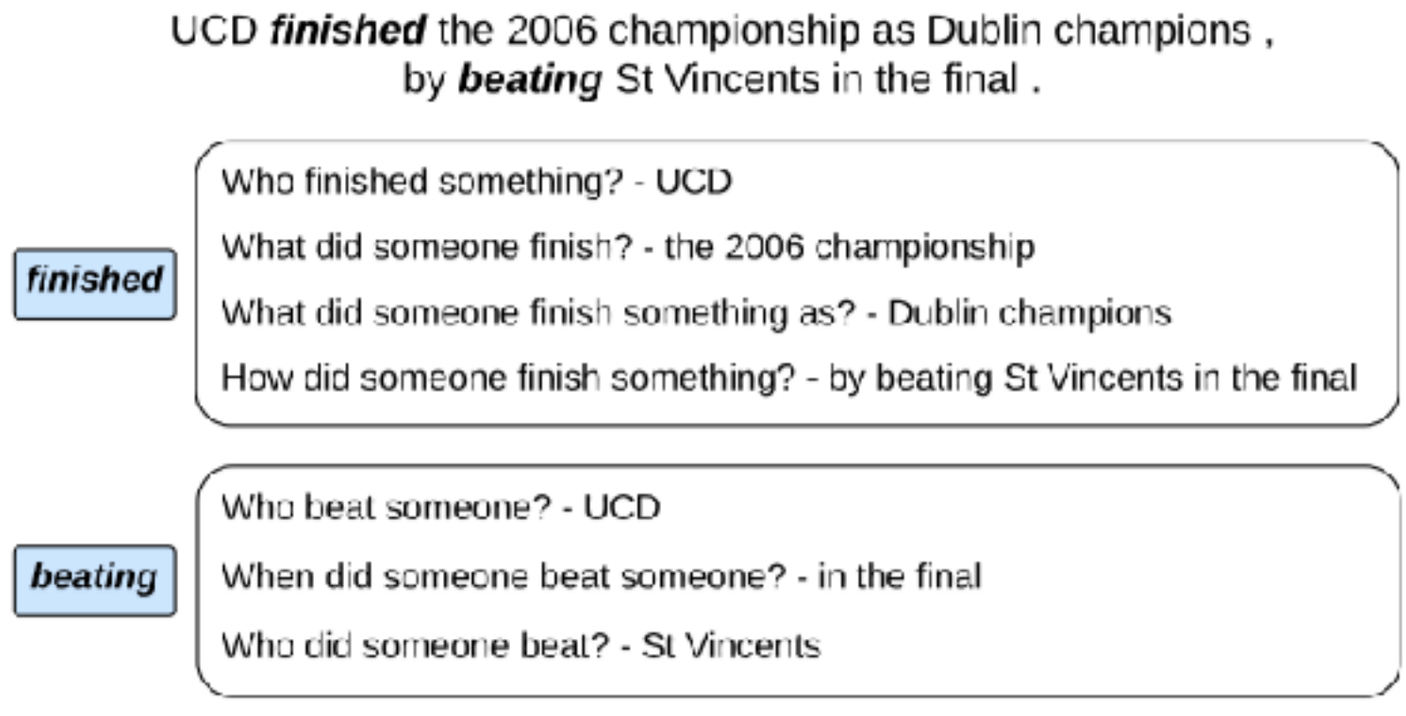
- Can extract any variety of relations, but does not abstract

Rule-based Open IE

- e.g. TextRunner (Banko et al. 2007), ReVerb (Fader et al. 2011)
- Use parser to extract according to rules
    - e.g. relation must contain a predicate, subject object must be noun phrases, etc.
- Train a fast model to extract over large amounts of data
- Aggregate multiple pieces of evidence (heuristically) to find common, and therefore potentially reliable, extractions

Neural Models for Open IE

- Unfortunately, heuristics are still not perfect
- Possible to create relatively large datasets by asking simple questions (He et al. 2015):



- Can be converted into OpenIE extractions, for use in supervised neural BIO tagger (Stanovsky et al. 2018)

Learning Relations from Relations

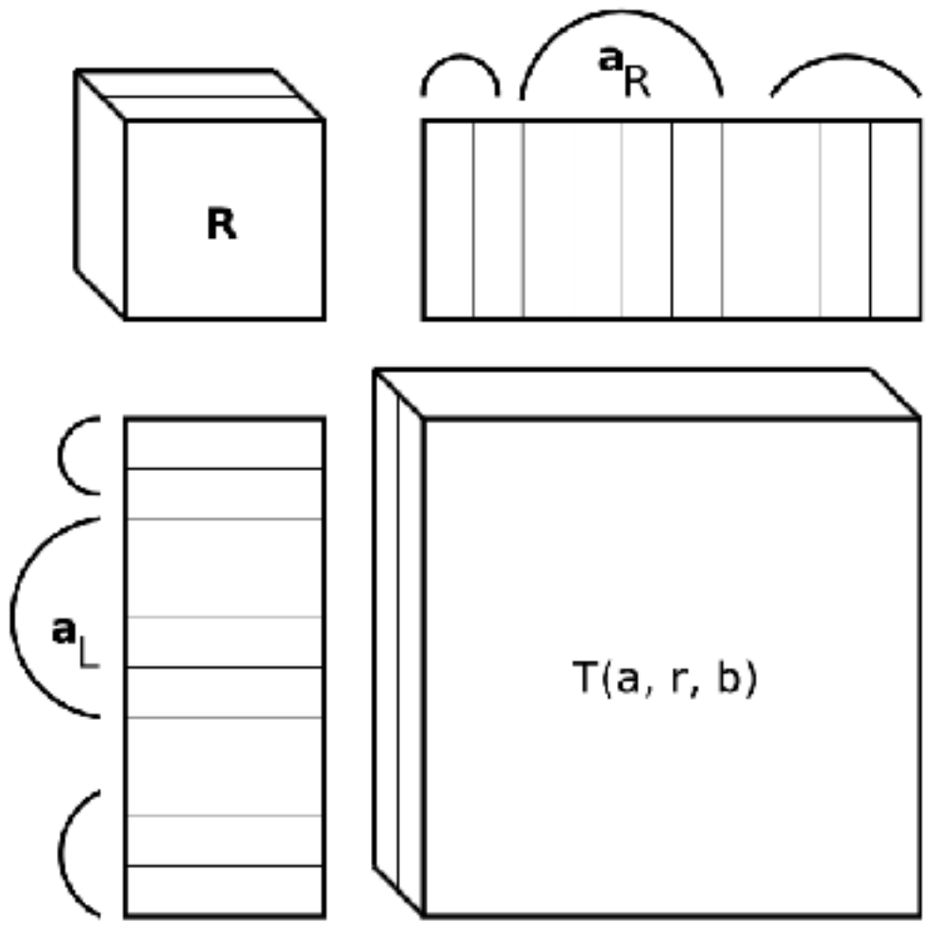
Modeling Word Embeddings vs. Modeling Relations

- Word embeddings give information of the word in context, which is indicative of KB traits
- However, other relations (or combinations thereof) are also indicative
  - This is a *link prediction* problem in graphs

Tensor Decomposition

(Sutskever et al. 2009)

- Can model relations by decomposing a tensor containing entity/relation/entity tuples

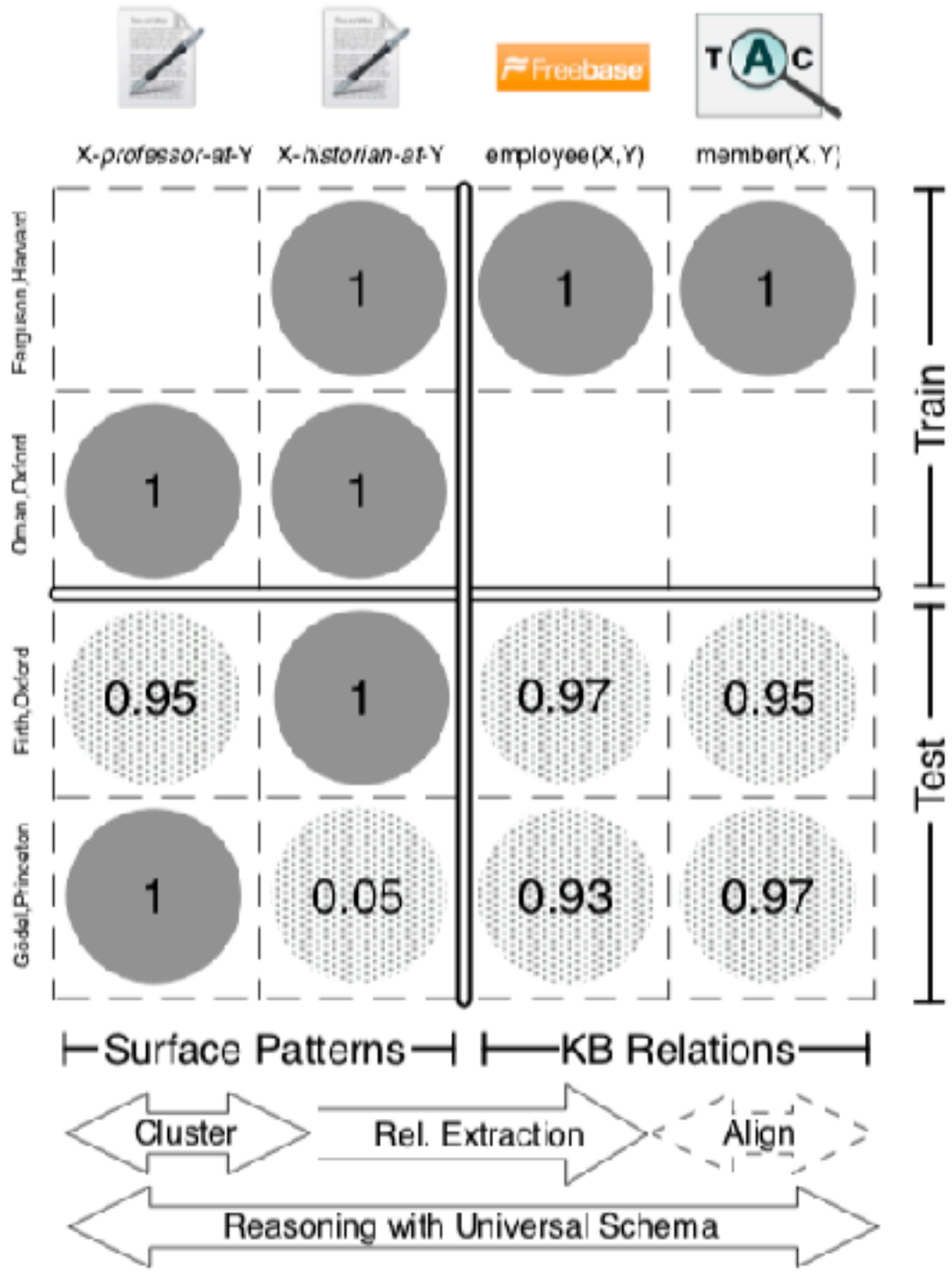


Matrix Factorization to Reconcile Schema-based and Open IE Extractions

(Riedel et al. 2013)

- What to do when we have a knowledge base, and text from OpenIE extractions?

- **Universal schema:** embed relations from multiple schema in the same space



Modeling Relation Paths
(Lao and Cohen 2010)

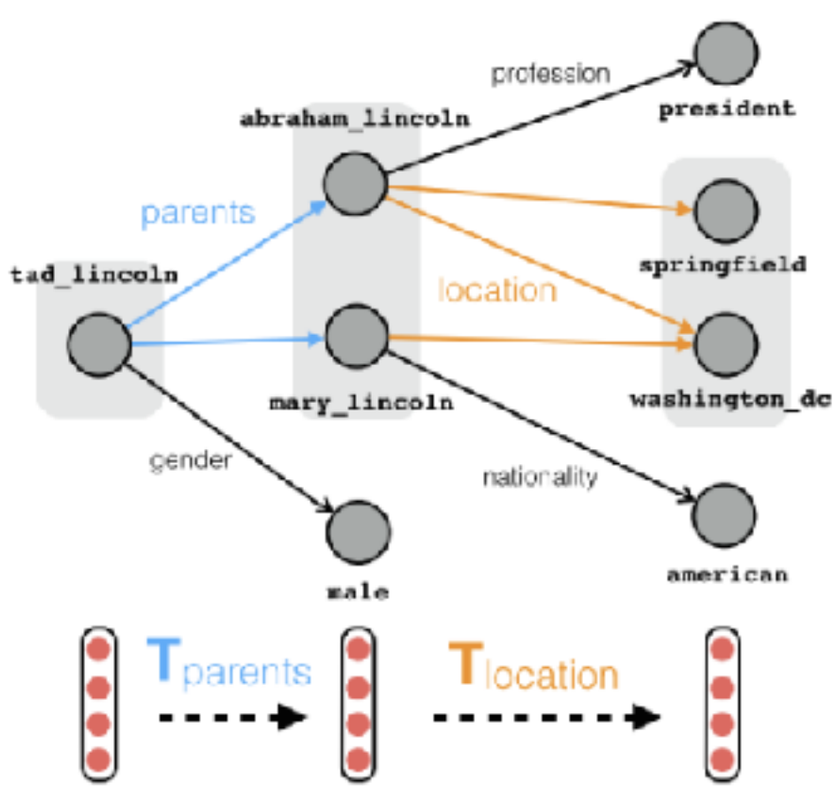
- Multi-step paths can be informative for indicating individual relations

- e.g. “given word, recommend venue in which to publish the paper”

ID	Weight	Feature
1	26.9	$word \xrightarrow{HasTitle^{-1}} paper \xrightarrow{In} journal$
2	4.5	$word \xrightarrow{HasTitle^{-1}} paper \xrightarrow{FirstAuthor} author \xrightarrow{FirstAuthor^{-1}} paper \xrightarrow{In} journal$
3	2.8	$word \xrightarrow{HasTitle^{-1}} paper \xrightarrow{AnyAuthor} author \xrightarrow{AnyAuthor^{-1}} paper \xrightarrow{In} journal$
4	1.1	$gene \xrightarrow{GeneticallyRelated} gene \xrightarrow{HasGene^{-1}} paper \xrightarrow{In} journal$
5	0.9	$gene \xrightarrow{HasGene^{-1}} paper \xrightarrow{In} journal$
6	0.6	$e^* \xrightarrow{AnyPaper} paper \xrightarrow{Cite} paper \xrightarrow{In} journal$

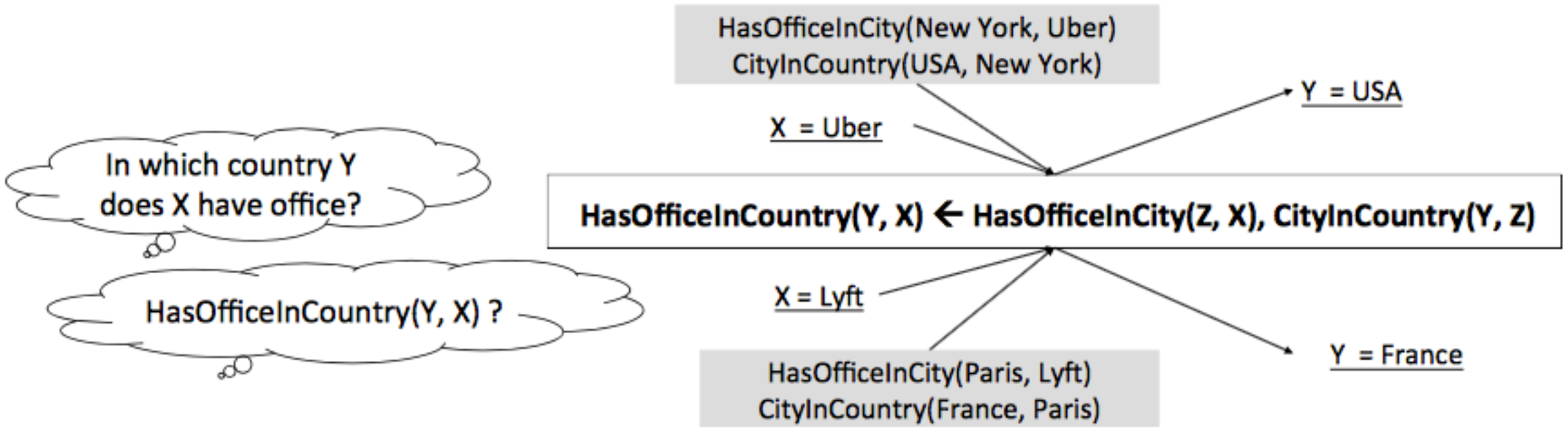
Optimizing Relation Embeddings over Paths (Guu et al. 2015)

- Traveling over relations might result in error propagation
- Simple idea: optimize so that after traveling along a path, we still get the correct entity



Differentiable Logic Rules
(Yang et al. 2017)

- Consider whole paths in a differentiable framework



- Treat path as a sequence of matrix multiplies, where the rule weight is α

$$\sum_{l} \alpha_l \Pi_{\mathbf{k} \in \beta_l} \mathbf{M}_{\mathbf{R_k}}$$

Questions?