CS11-711 Advanced NLP

Learning From/For Knowledge Bases

Graham Neubig

Carnegie Mellon University

Language Technologies Institute

https://phontron.com/class/anlp-fall2024/

Some slides by Zhengbao Jiang

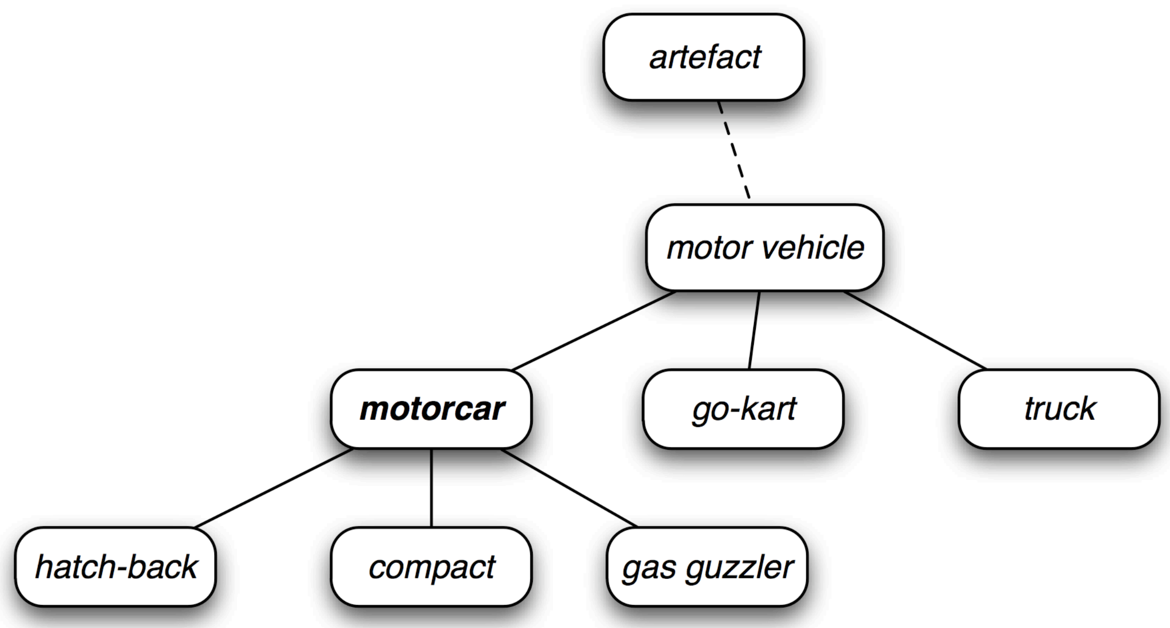
Knowledge Bases

- 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

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

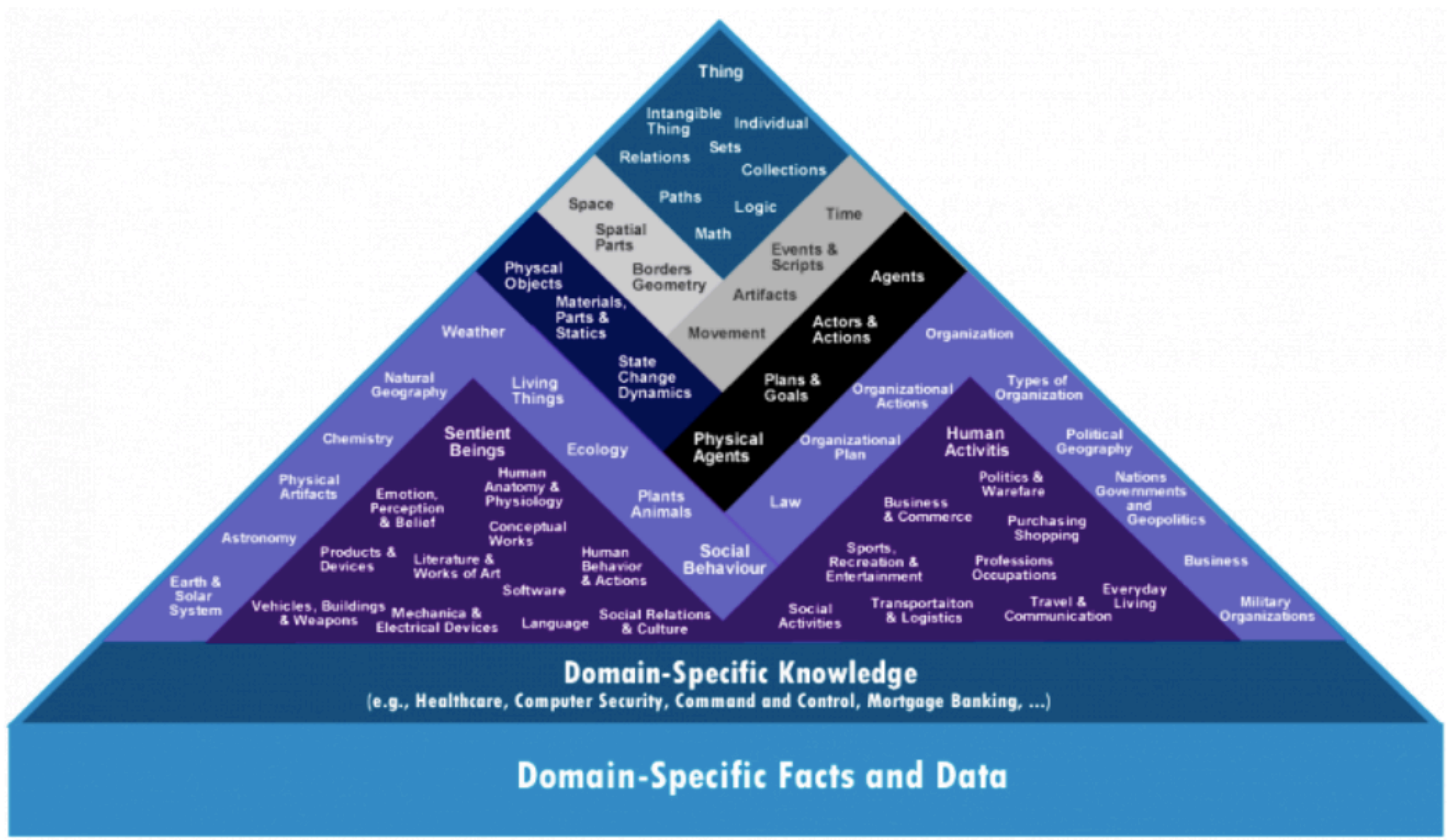
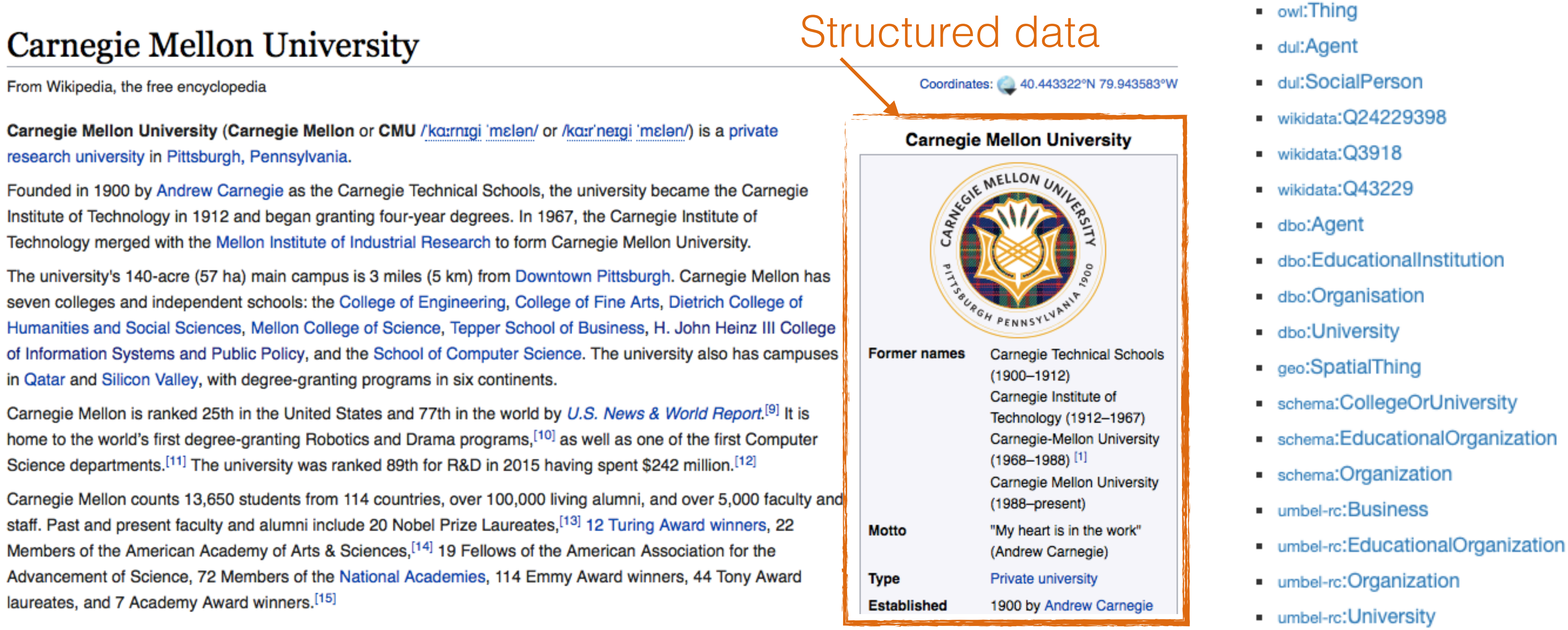


Image Credit: NLTK

DBPedia (Auer et al. 2007)

- Extraction of structured data from Wikipedia



Structured data

Carnegie Mellon University

From Wikipedia, the free encyclopedia

Coordinates: 40.443322°N 79.943583°W

Carnegie Mellon University (**Carnegie Mellon** or **CMU** /ˈkɑːrnɪɡi ˈmɛlən/ or /kɑːrˈneɪɡi ˈmɛlən/) is a private research university in Pittsburgh, Pennsylvania.

Founded in 1900 by Andrew Carnegie as the Carnegie Technical Schools, the university became the Carnegie Institute of Technology in 1912 and began granting four-year degrees. In 1967, the Carnegie Institute of Technology merged with the Mellon Institute of Industrial Research to form Carnegie Mellon University.

The university's 140-acre (57 ha) main campus is 3 miles (5 km) from Downtown Pittsburgh. Carnegie Mellon has seven colleges and independent schools: the College of Engineering, College of Fine Arts, Dietrich College of Humanities and Social Sciences, Mellon College of Science, Tepper School of Business, H. John Heinz III College of Information Systems and Public Policy, and the School of Computer Science. The university also has campuses in Qatar and Silicon Valley, with degree-granting programs in six continents.

Carnegie Mellon is ranked 25th in the United States and 77th in the world by *U.S. News & World Report*.[9] It is home to the world's first degree-granting Robotics and Drama programs,[10] as well as one of the first Computer Science departments.[11] The university was ranked 89th for R&D in 2015 having spent $242 million.[12]

Carnegie Mellon counts 13,650 students from 114 countries, over 100,000 living alumni, and over 5,000 faculty and staff. Past and present faculty and alumni include 20 Nobel Prize Laureates,[13] 12 Turing Award winners, 22 Members of the American Academy of Arts & Sciences,[14] 19 Fellows of the American Association for the Advancement of Science, 72 Members of the National Academies, 114 Emmy Award winners, 44 Tony Award laureates, and 7 Academy Award winners.[15]

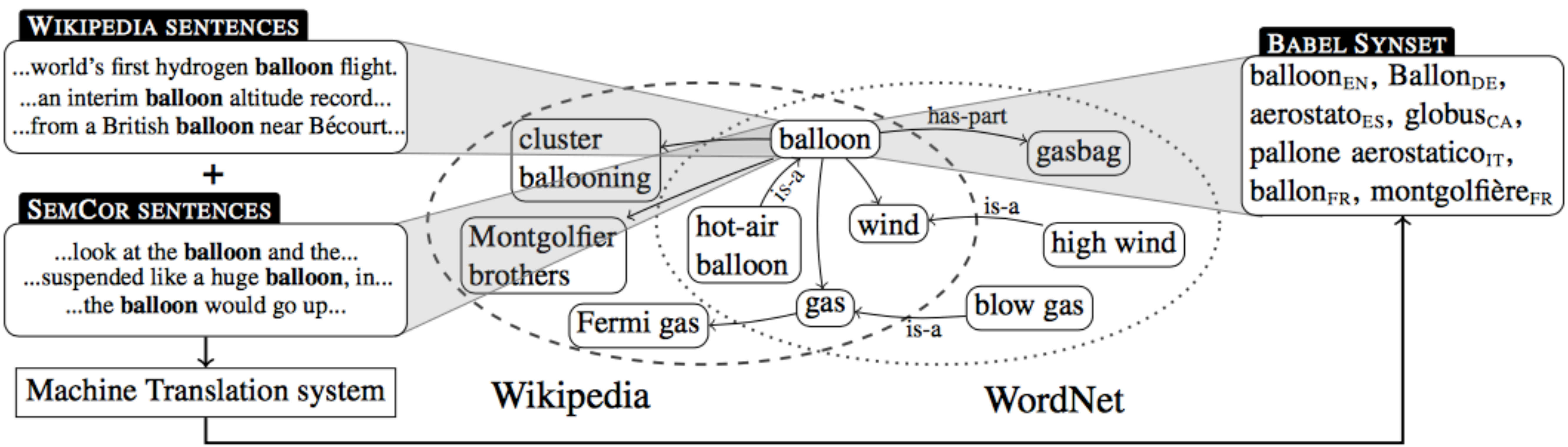
Carnegie Mellon University

Former names	Carnegie Technical Schools (1900–1912) Carnegie Institute of Technology (1912–1967) Carnegie-Mellon University (1968–1988) [1] Carnegie Mellon University (1988–present)
Motto	"My heart is in the work" (Andrew Carnegie)
Type	Private university
Established	1900 by Andrew Carnegie

- owl:Thing
- dul:Agent
- dul:SocialPerson
- wikidata:Q24229398
- wikidata:Q3918
- wikidata:Q43229
- dbo:Agent
- dbo:EducationalInstitution
- dbo:Organisation
- dbo:University
- geo:SpatialThing
- schema:CollegeOrUniversity
- schema:EducationalOrganization
- schema:Organization
- umbel-rc:Business
- umbel-rc:EducationalOrganization
- umbel-rc:Organization
- umbel-rc:University

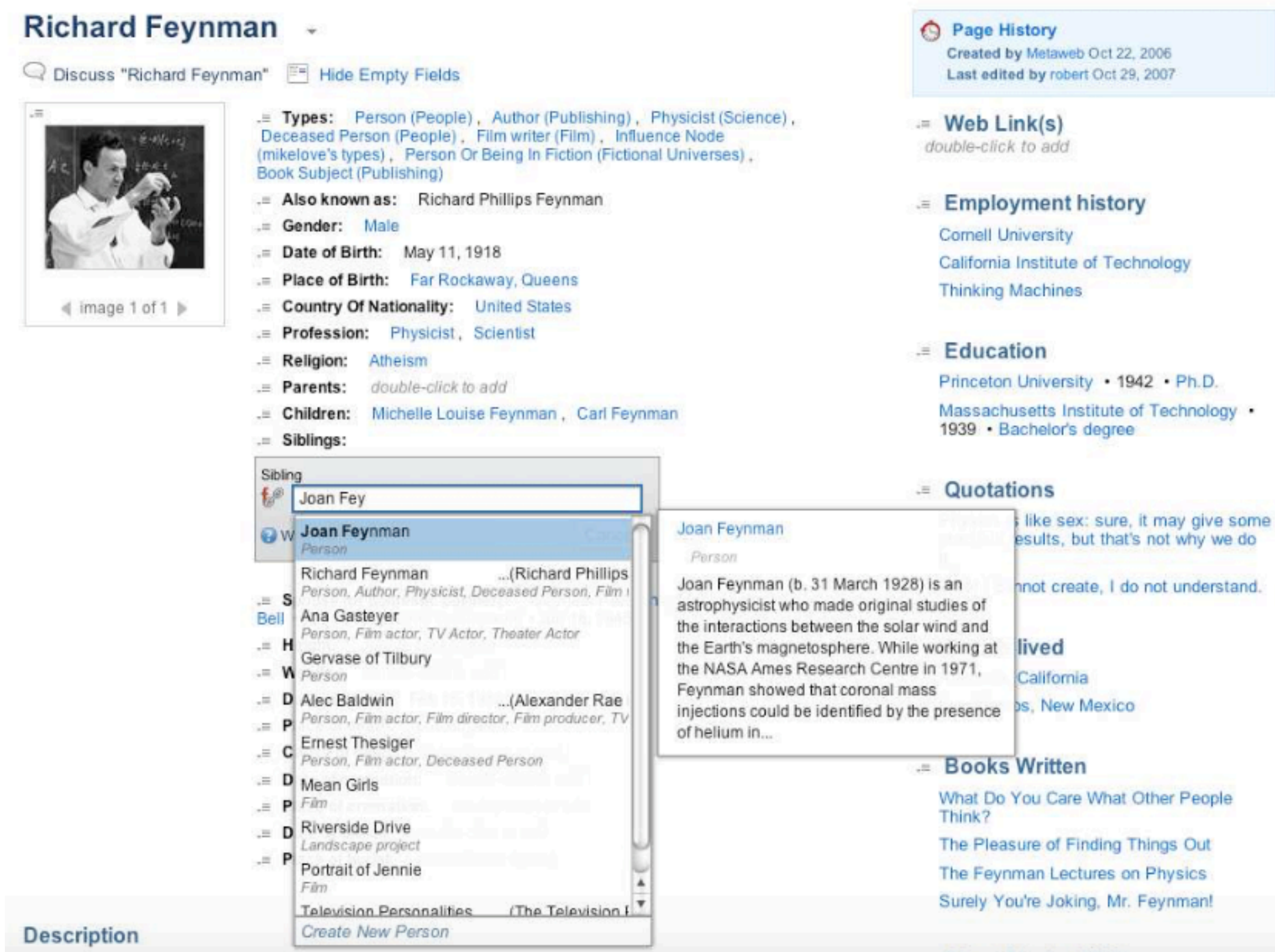
BabelNet
(Navigli and Ponzetto 2008)

- A meta-database including various sources such as WordNet and Wikipedia, but augmented with multi-lingual information



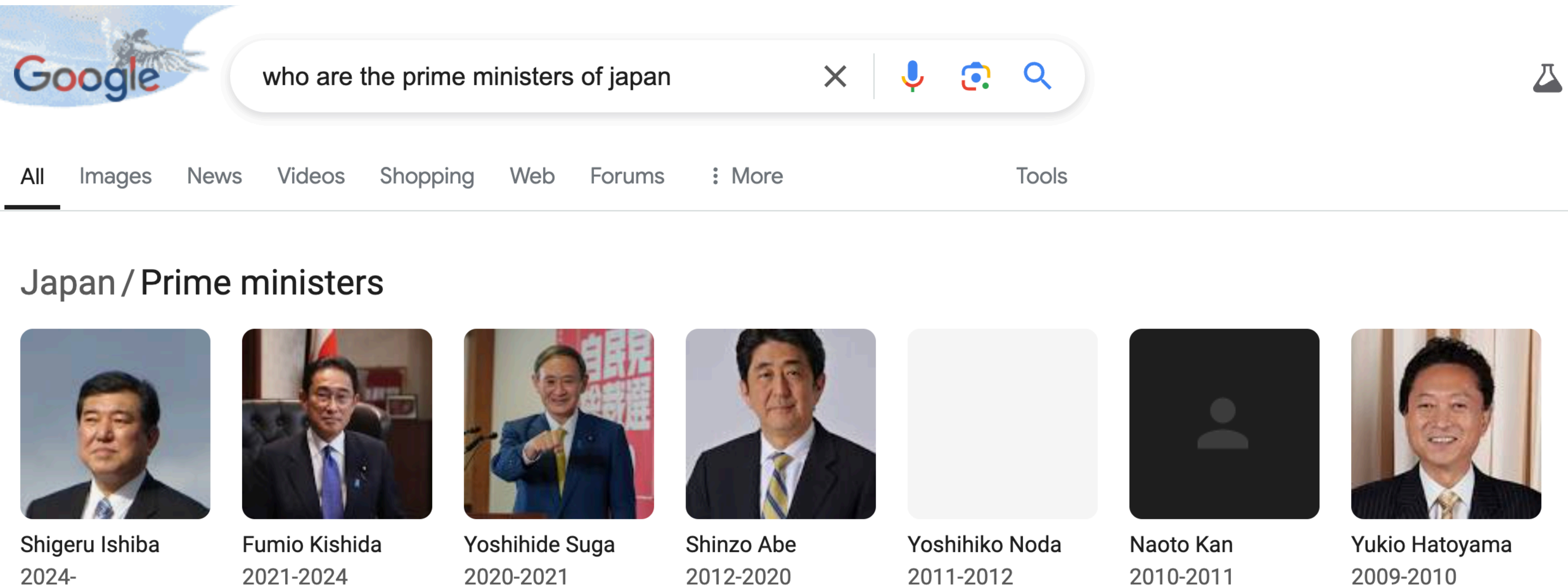
WikiData (Bollacker et al. 2008)

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



Querying Structured Data

Answering Questions w/ Knowledge Bases



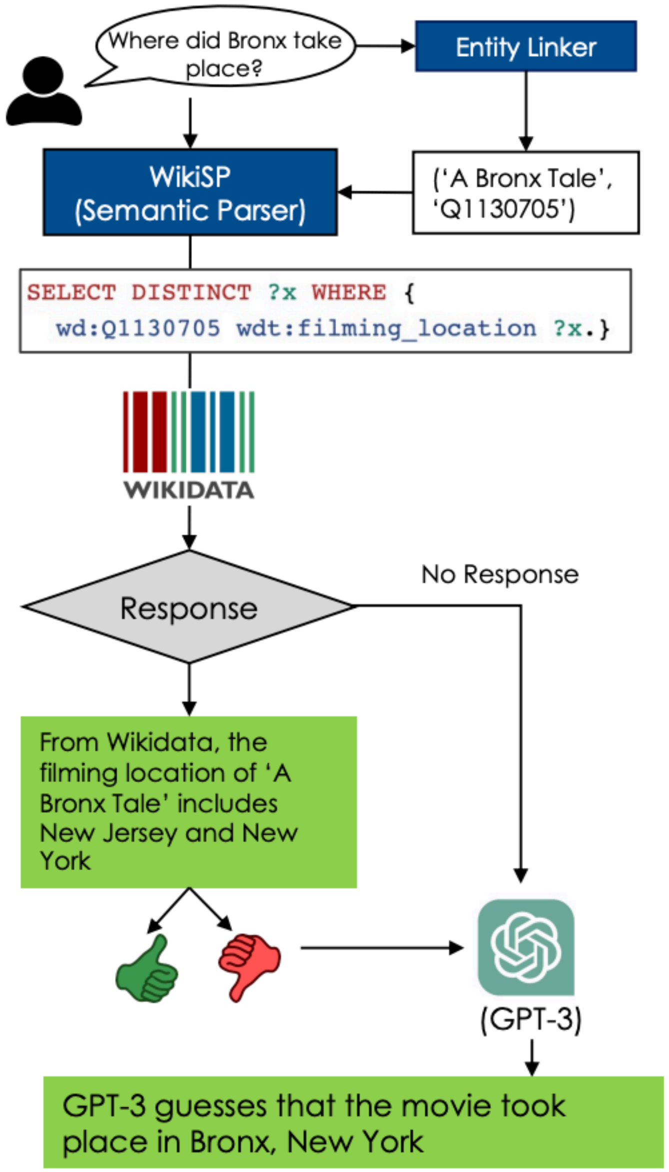
SPARQL: KB Query Language

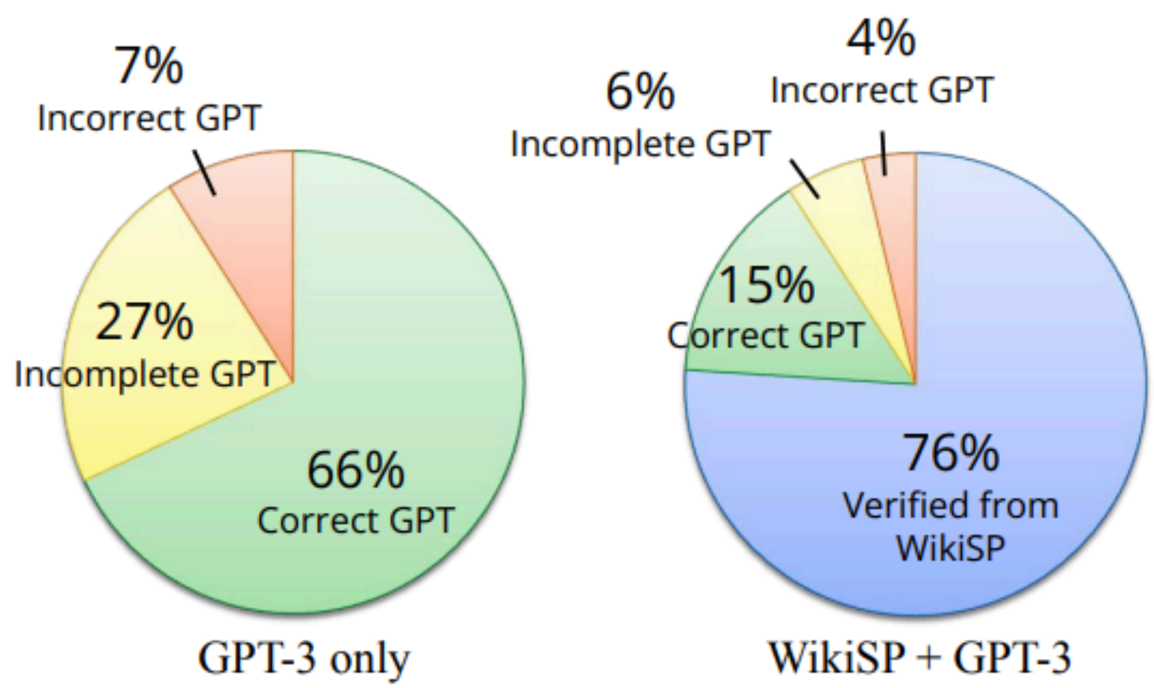
```
PREFIX wd: <http://www.wikidata.org/entity/>
PREFIX wdt: <http://www.wikidata.org/prop/direct/>
PREFIX p: <http://www.wikidata.org/prop/>
PREFIX ps: <http://www.wikidata.org/prop/statement/>
PREFIX pq: <http://www.wikidata.org/prop/qualifier/>

SELECT DISTINCT ?pmLabel ?startDate ?endDate
WHERE {
  ?pm p:P39 ?position .        # position held statement
  ?position ps:P39 wd:Q274948 . # position is Prime Minister of Japan

  OPTIONAL {
    ?position pq:P580 ?startDate .  # start date qualifier
  }

  OPTIONAL {
    ?position pq:P582 ?endDate .    # end date qualifier
  }

  SERVICE wikibase:label {
    bd:serviceParam wikibase:language "en" .
    ?pm rdfs:label ?pmLabel .
  }
}
ORDER BY DESC(?startDate)
```

pmLabel	startDate
Shigeru Ishiba	1 October 2024
Fumio Kishida	13 September 2023
Fumio Kishida	10 August 2022
Fumio Kishida	10 November 2021
Fumio Kishida	4 October 2021
Yoshihide Suga	16 September 2020
Shinzō Abe	1 November 2017
Shinzō Abe	24 December 2014
Shinzō Abe	26 December 2012
Yoshihiko Noda	2 September 2011
Naoto Kan	8 June 2010
Yukio Hatoyama	16 September 2009

Using Structured Data w/ Language Models

- e.g. WikiSP (Xu et al. 2023)





Probing Factual Knowledge in LMs

Probing Knowledge in LMs

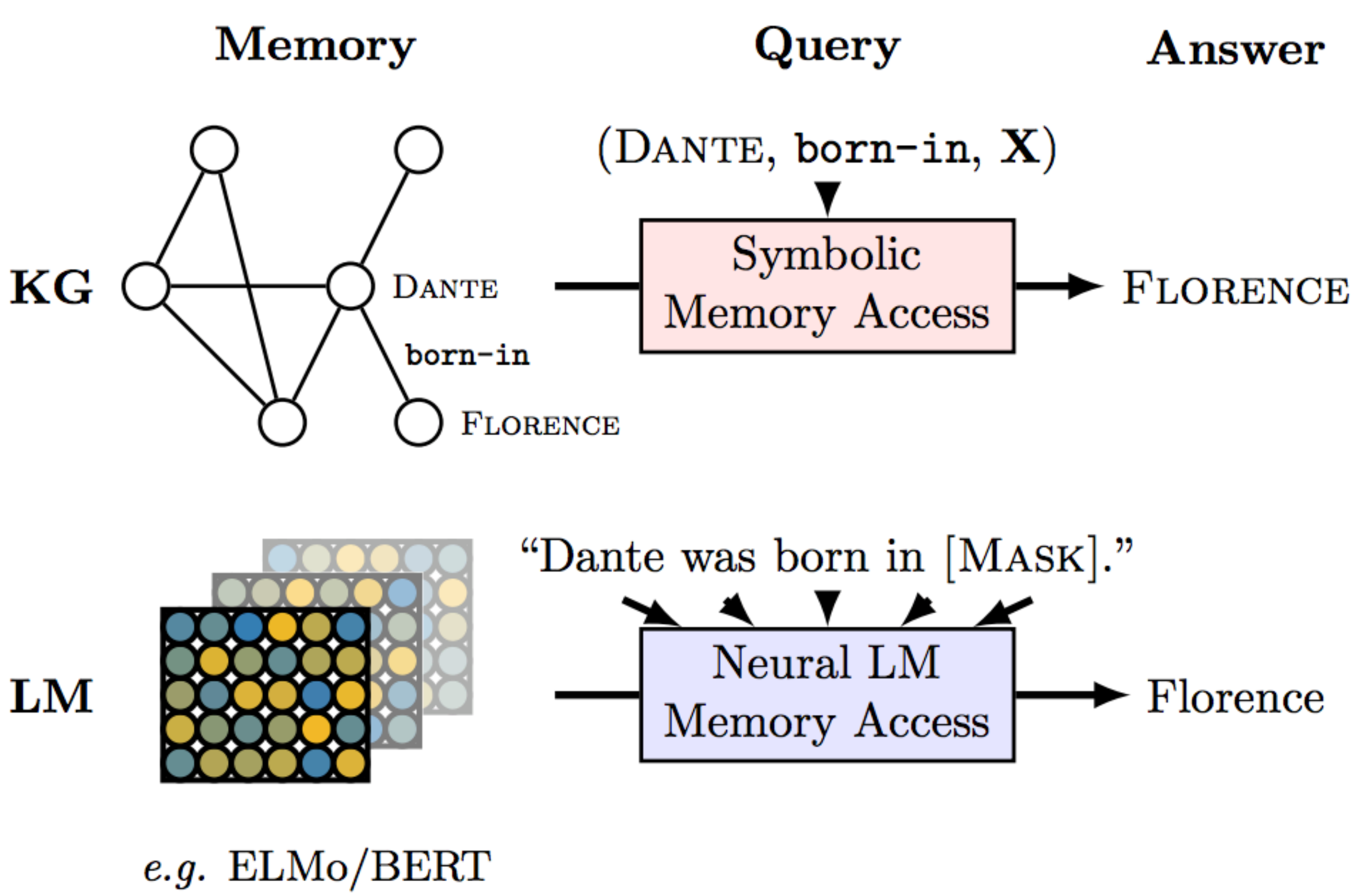
- Traditional QA/MRC models usually refer to external resources to answer questions, e.g., Wikipedia articles or KGs.

- Do LMs pre-trained on a large text corpus already capture those knowledge?

LMs as KBs?

(Petroni et al. 2019)

- Structured queries (e.g., SQL) to query KBs.
- Natural language prompts to query LMs.



LMs as KBs?
(Petroni et al. 2019)

- LAMA benchmark
  - Manual prompts for 41 relations: “[X] was founded in [Y].”
  - Fill in subjects and have LMs (e.g., BERT) predict objects: “Bloomberg L.P. was founded in [MASK].”
  - Accuracy: ELMo 7.1%, Transformer-XL 18.3%, BERT-base 31.1%

Mask 1 Predictions:

5.2% **Chicago**

4.1% **London**

2.8% **Toronto**

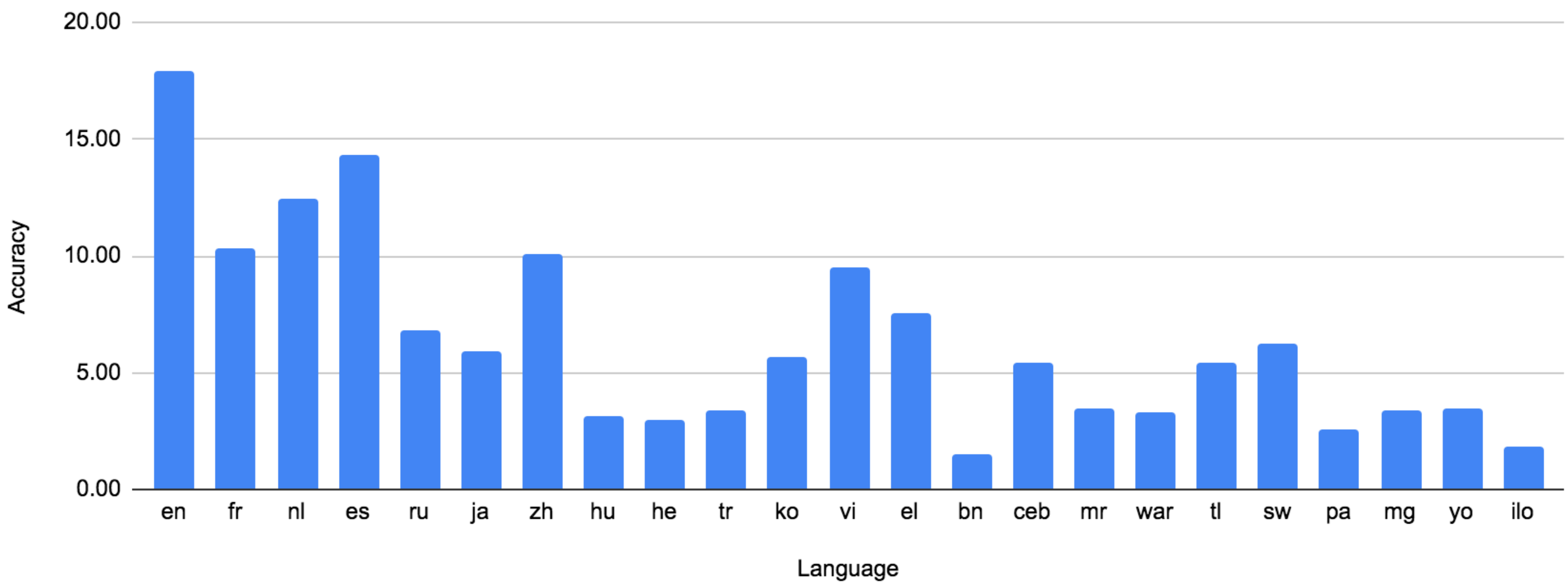
2.3% **c**

1.6% **India**

https://demo.allennlp.org/masked-lm/s/bloomberg-lp-was-founded-mask/I5Q1P2T5Z0

X-FACTR: Multilingual Factual Knowledge Probing (Jiang et al. 2020)

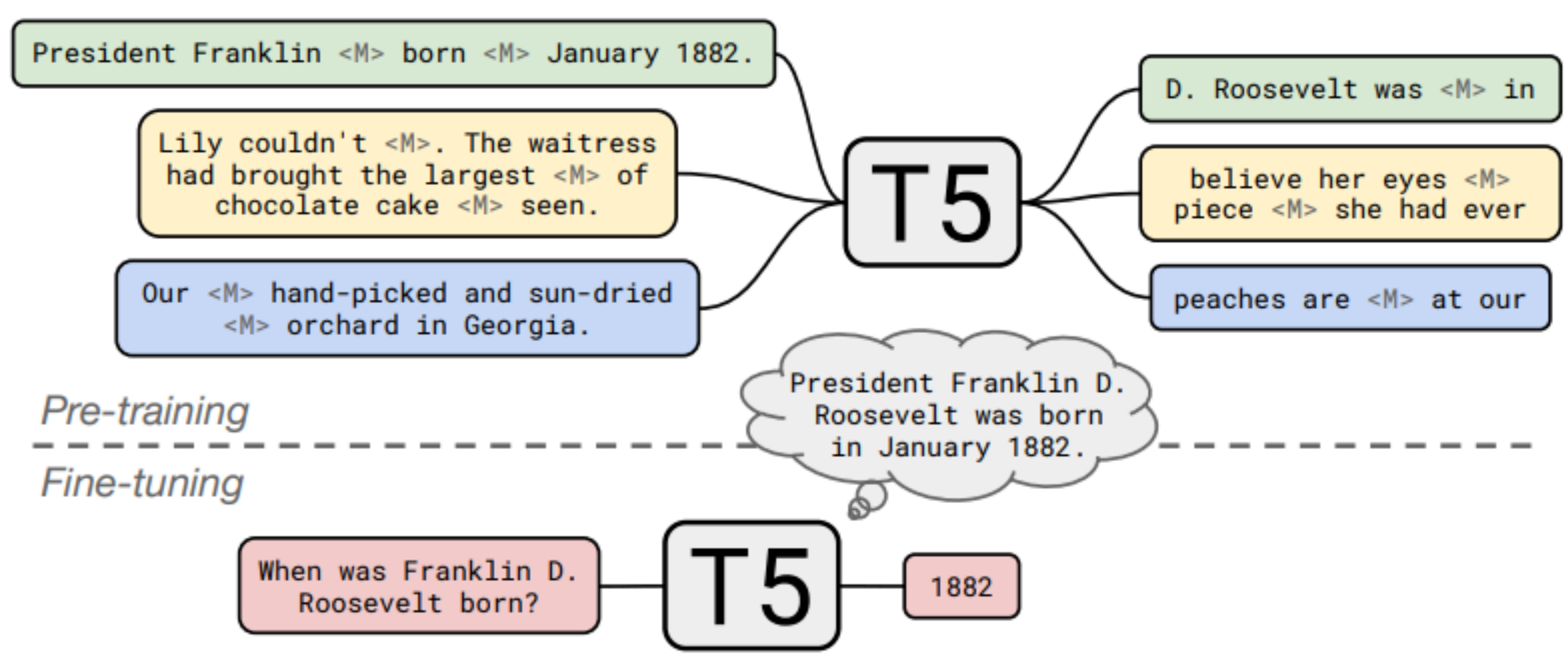
- Overall, factual knowledge in LMs is still limited, especially for low-resource languages.



Max performance of M-BERT, XLM, XLM-R

Close-book T5: Directly Fine-tune with QA Pairs (Roberts et al. 2020)

- Generate answers given questions without additional context.
- Performs even better than QA models with retrieved context.

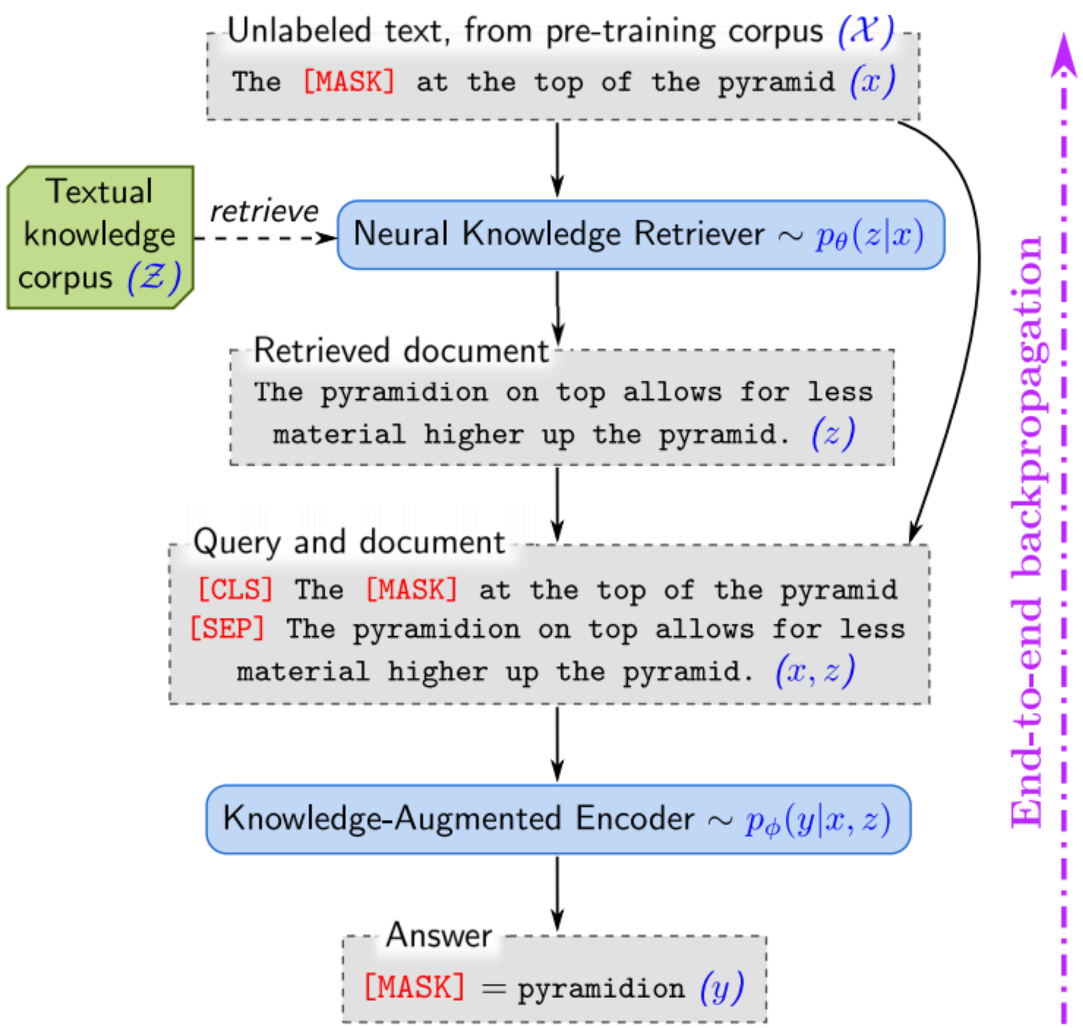


Nonparametric Models Outperform Parametric Models

- For knowledge-intensive tasks like QA, nonparametric models (w/ retrieved context) outperform parametric models (w/o context) by a large margin.

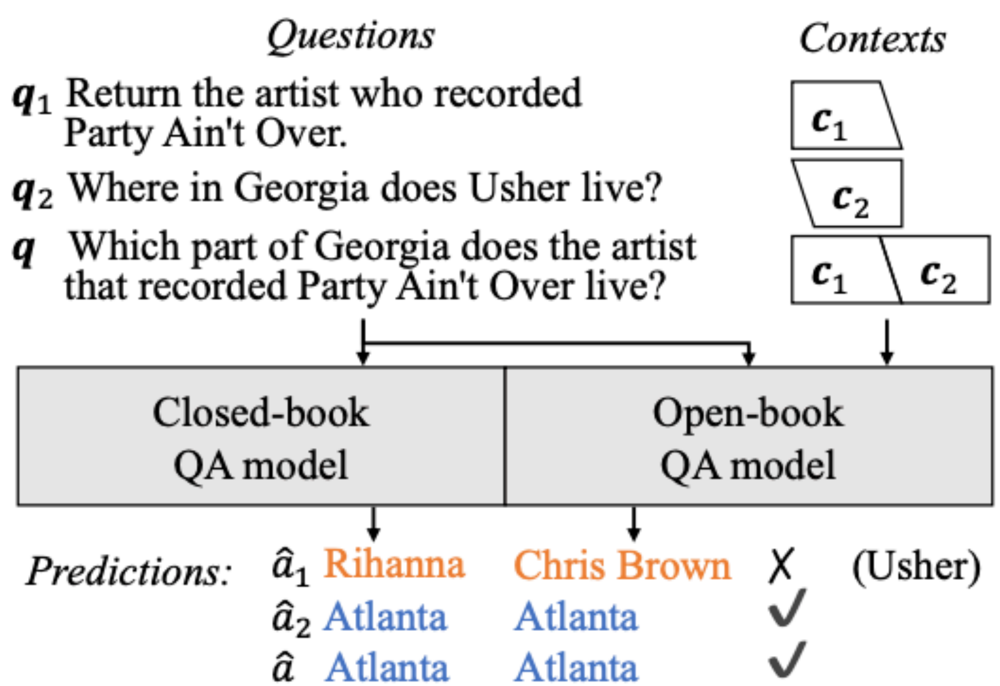
- For example, REALM (Guu et al. 2020), RAG (Lewis et al. 2020) on the NaturalQuestion datasets.

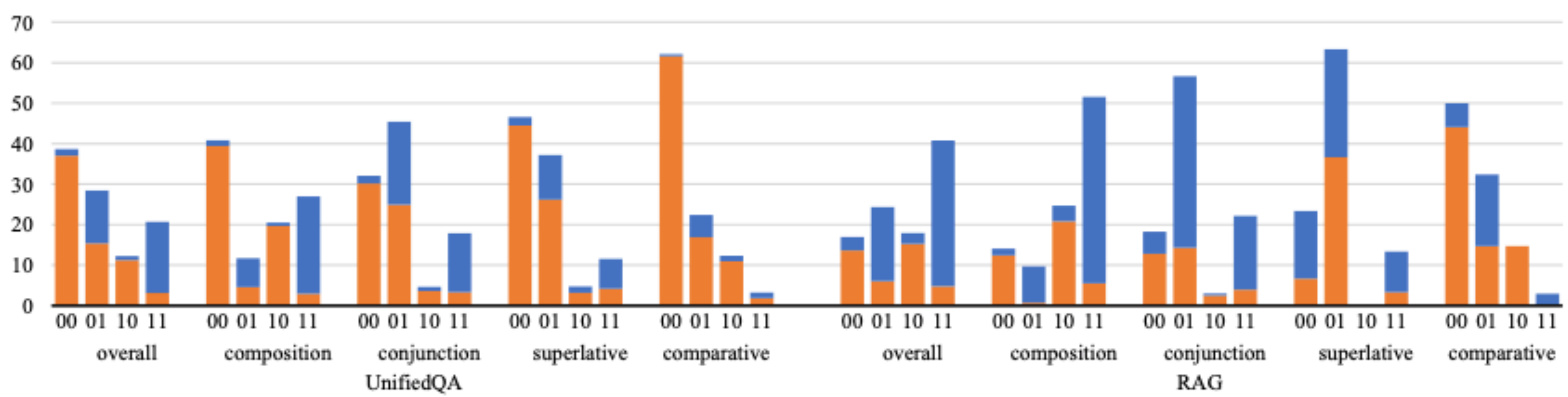
Close-book T5	34.5
REALM	40.4
RAG	44.5



Multi-hop Factual Reasoning in LMs

(Jiang et al. 2022)





UnifiedQA	overall 00	01	10	11	composition 00	01	10	11	conjunction 00	01	10	11	superlative 00	01	10	11	comparative 00	01	10	11
1. Correct %	2	13	1	18	1	7	1	24	2	21	1	15	2	11	2	7	0	5	1	1
0. Incorrect %	37	15	11	3	39	5	20	3	30	25	4	3	45	26	3	4	62	17	11	2
1 / (1+0) %	4	46	8	85	3	61	4	89	6	45	22	81	4	30	33	64	1	24	11	43

RAG	overall 00	01	10	11	composition 00	01	10	11	conjunction 00	01	10	11	superlative 00	01	10	11	comparative 00	01	10	11
1. Correct %	3	18	3	36	2	9	4	46	5	42	0	18	17	27	0	10	6	18	0	3
0. Incorrect %	14	6	15	5	12	1	21	5	13	14	2	4	7	37	0	3	44	15	15	0
1 / (1+0) %	20	75	15	88	12	92	16	89	30	75	17	82	71	42	0	75	12	55	0	100

Figure 2: Correctness confusion matrices of two models on ComplexWebQuestions. Two binary codes on the X-axis indicates the correctness of the first/second single-hop question $s_1 s_2 = \{00, 01, 10, 11\}$. In the table, the first/second row indicates the percentage (%) of examples of which the multi-hop question is correctly/incorrectly answered $P(s = \{1, 0\}, s_1 s_2)$; the last row indicates the conditional success rate $P(s = 1 | s_1 s_2)$.

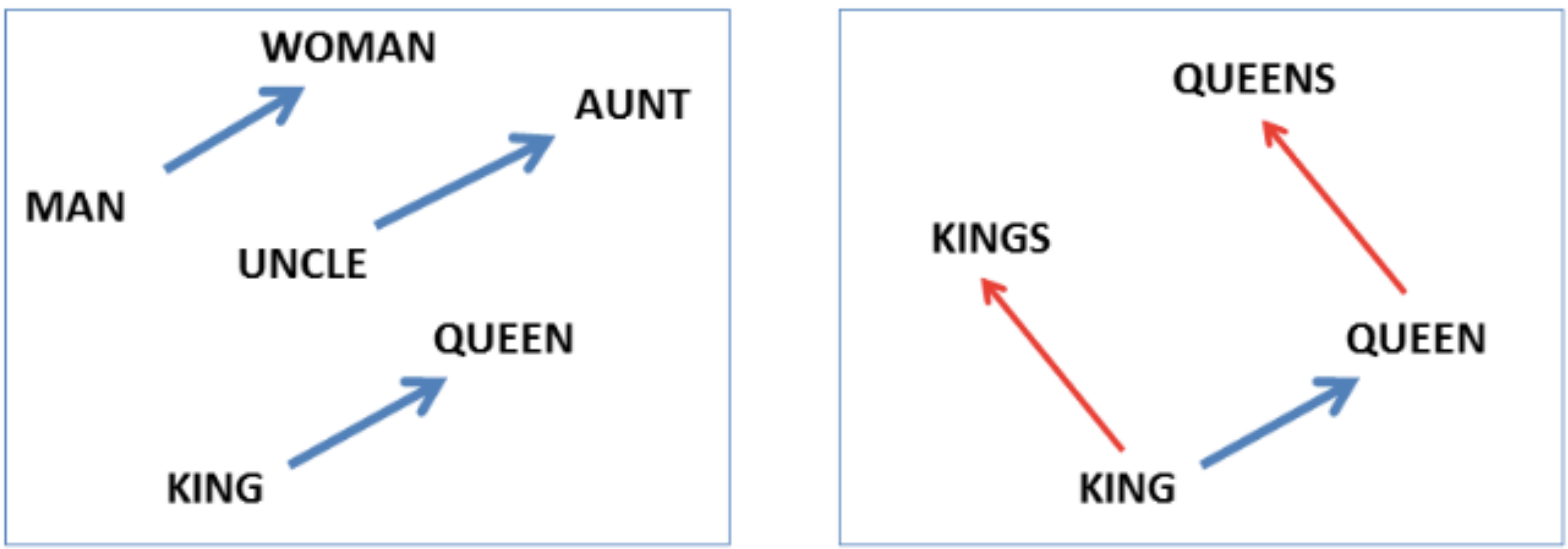
Knowledge Base Completion

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?

Consistency in Embeddings

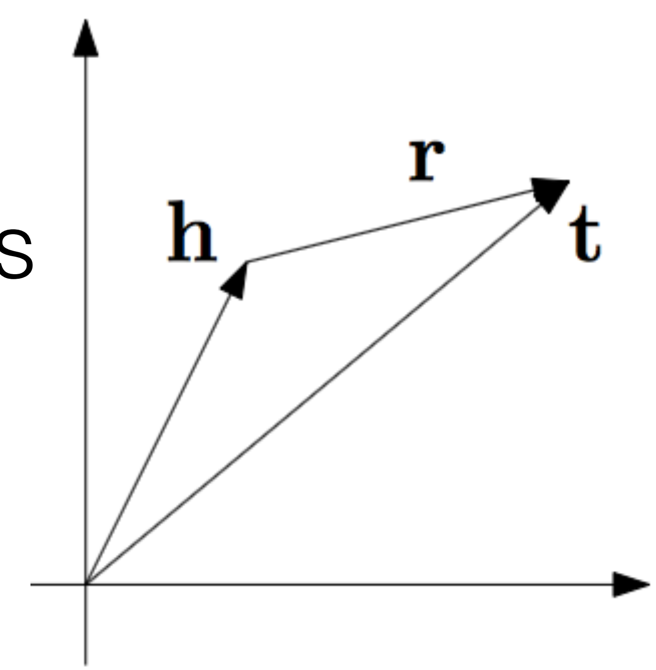
e.g. king-man+woman = queen (Mikolov et al. 2013)



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

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

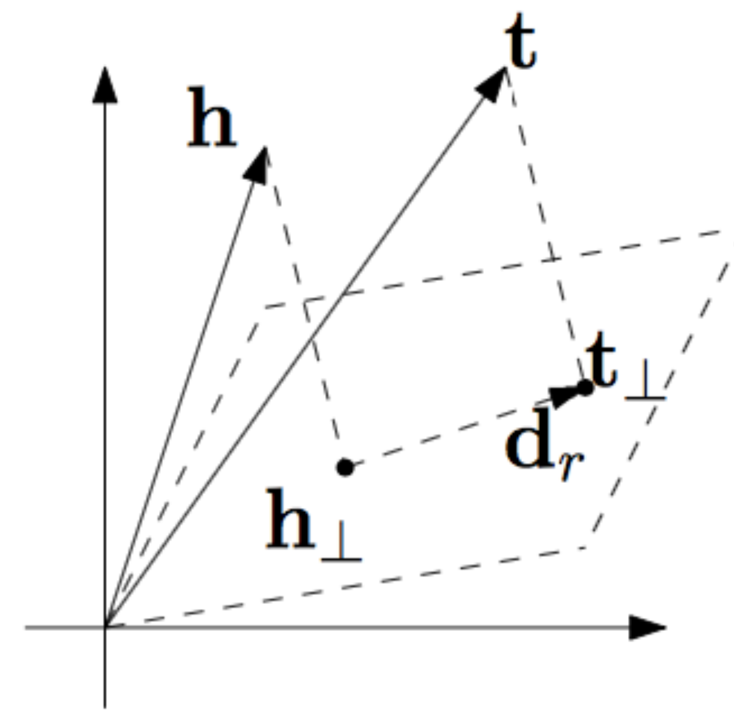


(a) TransE

Expressing Relations w/ Hyperplane Translation (Wang et al. 2014)

- Motivation: Not all dimensions are relevant to a particular relation
- Solution: project the entity vectors on a hyperplane specifically for that relation, then verify relation

$$\|(\mathbf{h} - \mathbf{w}_r^\top \mathbf{h} \mathbf{w}_r) + \mathbf{d}_r - (\mathbf{t} - \mathbf{w}_r^\top \mathbf{t} \mathbf{w}_r)\|_2^2$$



(b) TransH

- Also, TransR (Lin et al. 2015), which uses full matrix projection

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!

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

Supervised Relation Classification (Zeng et al. 2014)

- Extract features and classify
  - Lexical features of the entities themselves
  - Features of the whole span

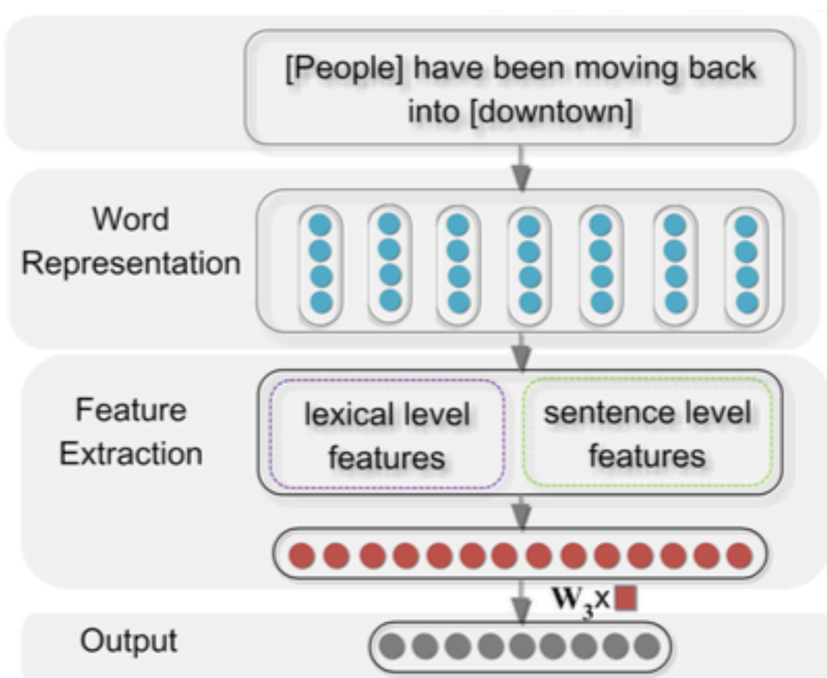


Figure 1: Architecture of the neural network used for relation classification.

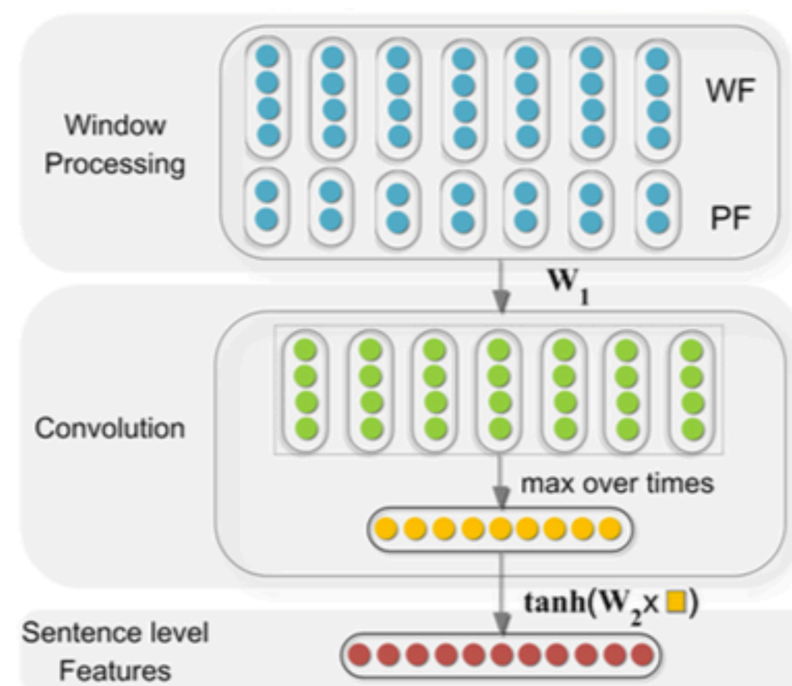
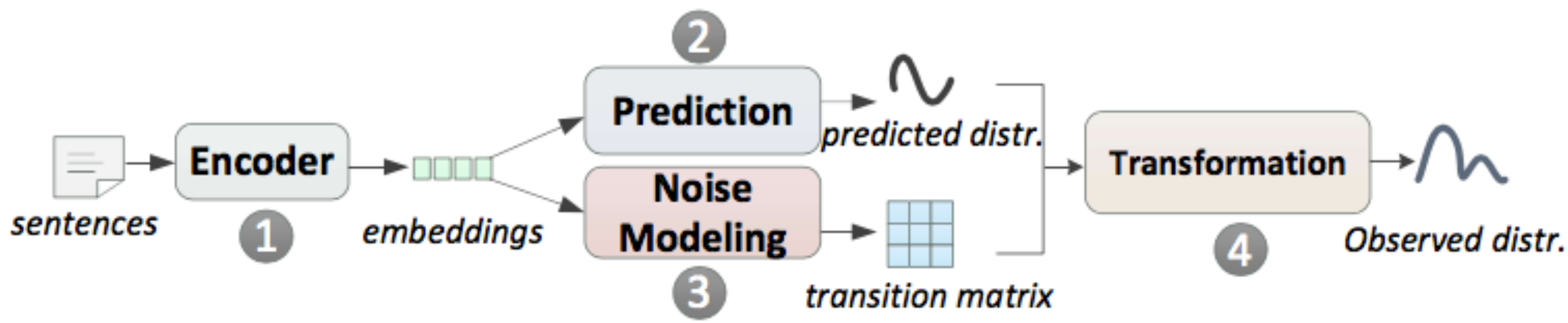


Figure 2: The framework used for extracting sentence level features.

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



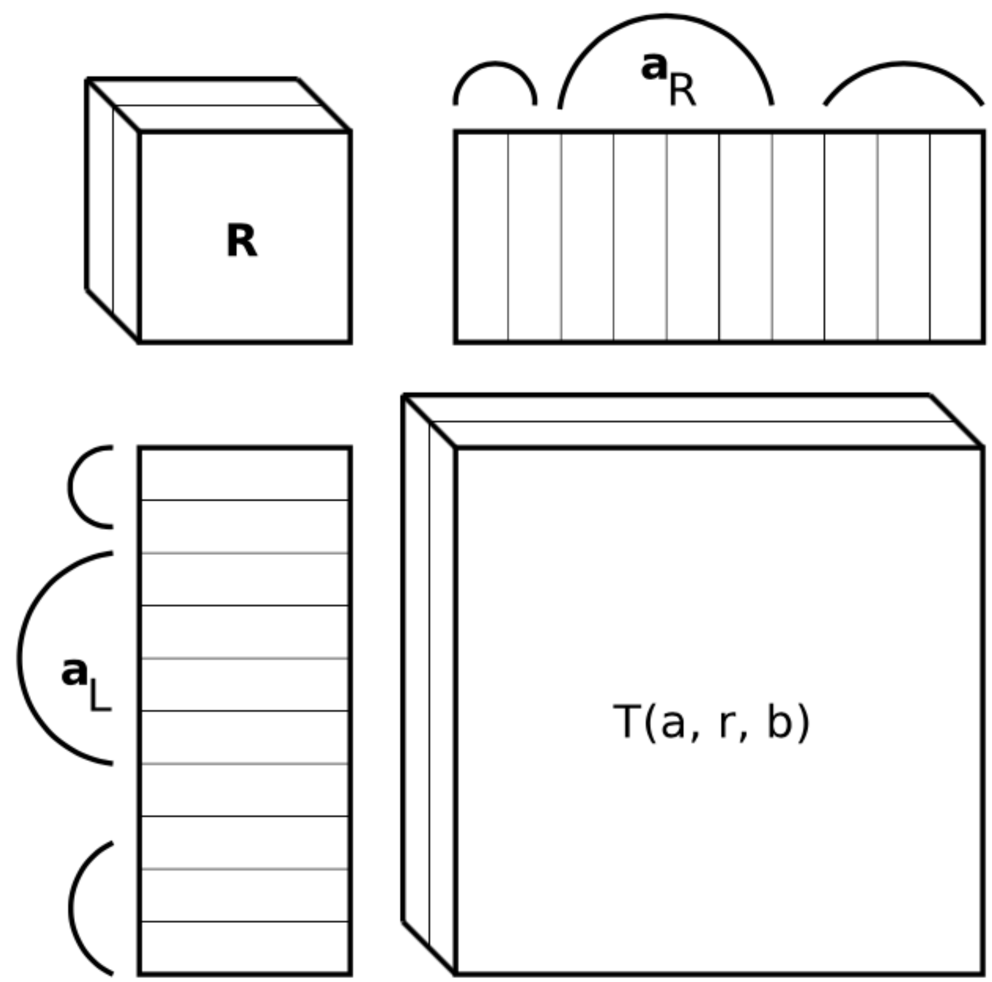
- By controlling the “transition matrix”, we can adjust to the amount of noise expected in the data
    - Trace normalization to try to make matrix close to identity
    - Start training w/ no transition matrix on data expected to be clean, then phase in on full data

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



Modeling Relation Paths

(Lao and Cohen 2010)

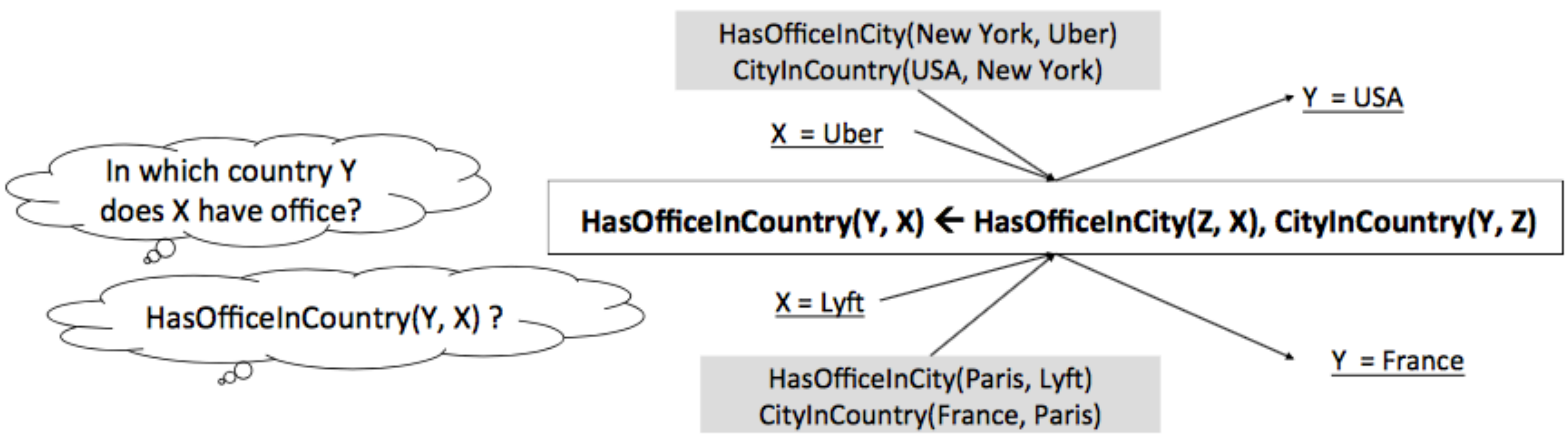
- Multi-step paths can be informative for indicating individual relations
- e.g. “given word, recommend venue in which to publish the paper”

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	$e^* \xrightarrow{AnyPaper} paper \xrightarrow{Cite} paper \xrightarrow{In} journal$

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

$$\sum_l \alpha_l \Pi_{\mathbf{k} \in \beta_l} \mathbf{M}_{\mathrm{R}_\mathbf{k}}$$

Schema-Free Extraction

Open Information Extraction
(Banko et al 2007)

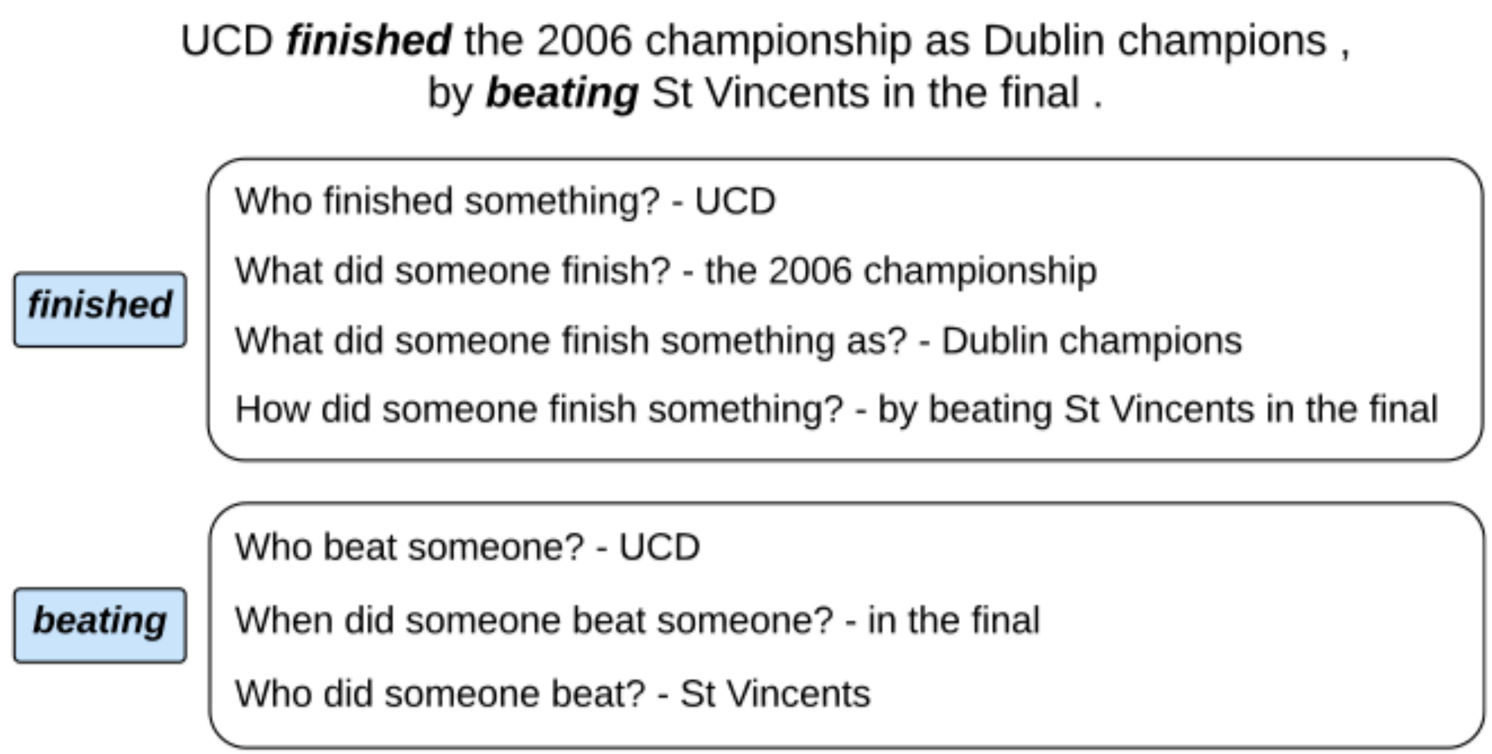
- Basic idea: **the text is the relation**
- e.g. "United has a hub in Chicago, which is the headquarters of United Continental Holdings"
  - {United; has a hub in; Chicago}
  - {Chicago; is the headquarters of; United Continental Holdings}
- Can extract any variety of relations, but does not abstract

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):



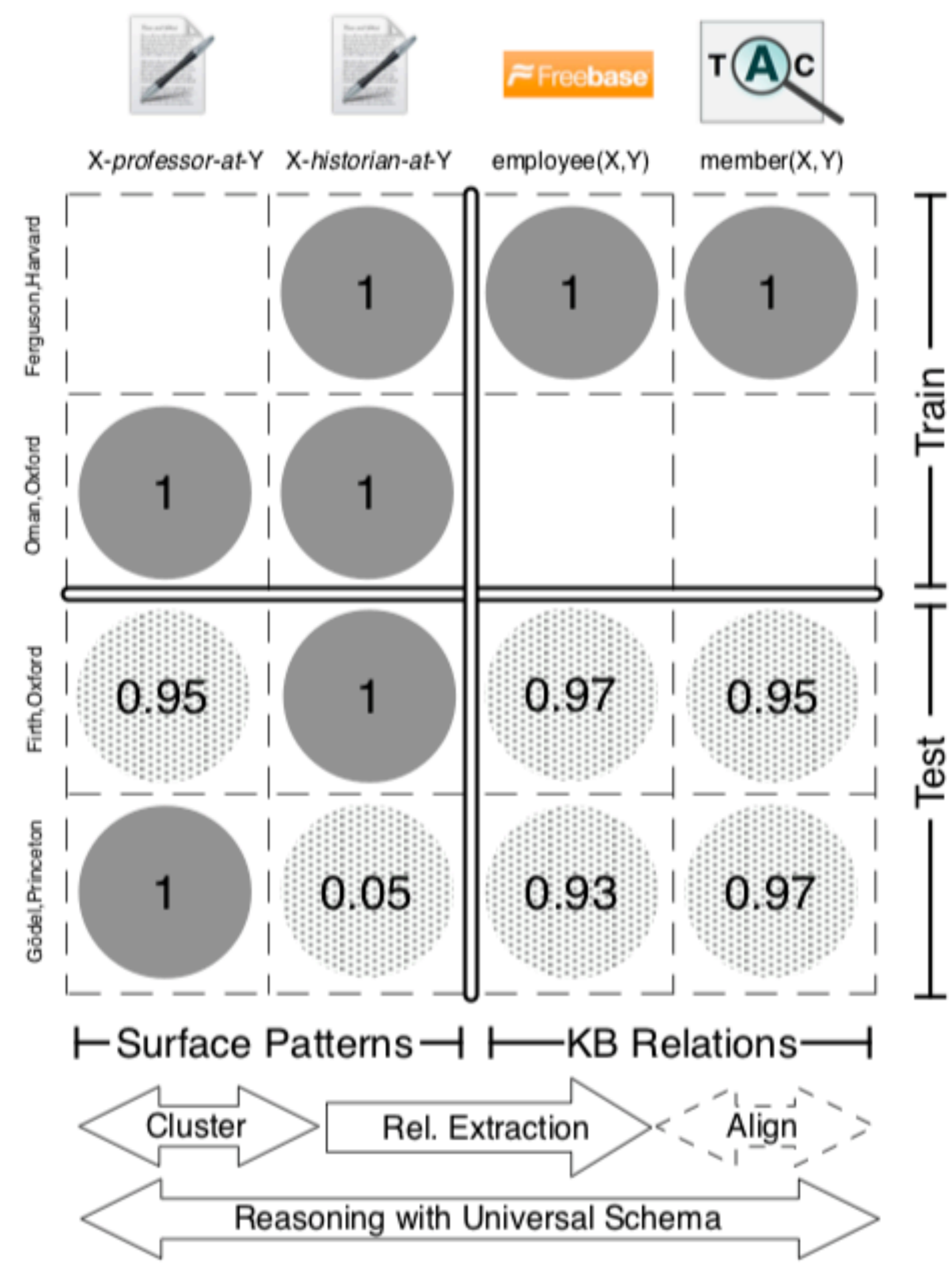
- Can be converted into OpenIE extractions, for use in supervised neural BIO tagger (Stanovsky et al. 2018)

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



Using Knowledge Bases to Inform Language Models

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

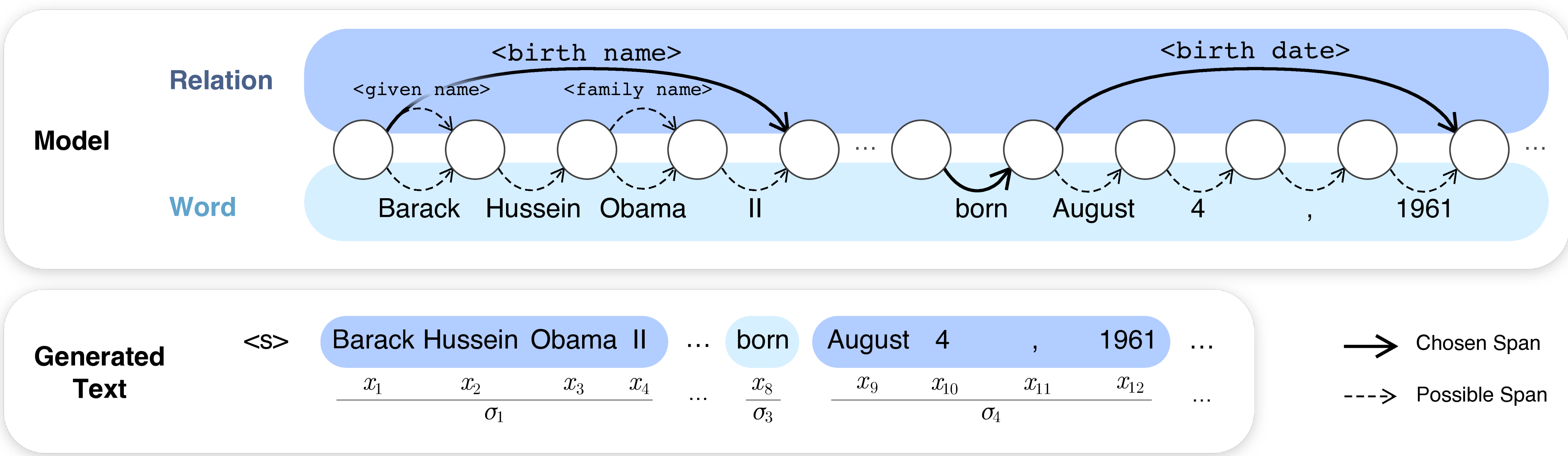
- Post-hoc transformation of embeddings
  - Advantage of being usable with any pre-trained embeddings
- Double objective of making transformed embeddings close to neighbors, and close to original embedding

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

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

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

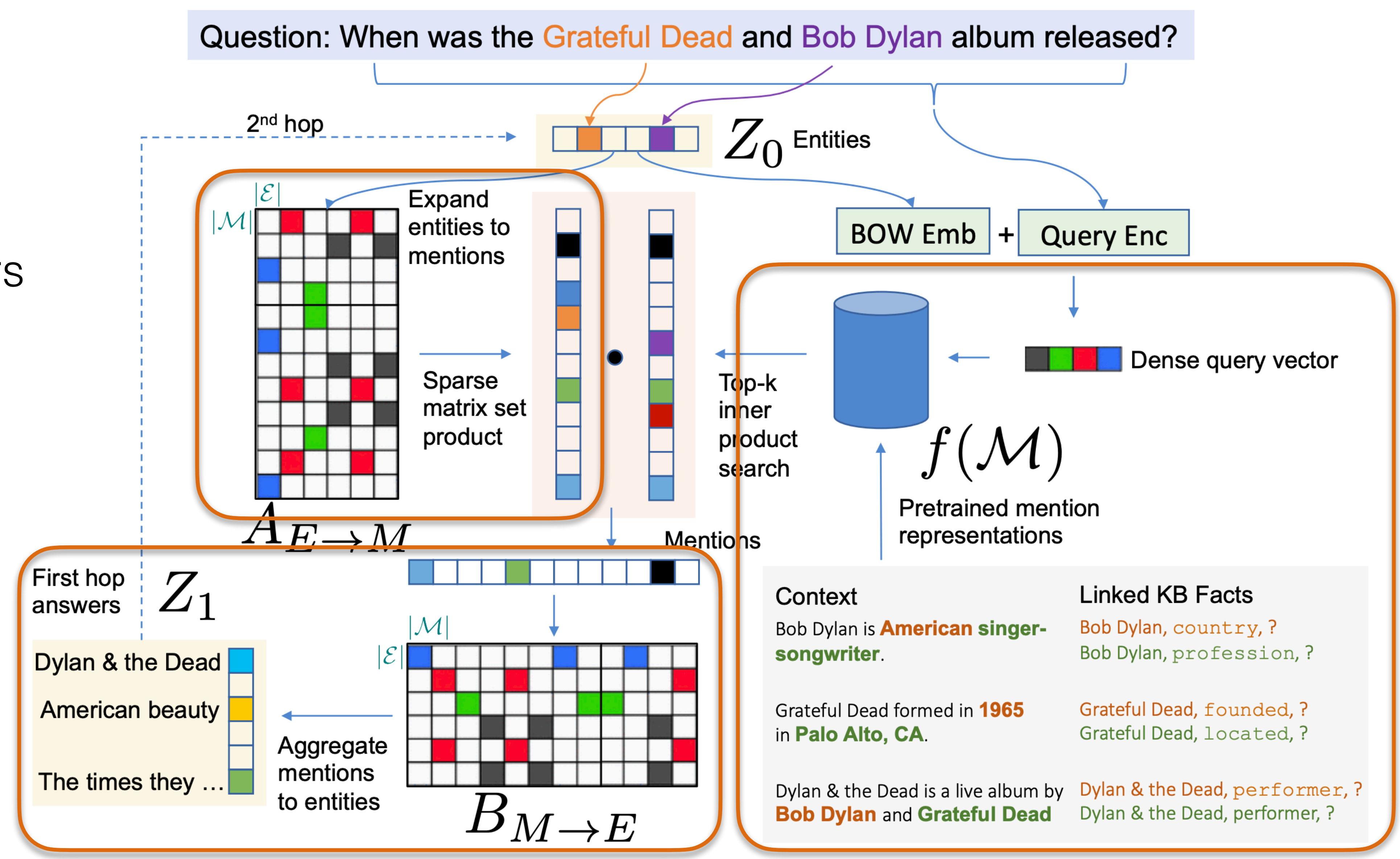
- Provide LMs with topical knowledge in the form of copiable graphs
  - Each (Wiki) text is given relevant KB taken from Wikidata
- Examine all possible decoding "paths" and maximize the marginal probability



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



Questions?