

CS769 Advanced NLP

# Information Extraction and Knowledge-based QA

Junjie Hu



Slides adapted from Graham, Zhengbao  
<https://junjiehu.github.io/cs769-spring23/>

# Goal for Today

- Types of Knowledge Bases (KB)
- Information Extraction (IE) for Constructing KB
  1. IE w/ Pre-defined Relations
  2. OpenIE w/o Pre-defined Relations
  3. Probing Knowledge in LMs
- Using KB to Inform Neural Nets
- Comparison between text/KB/LMs QA

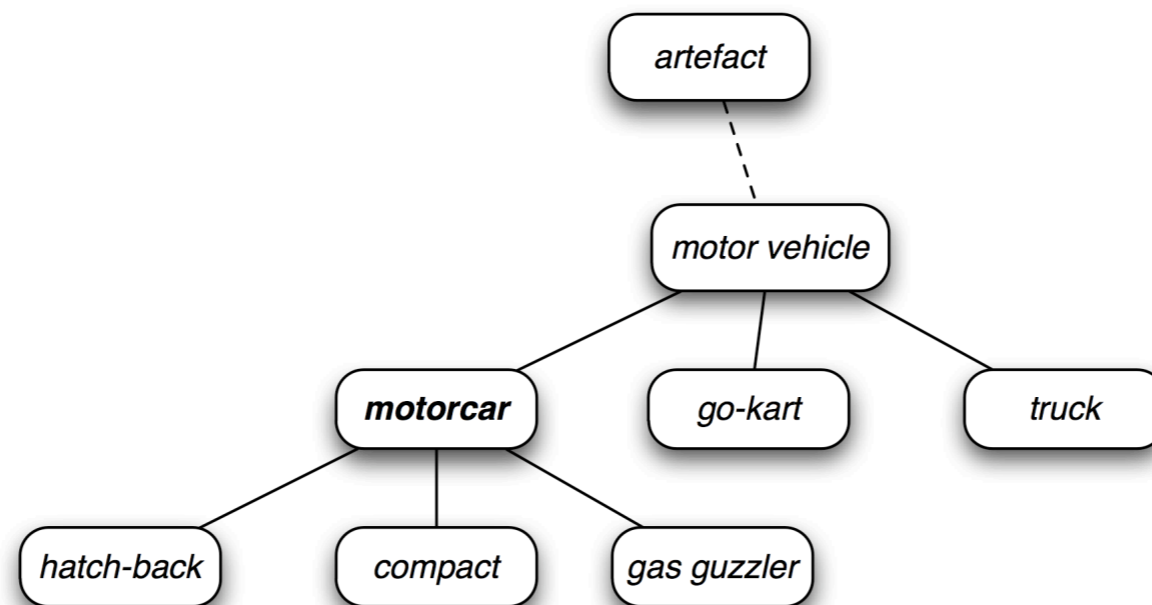
# Knowledge Bases

- 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)

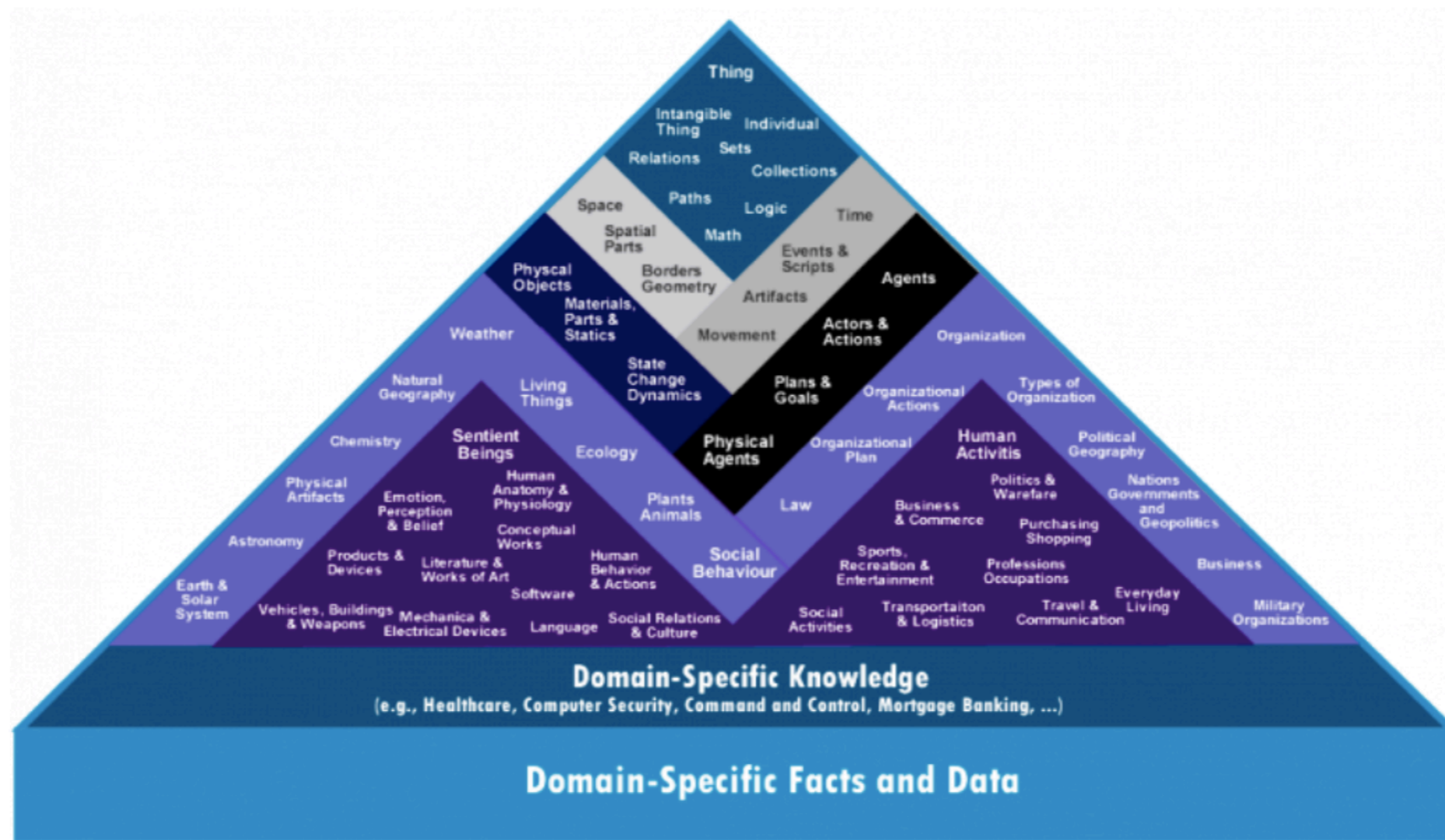
- WordNet is a large database of words including parts of speech, semantic relations



- 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)

# Cyc (Lenant 1995)

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



# DBPedia (Auer et al. 2007)

- Extraction of structured data from Wikipedia

## Carnegie Mellon University

From Wikipedia, the free encyclopedia

**Carnegie Mellon University** (**Carnegie Mellon** or **CMU** /kɑːrnɪɡi 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*.<sup>[9]</sup> It is home to the world's first degree-granting Robotics and Drama programs,<sup>[10]</sup> as well as one of the first Computer Science departments.<sup>[11]</sup> The university was ranked 89th for R&D in 2015 having spent \$242 million.<sup>[12]</sup>

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,<sup>[13]</sup> 12 Turing Award winners, 22 Members of the American Academy of Arts & Sciences,<sup>[14]</sup> 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.<sup>[15]</sup>

## Structured data

Coordinates: 40.443322°N 79.943583°W﻿ / ﻿

### Carnegie Mellon University



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

# WikiData (Bollacker et al. 2008)

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

**Richard Feynman** ▼

[Discuss "Richard Feynman"](#) [Hide Empty Fields](#)

 image 1 of 1

**Types:** [Person \(People\)](#), [Author \(Publishing\)](#), [Physicist \(Science\)](#), [Deceased Person \(People\)](#), [Film writer \(Film\)](#), [Influence Node \(mikelove's types\)](#), [Person Or Being In Fiction \(Fictional Universes\)](#), [Book Subject \(Publishing\)](#)

**Also known as:** [Richard Phillips Feynman](#)

**Gender:** [Male](#)

**Date of Birth:** [May 11, 1918](#)

**Place of Birth:** [Far Rockaway, Queens](#)

**Country Of Nationality:** [United States](#)

**Profession:** [Physicist](#), [Scientist](#)

**Religion:** [Atheism](#)

**Parents:** [double-click to add](#)

**Children:** [Michelle Louise Feynman](#), [Carl Feynman](#)

**Siblings:**

**Sibling**

- [Joan Fey](#)
- Joan Feynman** Person
- [Richard Feynman](#) ...(Richard Phillips Feynman, Author, Physicist, Deceased Person, Film writer)
- [Ana Gasteyer](#) Person, Film actor, TV Actor, Theater Actor
- [Gervase of Tilbury](#) Person
- [Alec Baldwin](#) ...(Alexander Rae Baldwin, Person, Film actor, Film director, Film producer, TV actor)
- [Ernest Thesiger](#) Person, Film actor, Deceased Person
- [Mean Girls](#) Film
- [Riverside Drive](#) Landscape project
- [Portrait of Jennie](#) Film
- [Television Personalities](#) (The Television Personalities)

[Create New Person](#)

**Page History**  
Created by [Melaweb](#) Oct 22, 2006  
Last edited by [robert](#) Oct 29, 2007

**Web Link(s)**  
*double-click to add*

**Employment history**  
[Cornell University](#)  
[California Institute of Technology](#)  
[Thinking Machines](#)

**Education**  
[Princeton University](#) • 1942 • Ph.D.  
[Massachusetts Institute of Technology](#) • 1939 • Bachelor's degree

**Quotations**  
[I like sex: sure, it may give some results, but that's not why we do it.](#)  
[I do not create, I do not understand.](#)

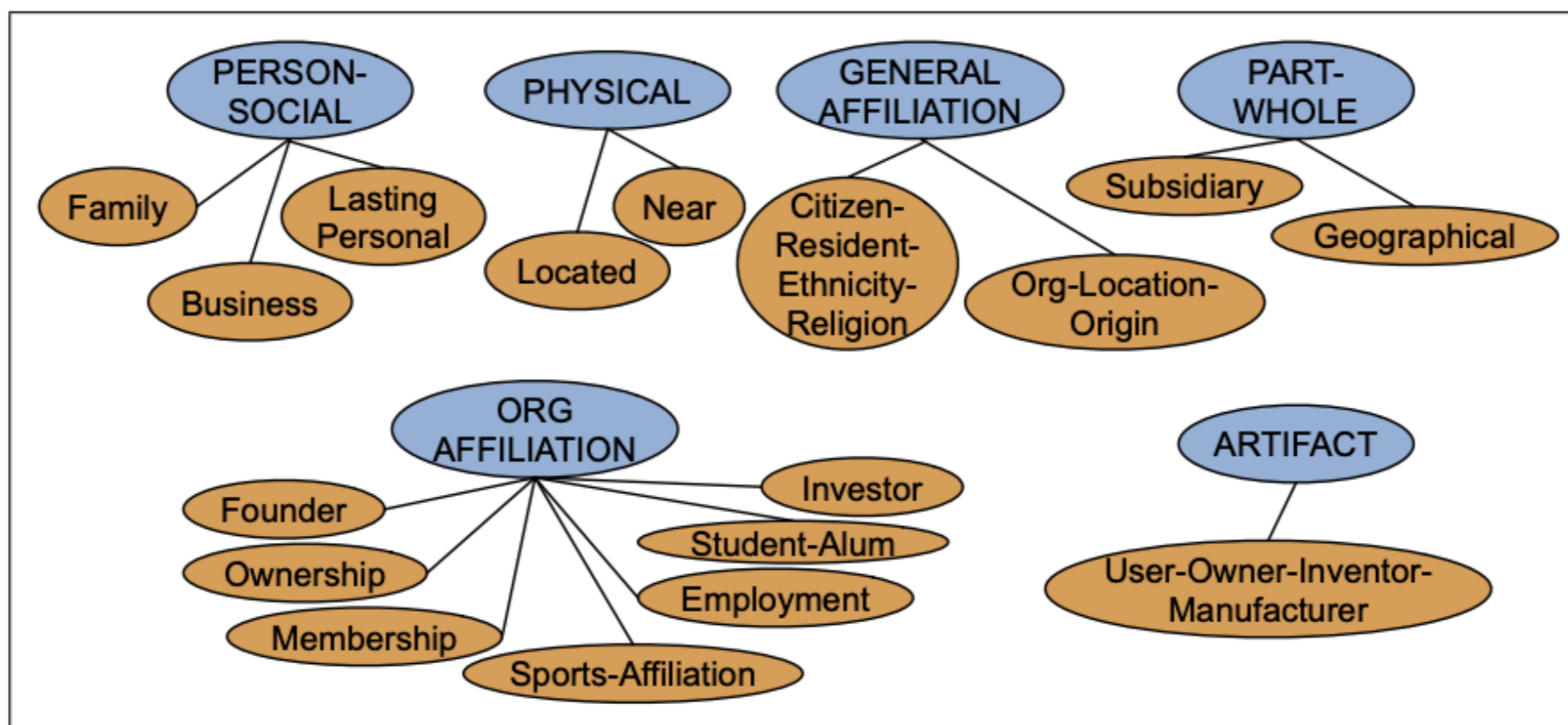
**Books Written**  
[What Do You Care What Other People Think?](#)  
[The Pleasure of Finding Things Out](#)  
[The Feynman Lectures on Physics](#)  
[Surely You're Joking, Mr. Feynman!](#)



# Information Extraction w/ Pre-defined Relations

# Pre-defined Relations

- Define a set of relations (a.k.a. schema) that we could extract for pairs of entities from text.



**Figure 17.1** The 17 relations used in the ACE relation extraction task.

Relations	Types	Examples
Physical-Located	PER-GPE	<b>He</b> was in <b>Tennessee</b>
Part-Whole-Subsidiary	ORG-ORG	<b>XYZ</b> , the parent company of <b>ABC</b>
Person-Social-Family	PER-PER	<b>Yoko</b> 's husband <b>John</b>
Org-AFF-Founder	PER-ORG	<b>Steve Jobs</b> , co-founder of <b>Apple...</b>

**Figure 17.2** Semantic relations with examples and the named entity types they involve.

# Supervised Relation Extraction Baseline

- **Training:**
  - Labeled dataset: a KB triple  $t = \langle e1, r, e2 \rangle$  on a sentence  $s$
  - Supervised training of models (e.g., logistic regression, NNs)
- **Test:**
  - Find any pairs of entities in a sentence
  - Apply the relation classifier on all entity pairs

**function** FINDRELATIONS(*words*) **returns** *relations*

*relations*  $\leftarrow$  *nil*

*entities*  $\leftarrow$  FINDENTITIES(*words*)

**forall** **entity pairs**  $\langle e1, e2 \rangle$  **in** *entities* **do**

**if** RELATED?(*e1, e2*)

*relations*  $\leftarrow$  *relations* + CLASSIFYRELATION(*e1, e2*)

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

- **Motivation:** Supervised baseline is still limited to the labeled data size.
- Given an entity-relation-entity triple, extract all text that matches these two entities, and use these texts to train the classifier

*[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]...*

- Extract hand-crafted features from this large corpus of (noisily) labeled text to train a system (e.g., multi-class logistic regression)

# 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

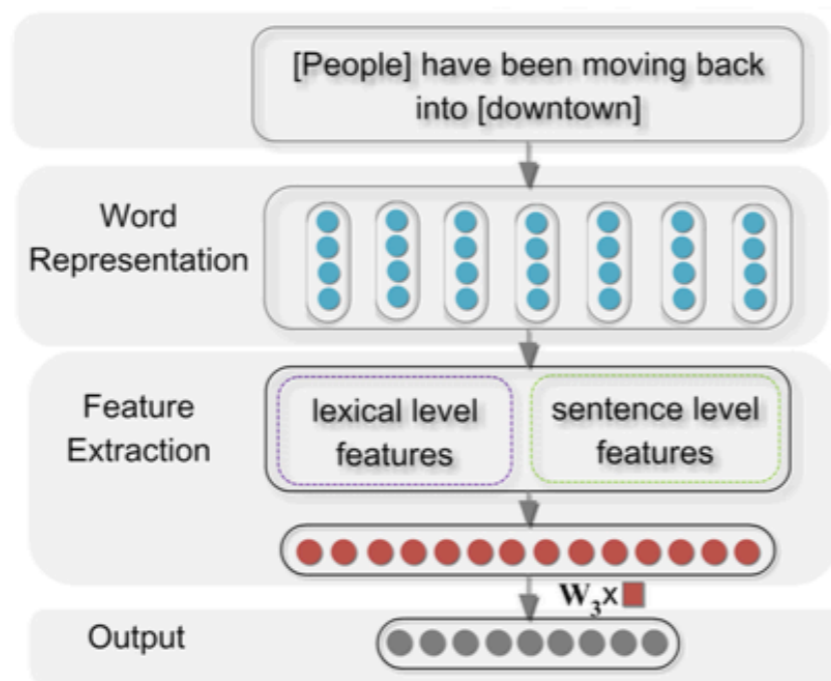


Figure 1: Architecture of the neural network used for relation classification.

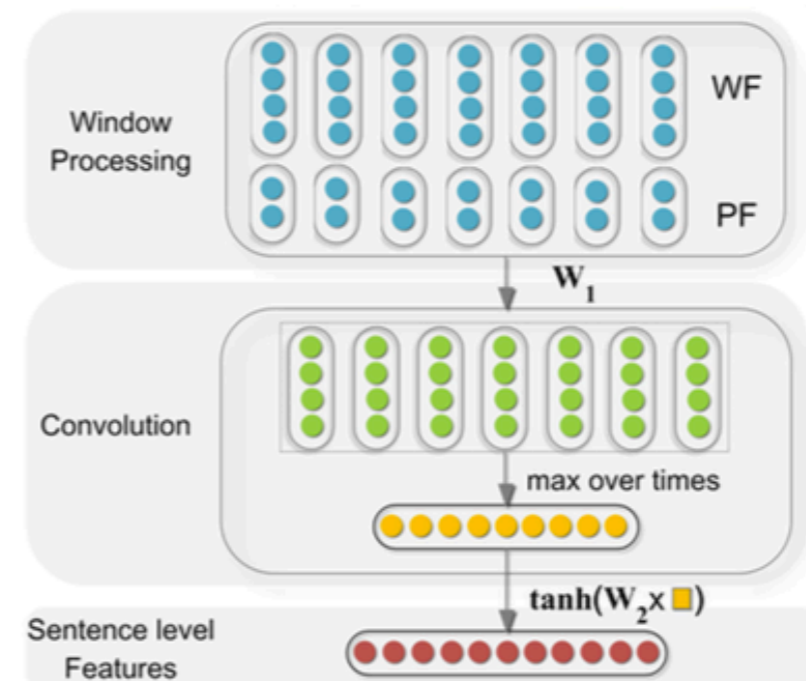
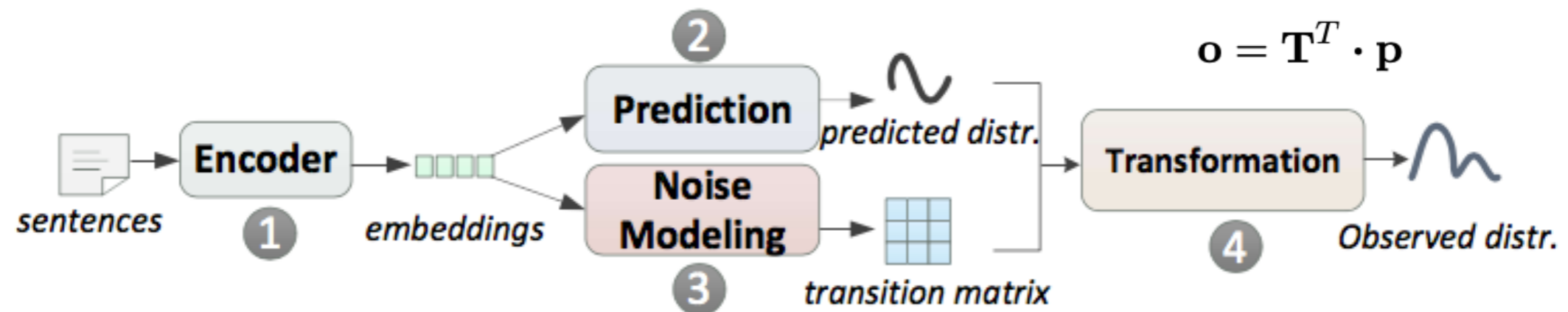


Figure 2: The framework used for extracting sentence level features.

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

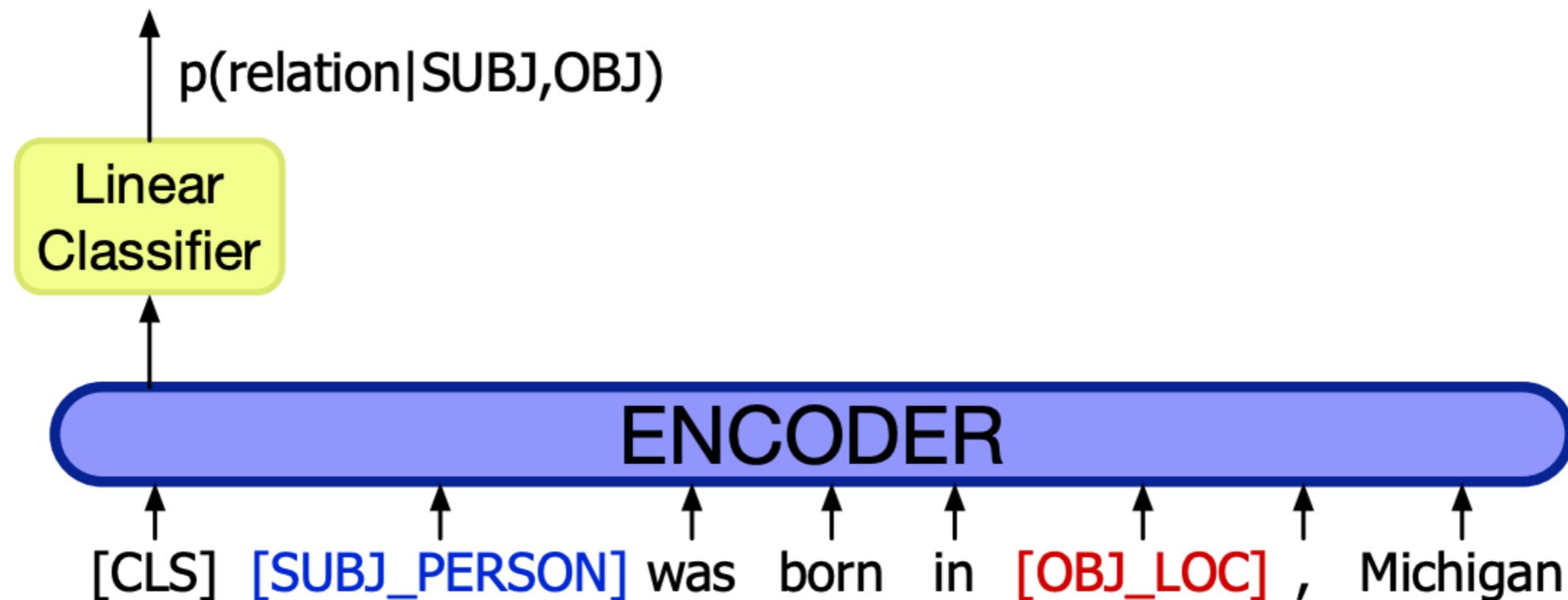
- Idea: there is noise in distant supervision labels, so we want to model it



- 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

# Relation Extraction w/ Pre-trained LMs

- Relation extraction as a linear layer on top of an encoder (e.g., BERT), with the subject and object entities replaced in the input by their NER tags (Zhang et al. 2017, Joshi et al. 2020).



# Schema-Free Extraction (Open IE)



# Open Information Extraction

(Banko et al 2007)

- Basic idea: **the text is the relation. No pre-defined set of relation types!**
- 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 relation strings, but does not abstract these relation strings to a relation type

# 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 to annotators (He et al. 2015):

UCD **finished** the 2006 championship as Dublin champions ,  
by **beating** St Vincents in the final .

**finished**

Who finished something? - UCD  
What did someone finish? - the 2006 championship  
What did someone finish something as? - Dublin champions  
How did someone finish something? - by beating St Vincents in the final

**beating**

Who beat someone? - UCD  
When did someone beat someone? - in the final  
Who did someone beat? - St Vincents

- Can be converted into OpenIE extractions

# Probing Knowledge in LMs

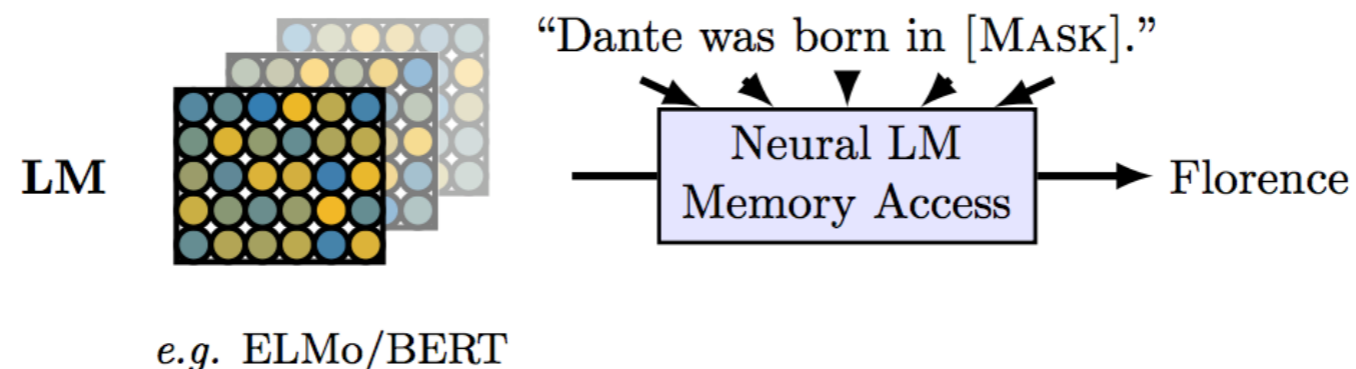
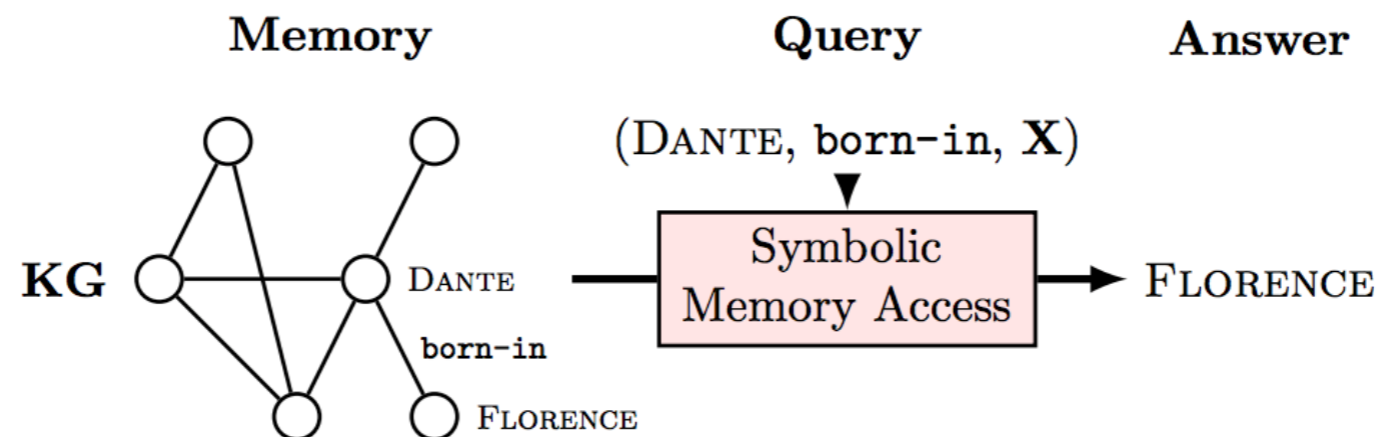
# Probing Knowledge in LMs

- 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**

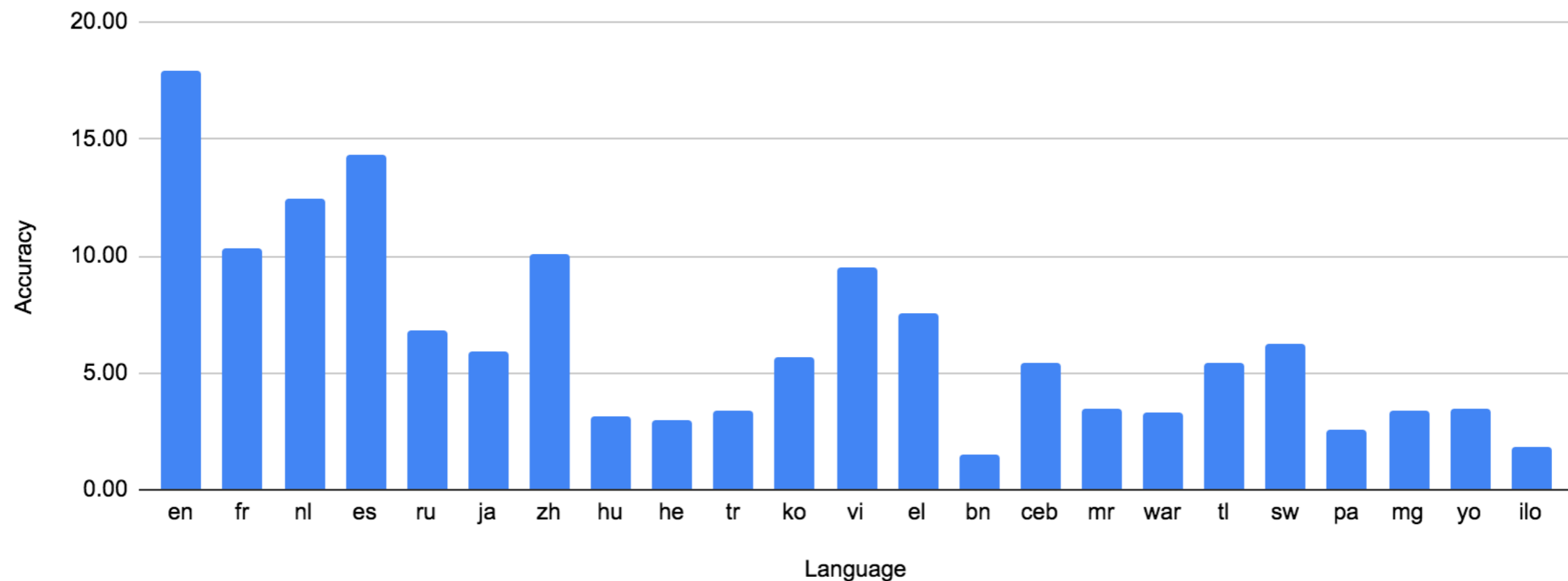
2.8% **Toronto**

2.3% **c**

1.6% **India**

# 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



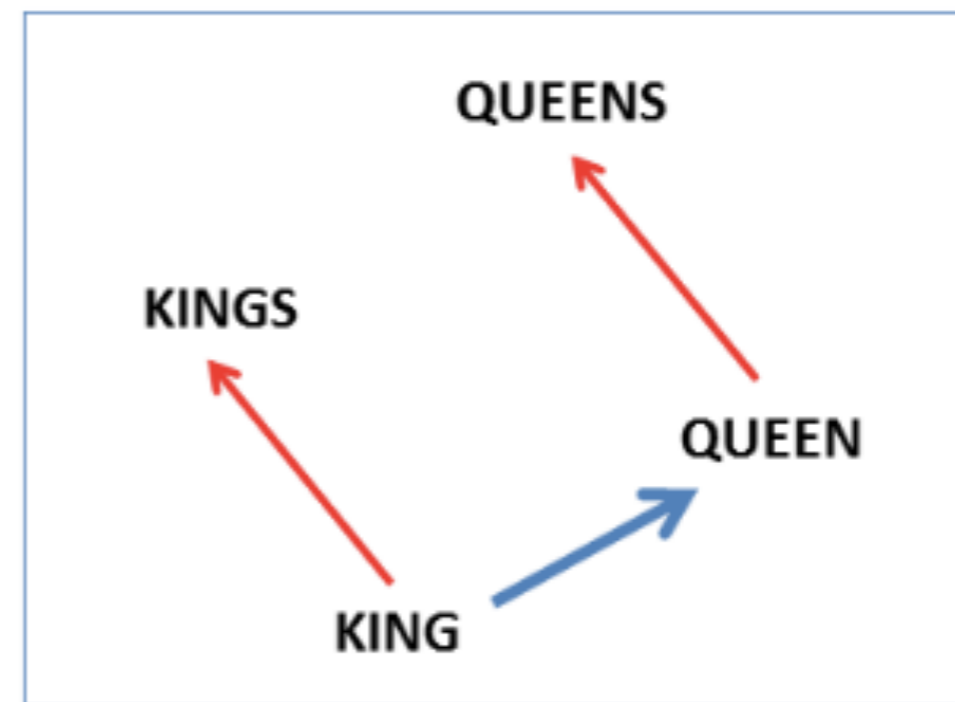
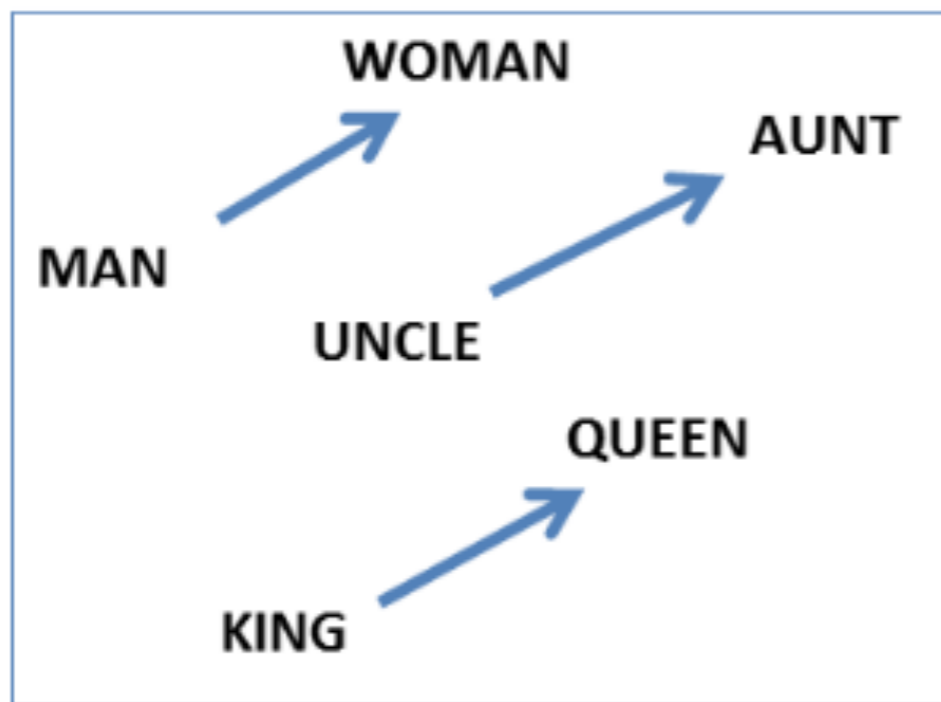
# Learning Representations for Knowledge Bases

# 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)

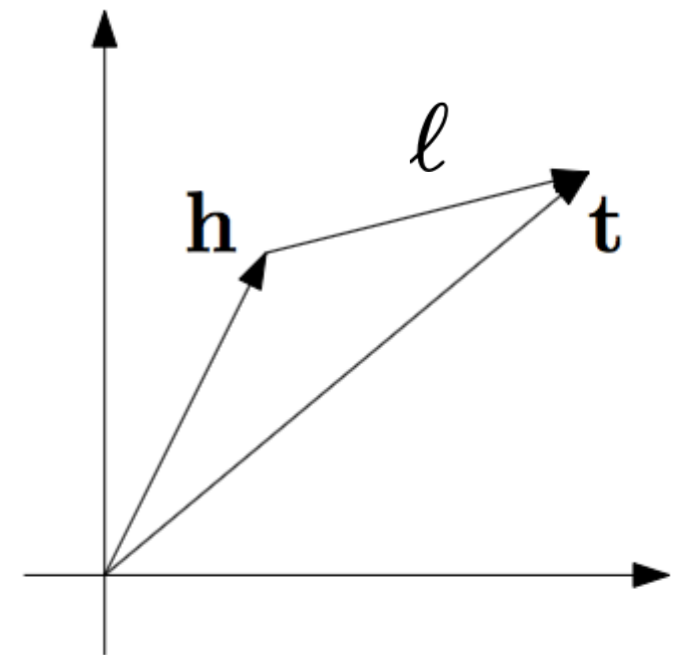


# Learning Knowledge Graph Embeddings (Bordes et al. 2013)

- Motivation: express triples as additive transformation
- Method: minimize the distance of existing triples with a margin-based loss that

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

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

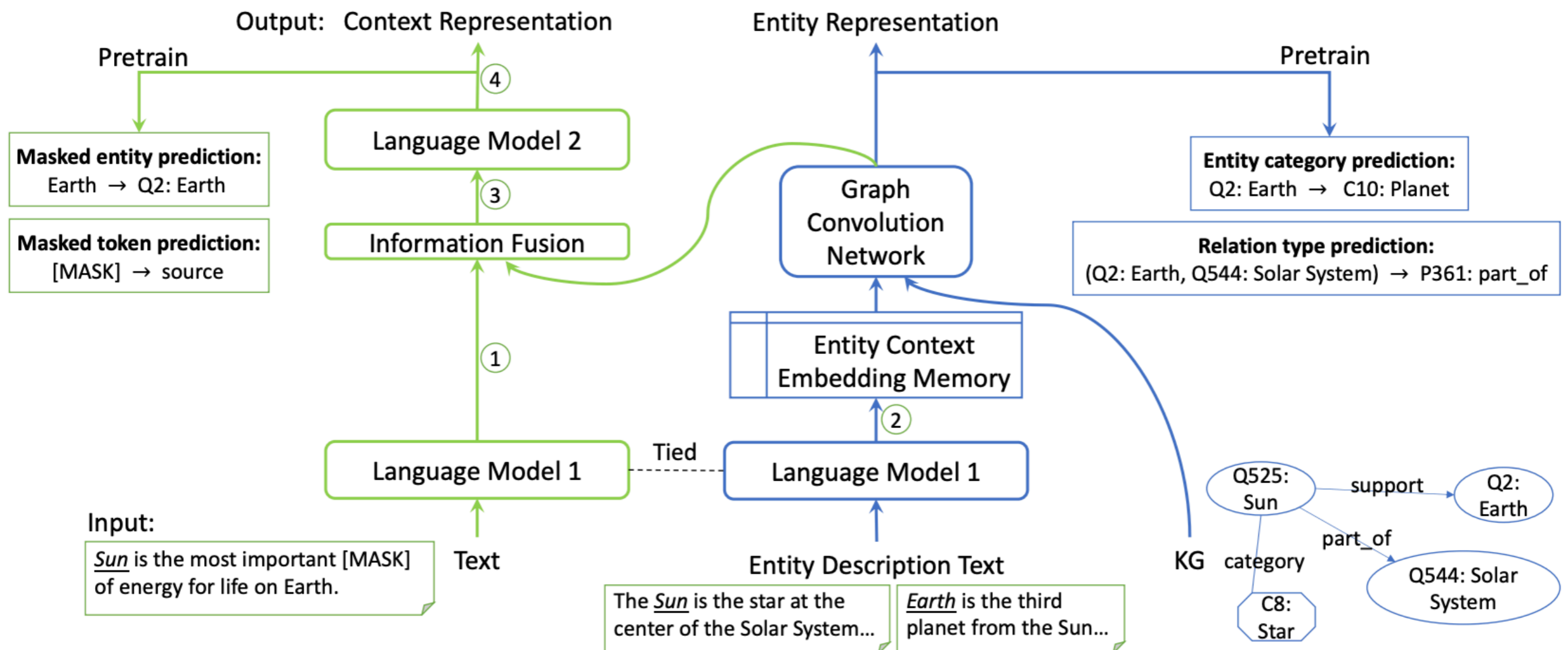


(a) TransE

# Joint Text-Graph Pre-training

(JAKET, Yu et al. 2022)

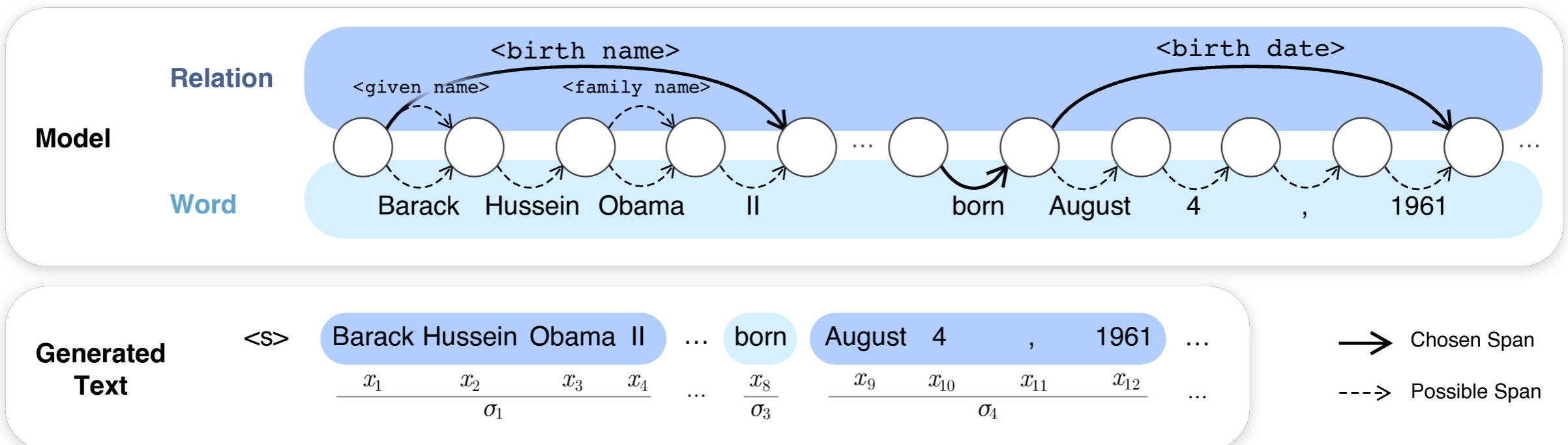
- Inspired by masked language model pre-training, we can pre-train a graph neural network for a knowledge graph
- Self-supervised tasks: **Entity category prediction and relation type prediction**



# Using Knowledge Bases to Inform Neural Models

# 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



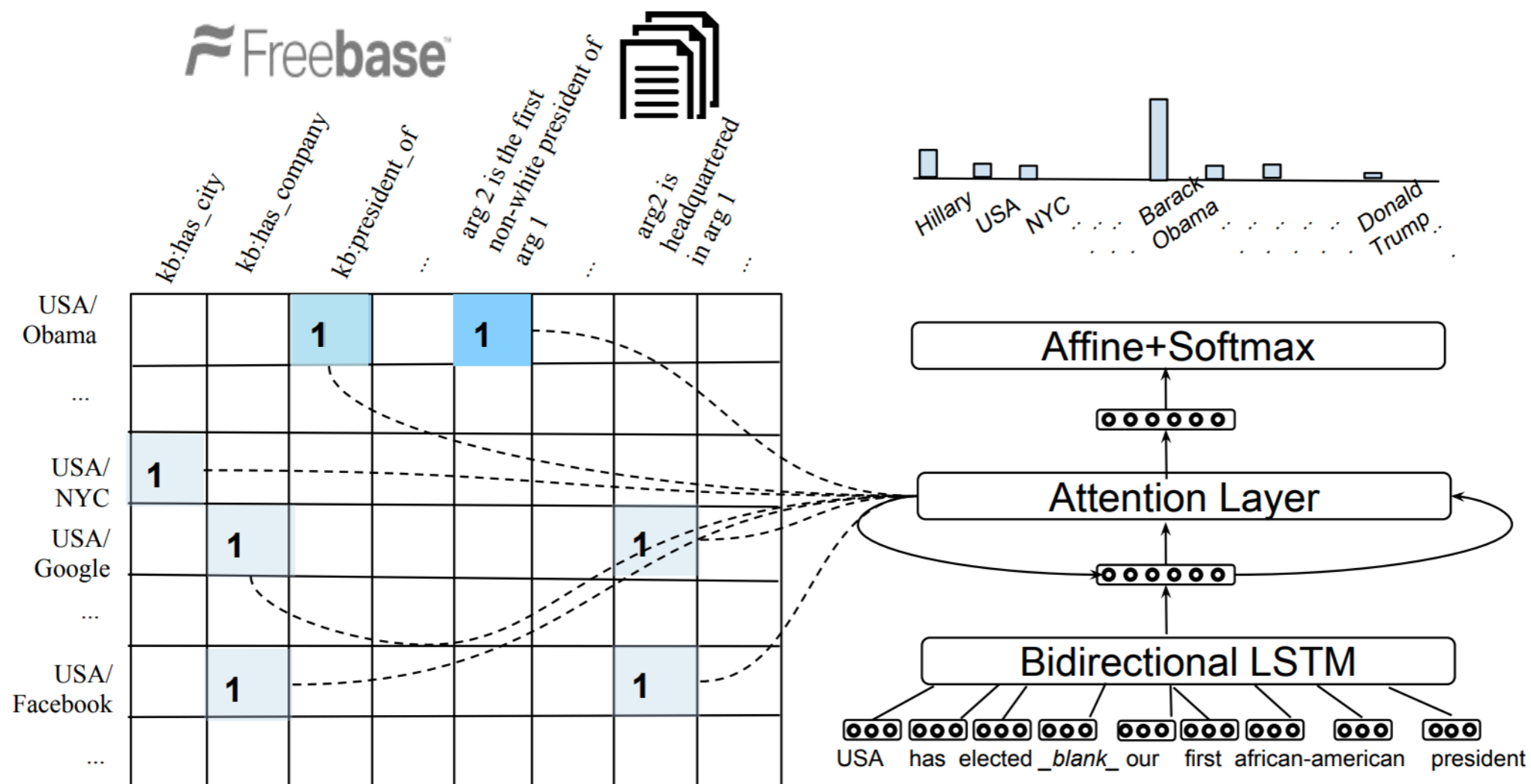
# Knowledge Base Question Answering (KBQA)

- Construct a KB from texts or other resources either manually or automatically
- **Symbolic method (semantic parsing)**: convert a natural language query into a structured format (e.g., SQL) to query the KB
- **Neural symbolic method**: embed a natural language query and the knowledge base information into an embedding space, learn some integration modules to combine information, and make answer prediction



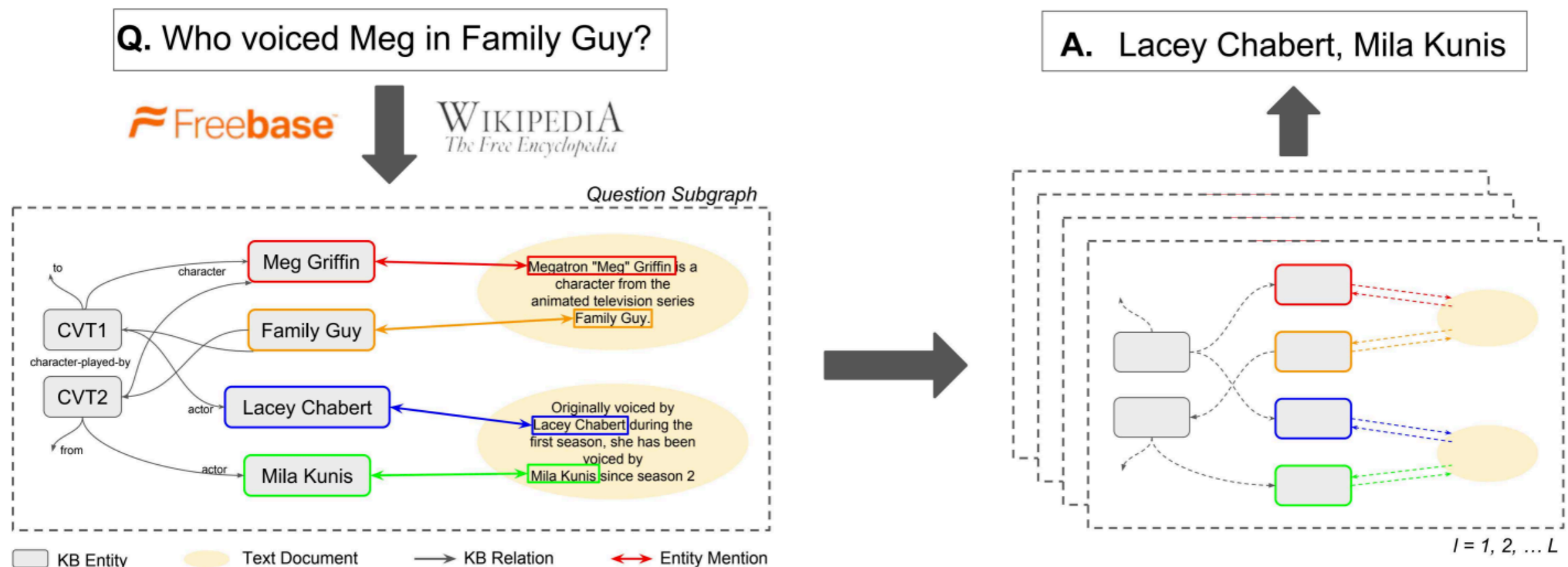
# QA on KB w/ Universal Schema and Memory Network (Das et al. 2017)

- Represent each KB entity as a row in a memory matrix
- Use attention to retrieve relevant entities for QA



# QA w/ KB and Text

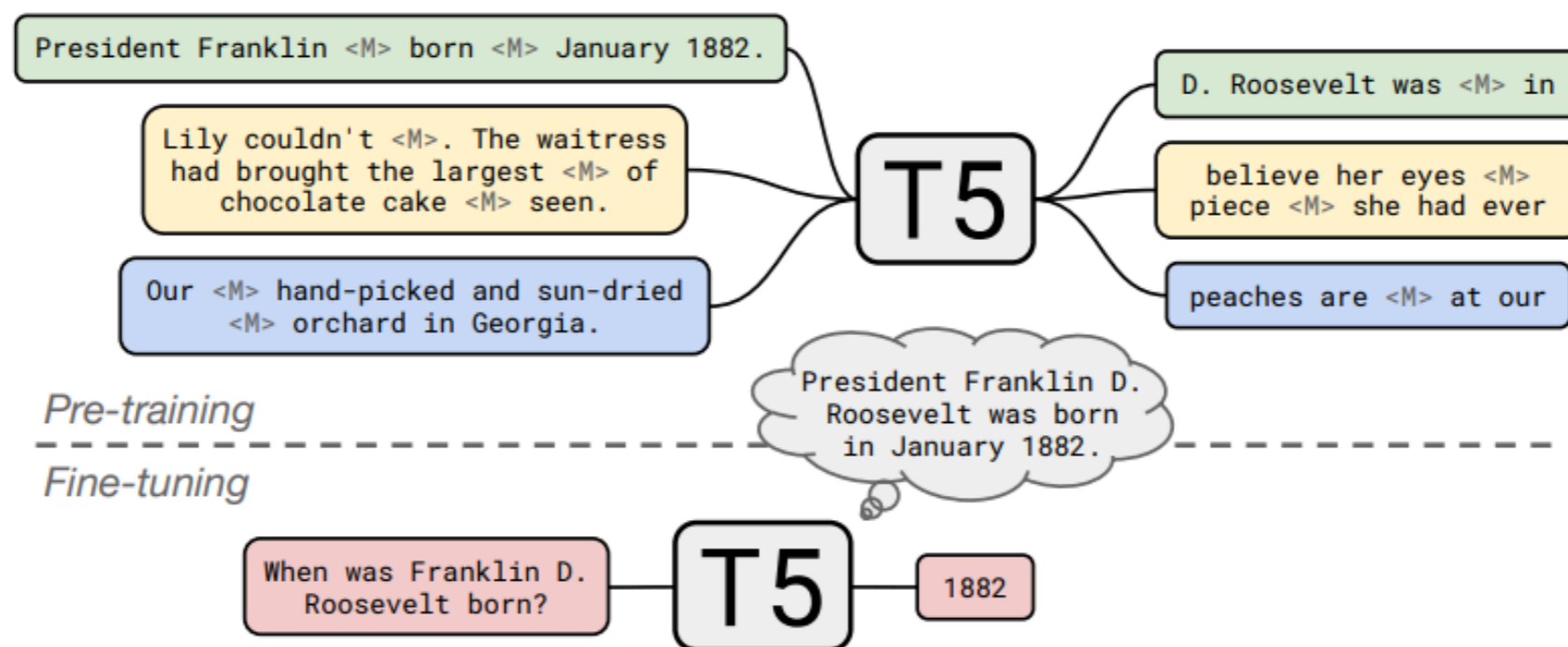
- **Entity linking:** Link named entities extracted from text to their corresponding KB entities
- Use Graph NNs to encode the graph where each node in a graph can either be a KB entity or the entity with textual context.



# Comparison between QA Methods

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

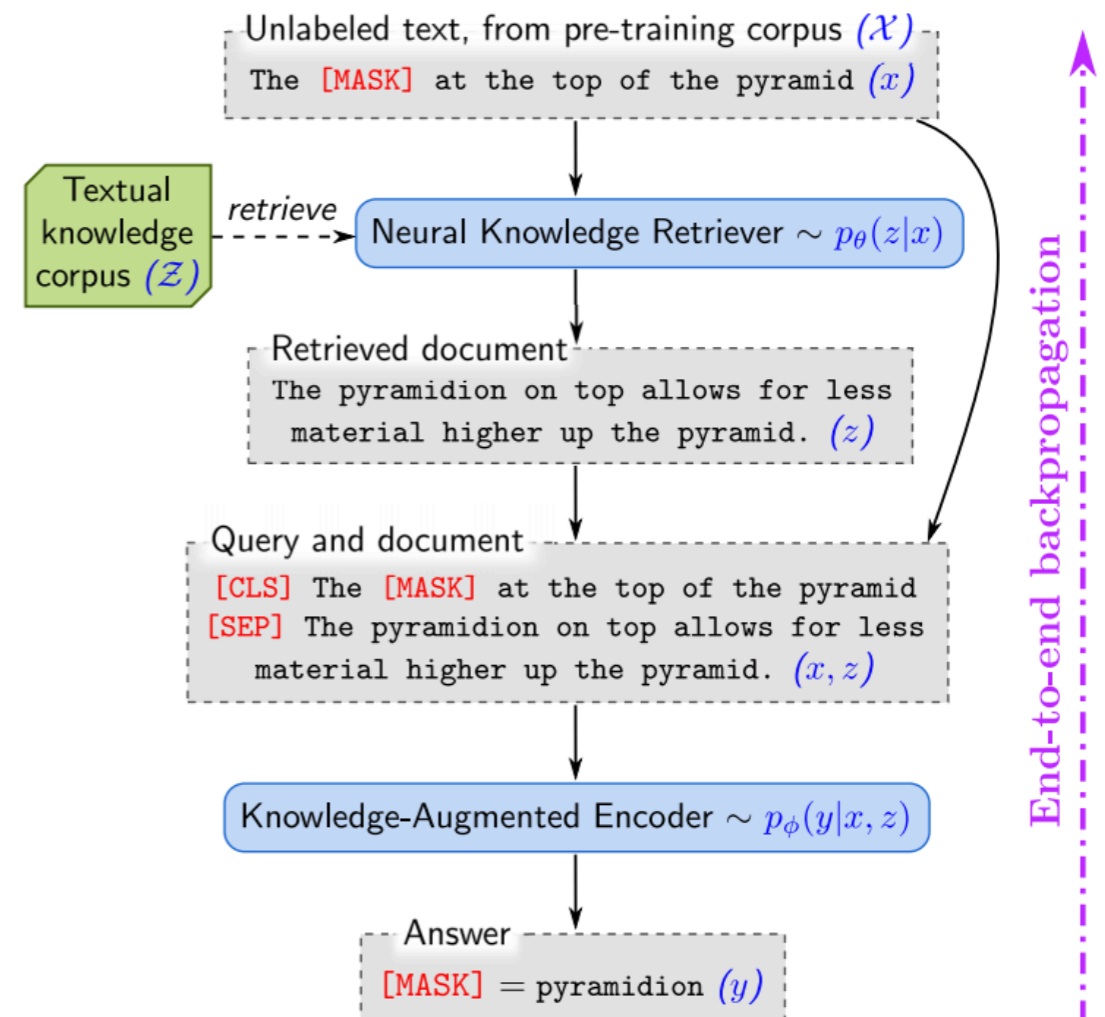
- Generate answers given questions without additional context.
- Performs even better than QA models with retrieved context.



# 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



# Comparison

- KBQA
  - Low coverage of knowledge
  - Faithful and interpretable
  - Dense structured
- TextQA
  - Wide coverage of knowledge
  - Misinformation
  - Massive raw texts
  - Enhanced with a text retrieval model
- LM-QA
  - Wide coverage of knowledge
  - Misinformation & out-dated information
  - Large model size
  - Black-box, not controllable

Questions?