

CS11-747 Neural Networks for NLP

# Learning From/For Knowledge Bases

Graham Neubig



**Carnegie Mellon University**

Language Technologies Institute

Site

<https://phontron.com/class/nn4nlp2017/>

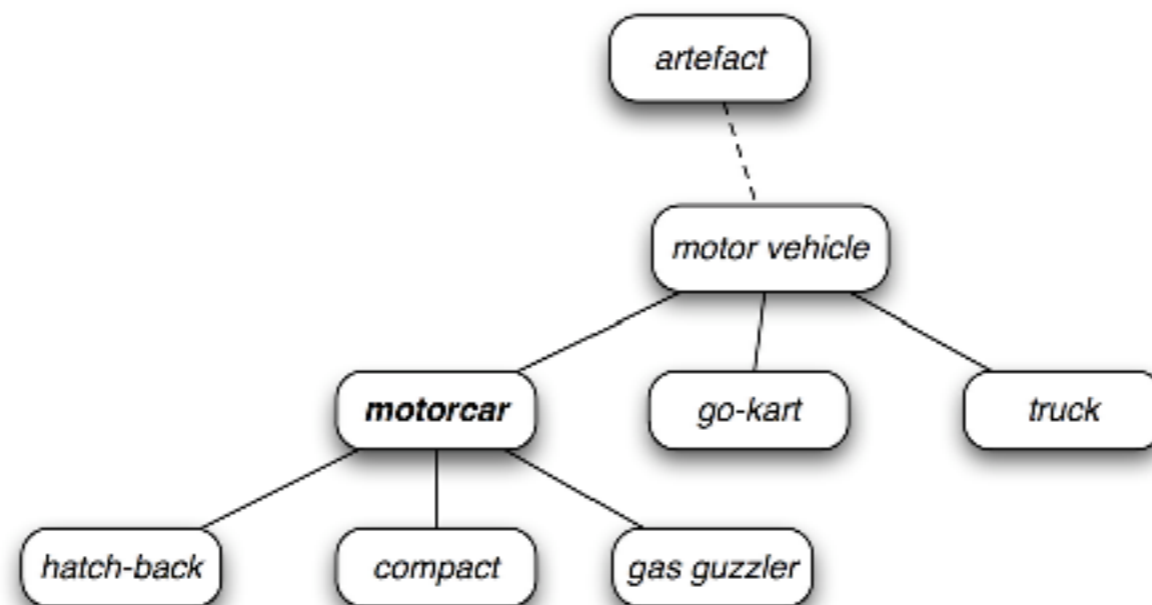
# Knowledge Bases

- 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?

# Types of Knowledge Bases

# WordNet (Miller 1995)

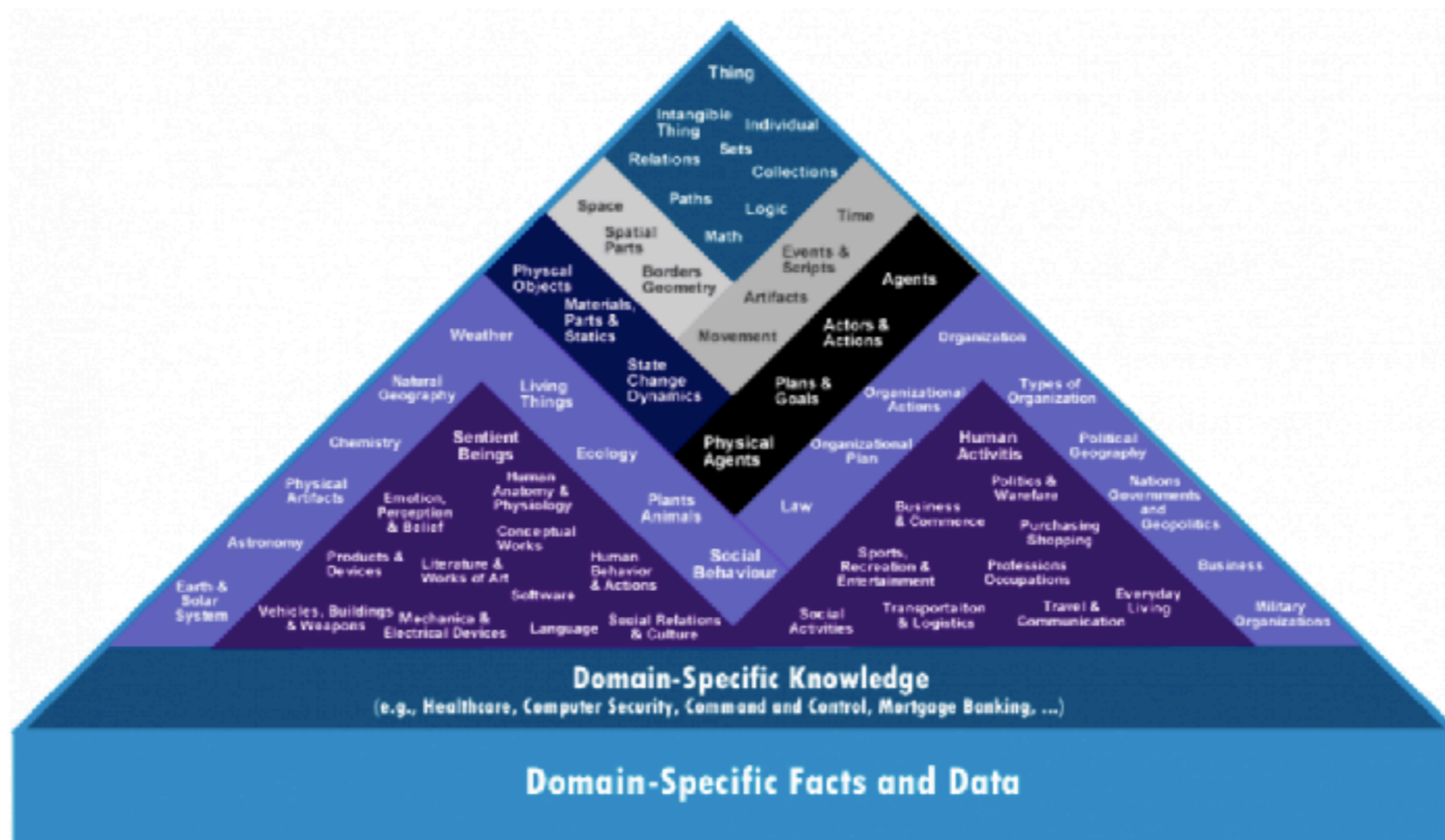
- WordNet is a large database of words including parts of speech, semantic relations



- 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** /kɑːrnɛɡi 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 hear: is in the work" (Andrew Carnegie)
<b>Type</b>	Private university
<b>Established</b>	1900 by Andrew Carnegie

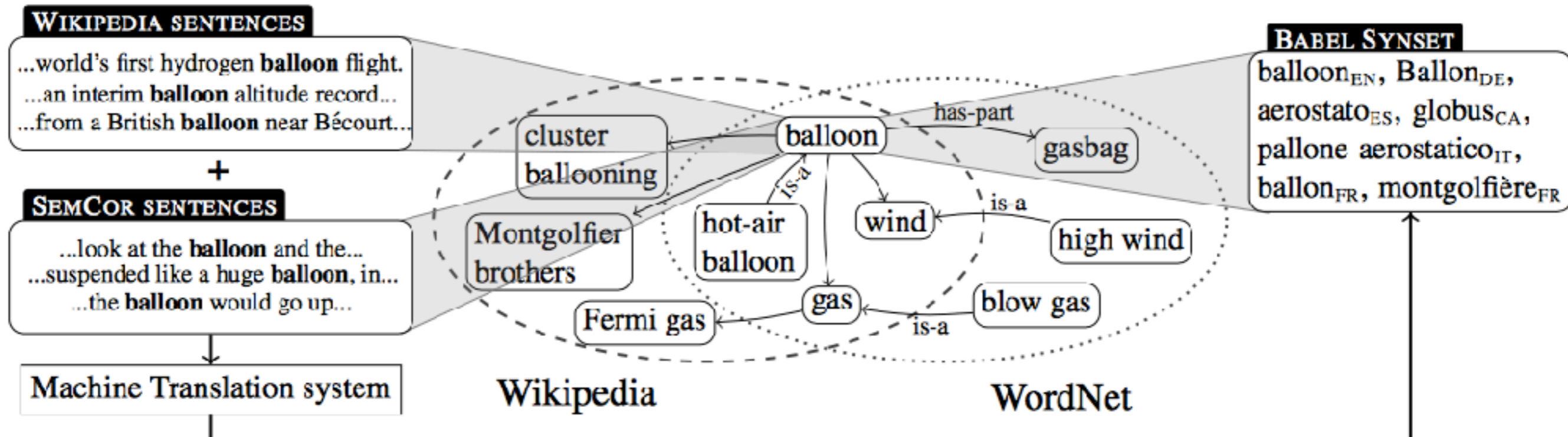
# YAGO (Suchanek et al. 2007)

- A meta-knowledge base, combining information from multiple sources (e.g. Wikipedia and WordNet)
- Expansions to include temporal/spatial information

# BabelNet

(Navigli and Ponzetto 2008)

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





# Freebase (Bollacker et al. 2008)

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

**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:**

- Joan Feynman (Richard Phillips Feynman)
- Ana Gasteyer (Richard Phillips Feynman)
- Gervase of Tilbury
- Alec Baldwin (Alexander Rae)
- Ernest Thesiger
- Mean Girls
- Riverside Drive
- Ferris Bueller's Day Off
- Television Personalities (The Television Personalities)

**Page History**  
Created by MetaWeb Oct 22, 2006  
Last edited by robert Oct 20, 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**  
like sex: sure, it may give some results, but that's not why we do it.  
I cannot 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!

**Description**

# WikiData

(Vrandečić and Krötzsch 2014)

- Knowledge base run by Wikimedia foundation and successor to FreeBase
- Incorporates many of the good points of previous work: multilingual, automatically extracted + curated, SPARQL interface

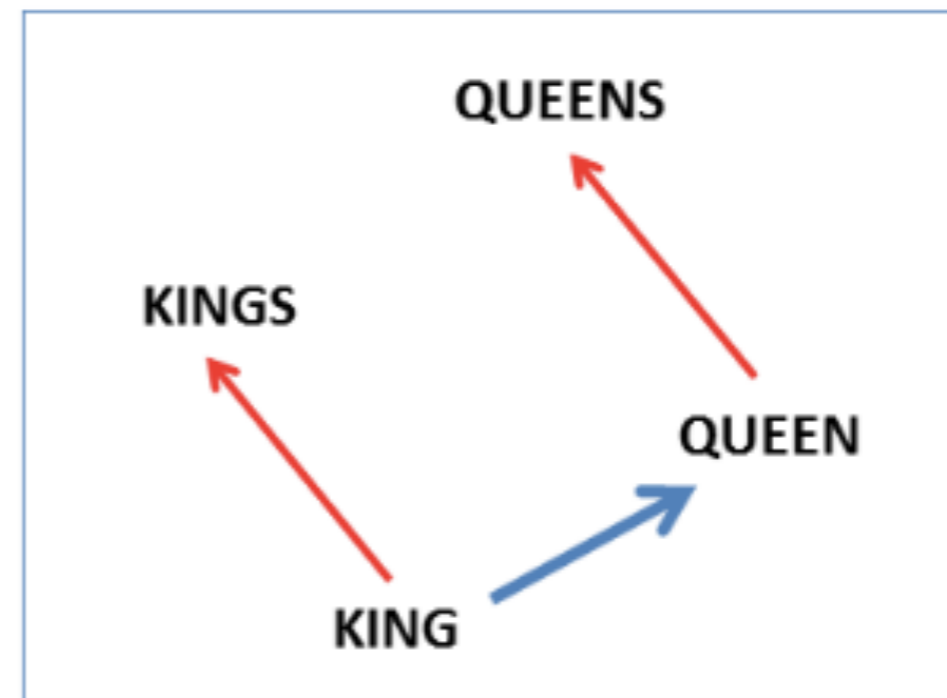
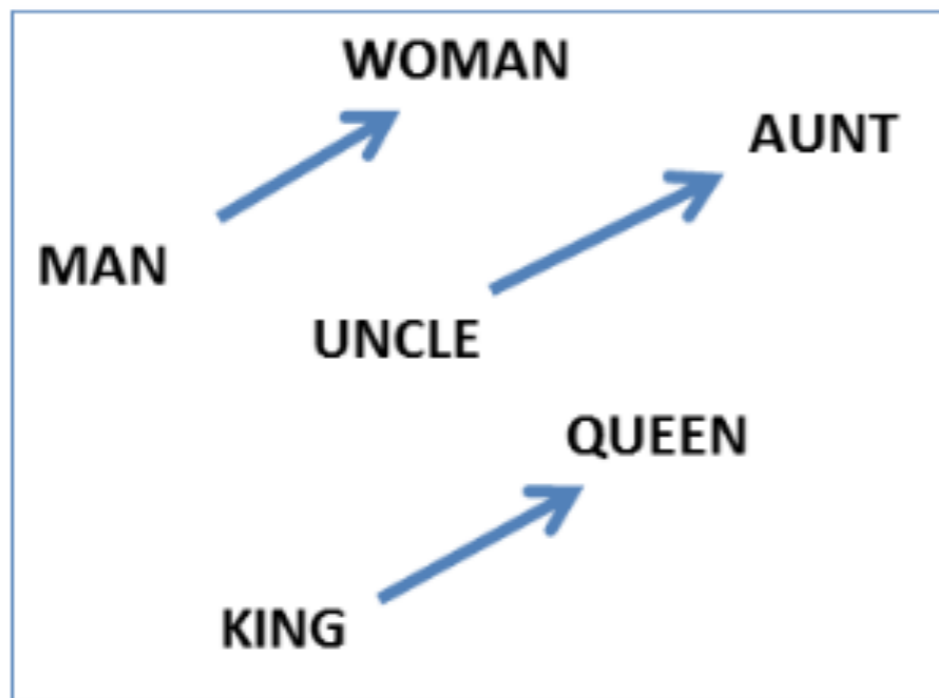
# Learning Relations from Embeddings

# 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

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

- 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 Relations from Embeddings (Bordes et al. 2013)

- Try to learn a transformation vector that shifts word embeddings based on their relation
- Optimize these vectors to minimize a margin-based loss

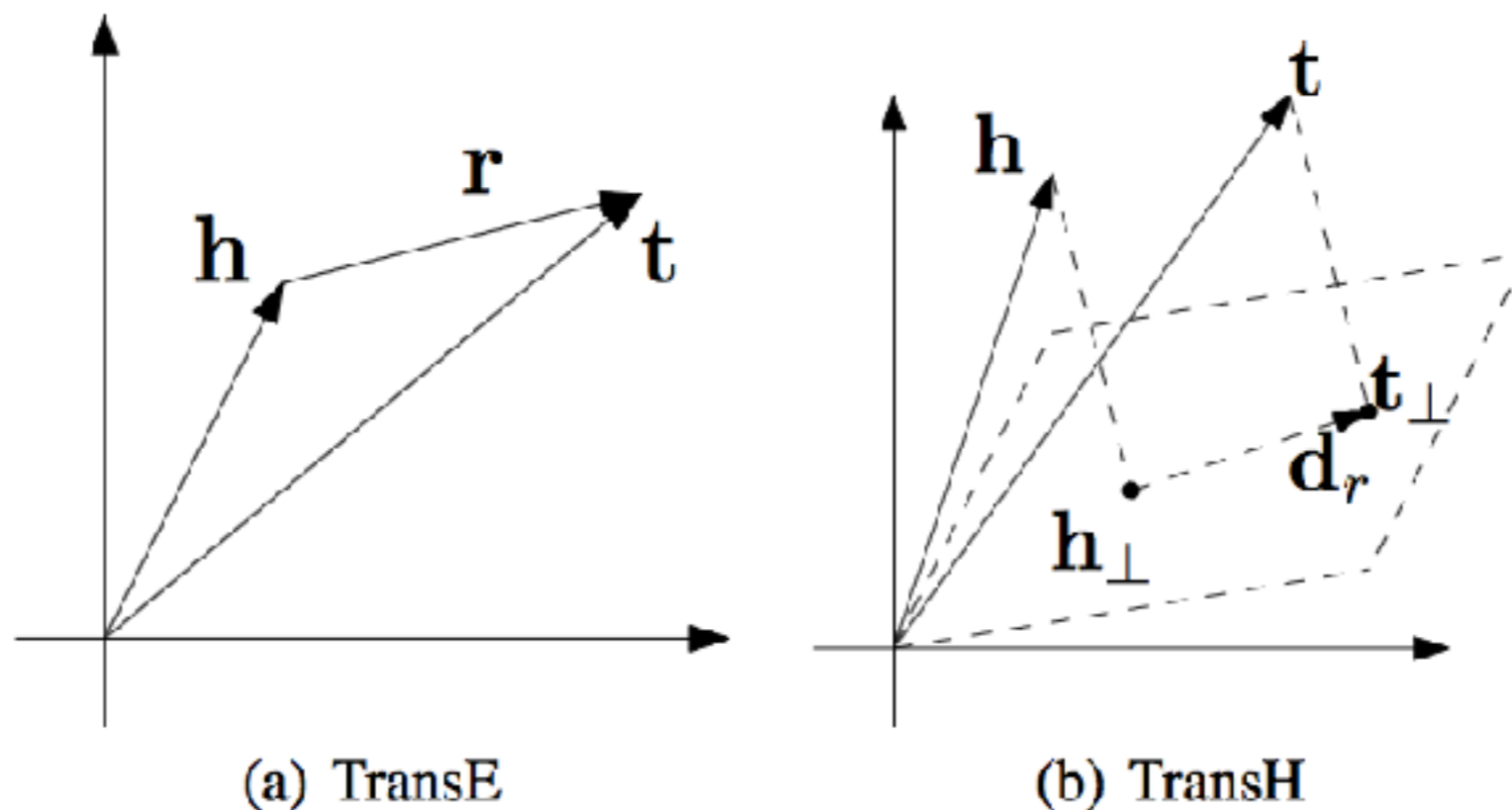
$$\mathcal{L} = \sum_{(h,\ell,t) \in S} \sum_{(h',\ell,t') \in S'_{(h,\ell,t)}} [\gamma + d(\mathbf{h} + \boldsymbol{\ell}, \mathbf{t}) - d(\mathbf{h}' + \boldsymbol{\ell}, \mathbf{t}')]_+$$

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

# Relation Extraction w/ Hyperplane

## Translation (Wang et al. 2014)

- Motivation: it is not realistic to assume that all dimensions are relevant to a particular relation
- Solution: project the word vectors on a hyperplane specifically for that relation, then verify relation



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

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



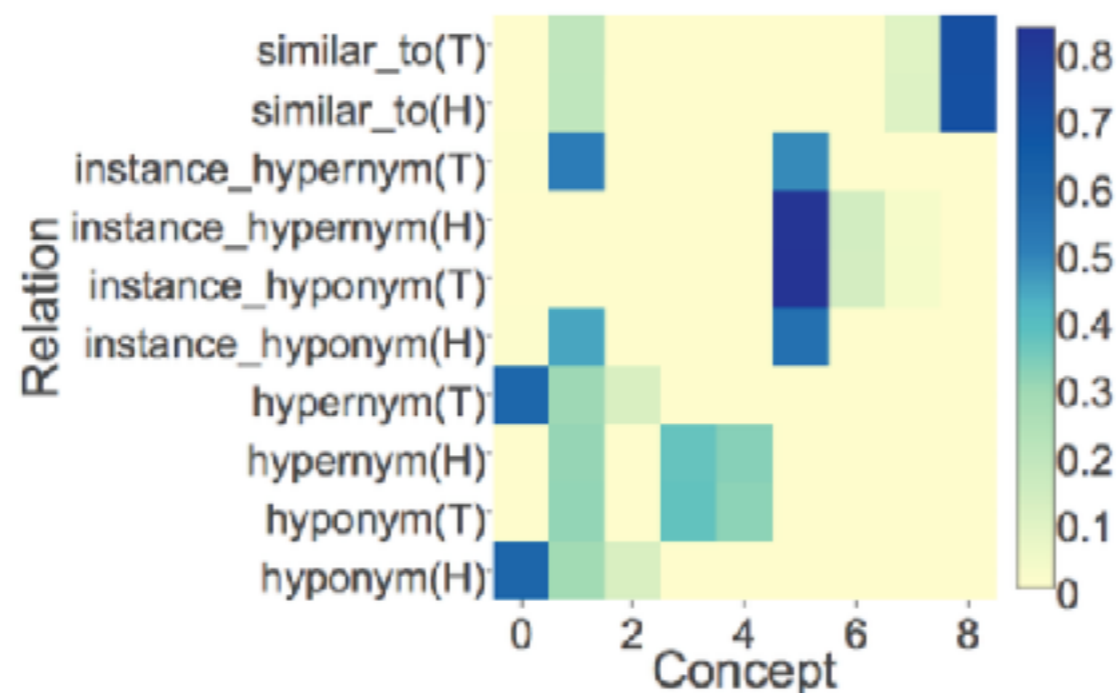
# Decomposable Relation

## 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  $\alpha$

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

- Better results, and also somewhat interpretable relations



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

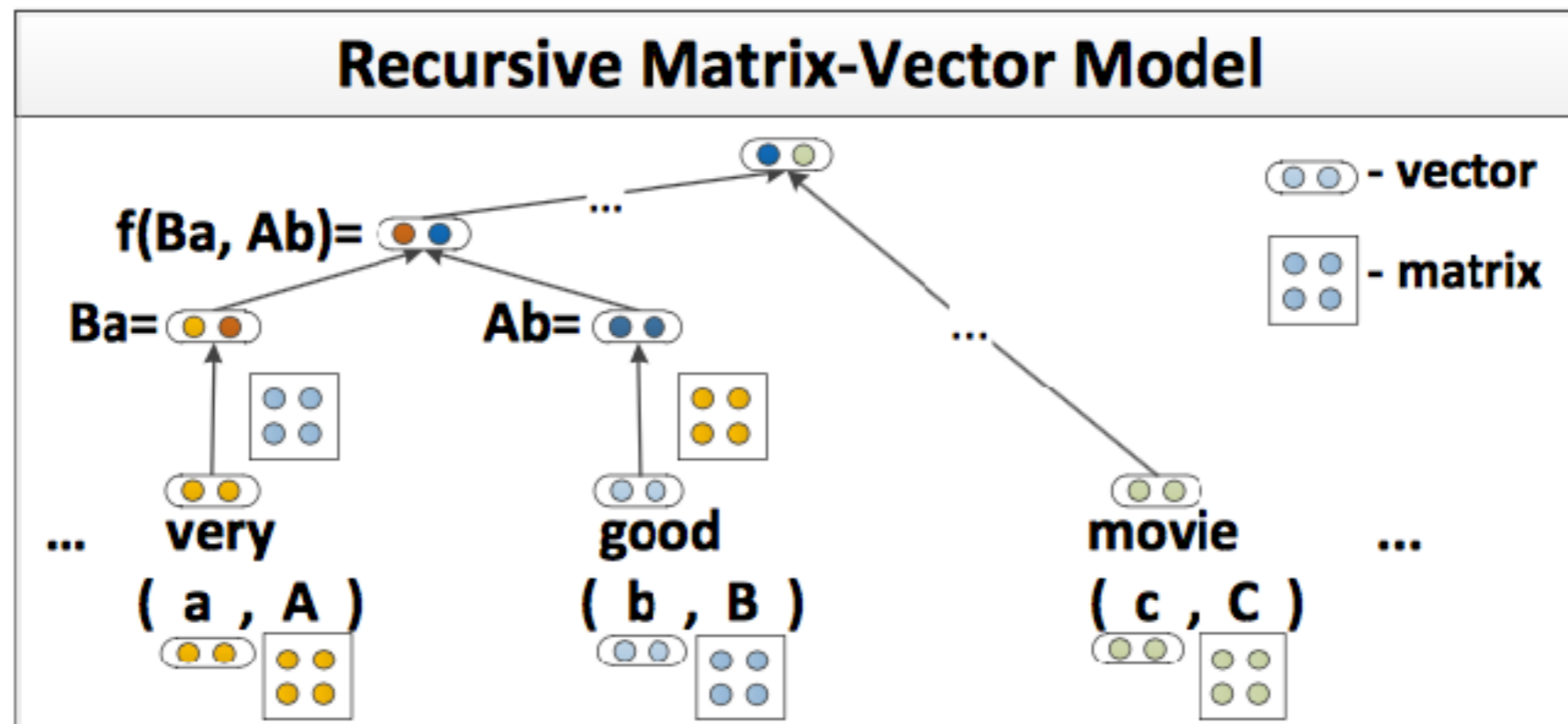
*[Steven Spielberg]'s film [Saving Private Ryan] is loosely based on the brothers' story.*

*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/ Recursive NNs (Socher et al. 2012)

- Create a syntax tree and do tree-structured encoding
- Classify the relation using the representation of the minimal constituent containing both words



# 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

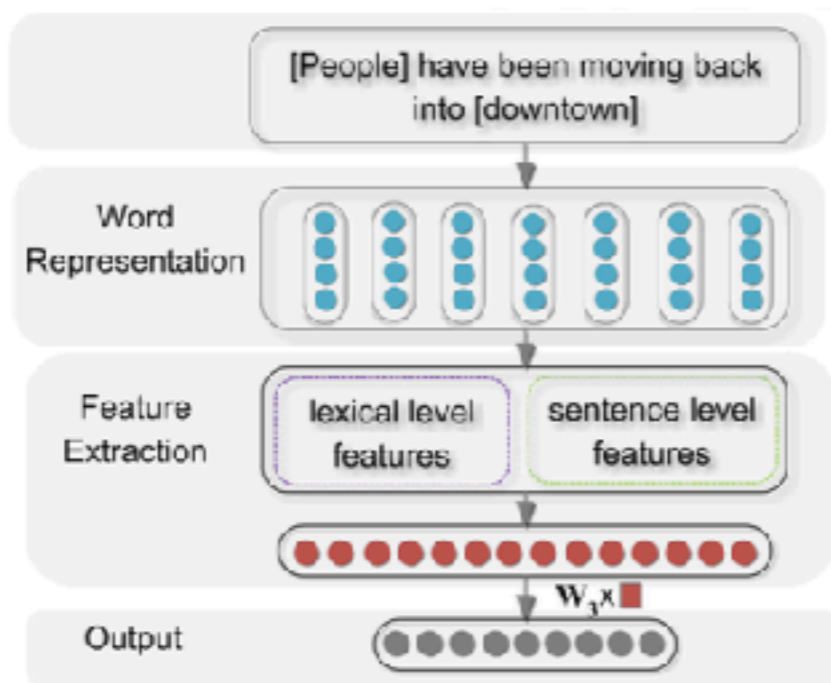


Figure 1: Architecture of the neural network used for relation classification.

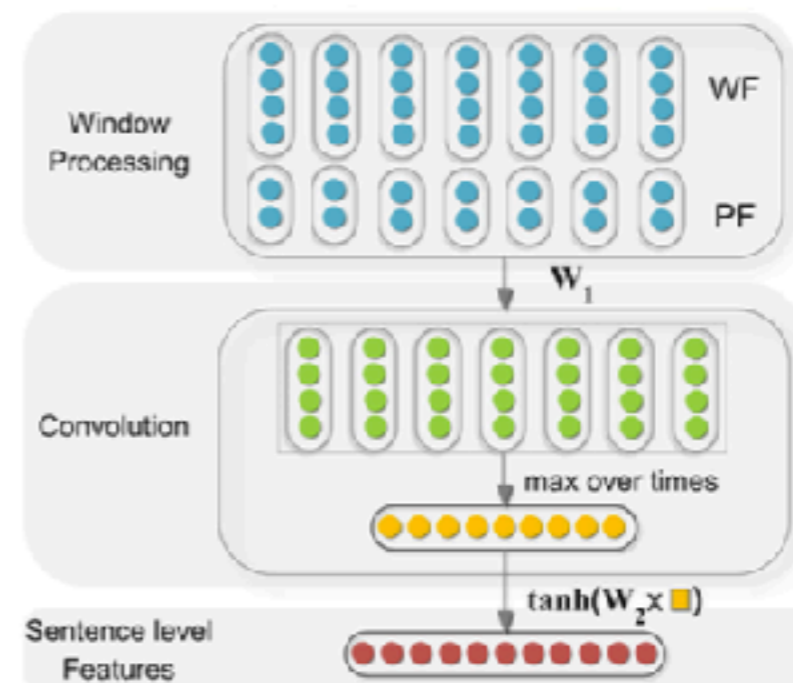


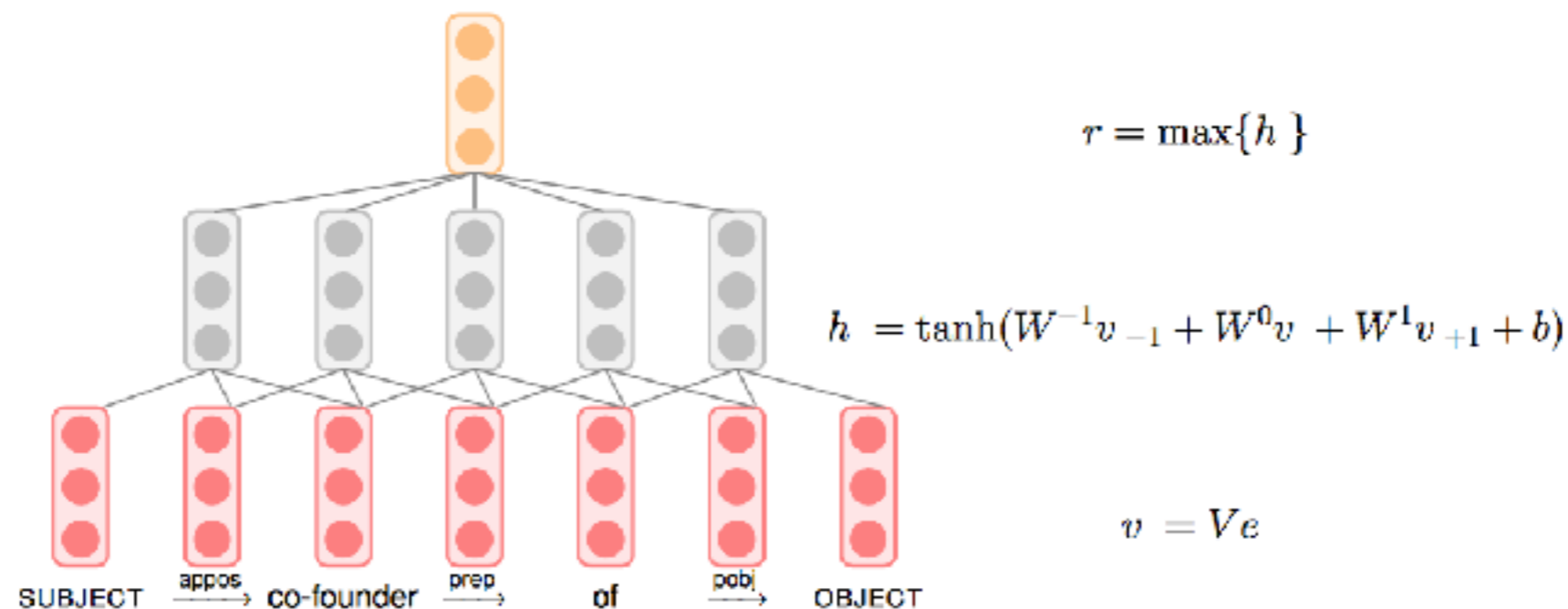
Figure 2: The framework used for extracting sentence level features.

# 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

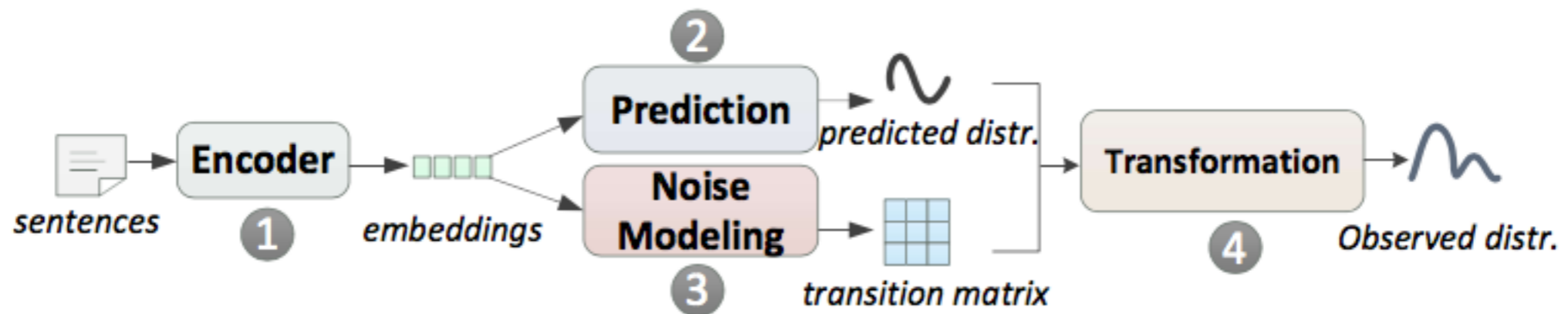
Textual Pattern	Count
SUBJECT $\xrightarrow{\text{appos}}$ founder $\xrightarrow{\text{prep}}$ of $\xrightarrow{\text{pobj}}$ OBJECT	12
SUBJECT $\xleftarrow{\text{nsubj}}$ co-founded $\xrightarrow{\text{dobj}}$ OBJECT	3
SUBJECT $\xrightarrow{\text{appos}}$ co-founder $\xrightarrow{\text{prep}}$ of $\xrightarrow{\text{pobj}}$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

- Model relations w/ CNN



# 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

# Learning from Relations Themselves



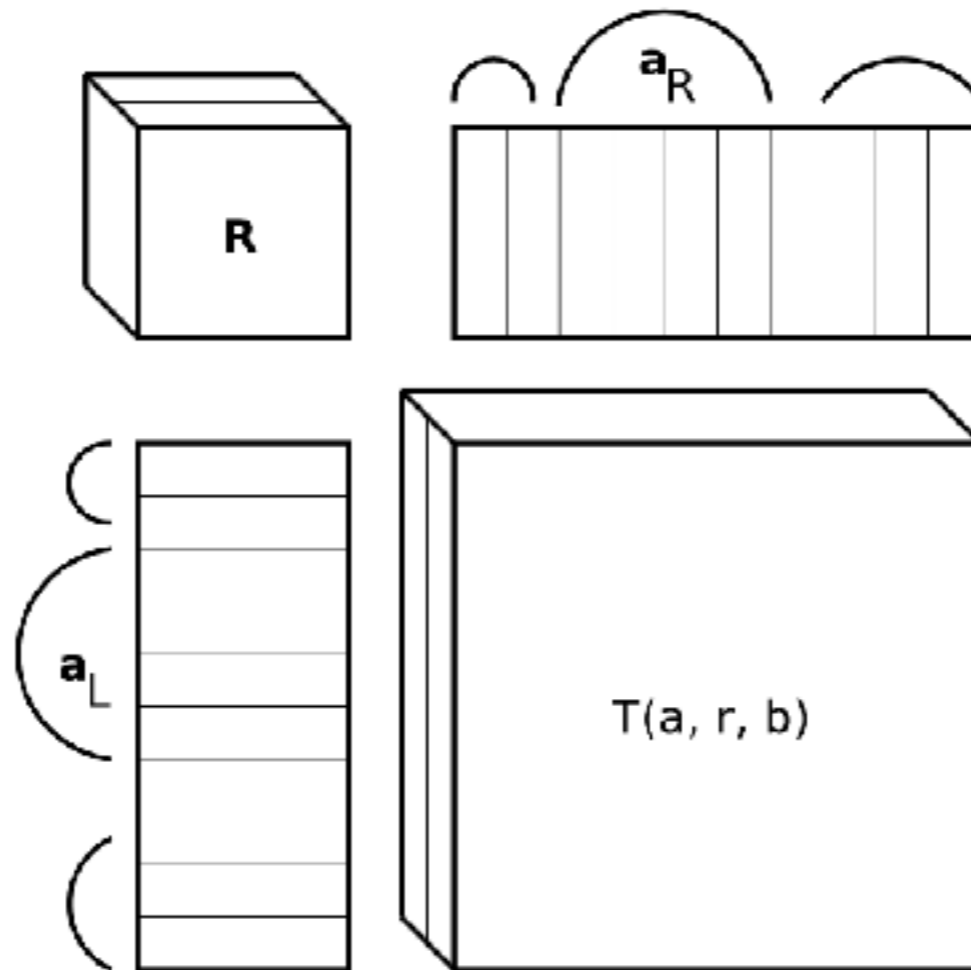
# Modeling Word Embeddings vs. Modeling Relations

- Word embeddings give information of the word in context, which is indicative of KB traits
- However, other relations (or combinations thereof) are also indicative

# 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”

---

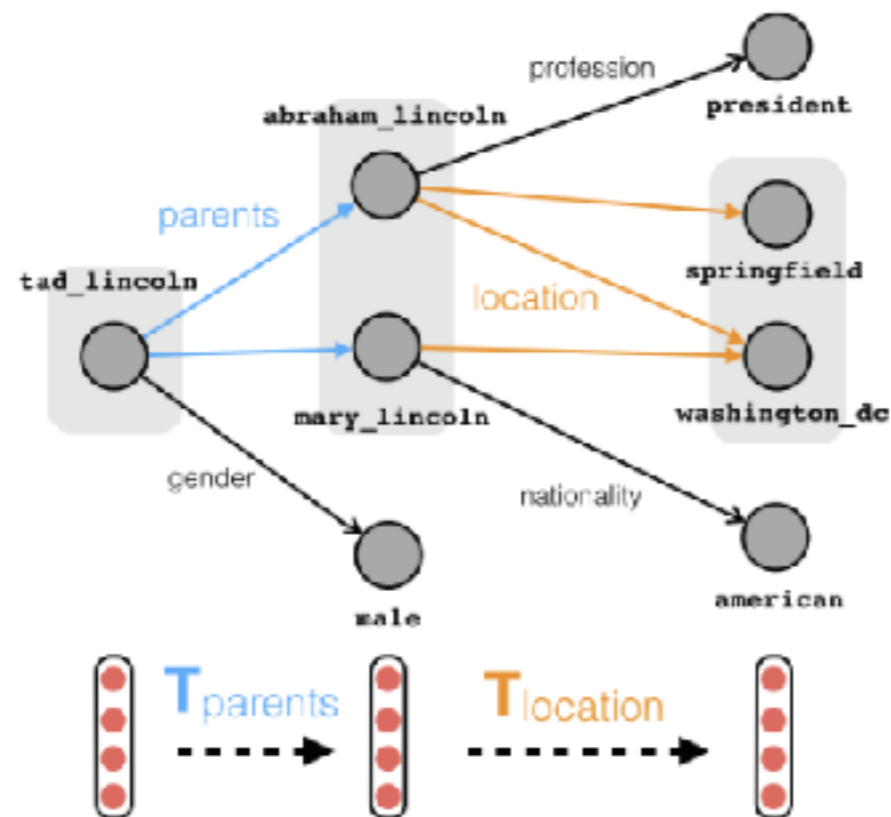
## ID Weight Feature

---

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

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

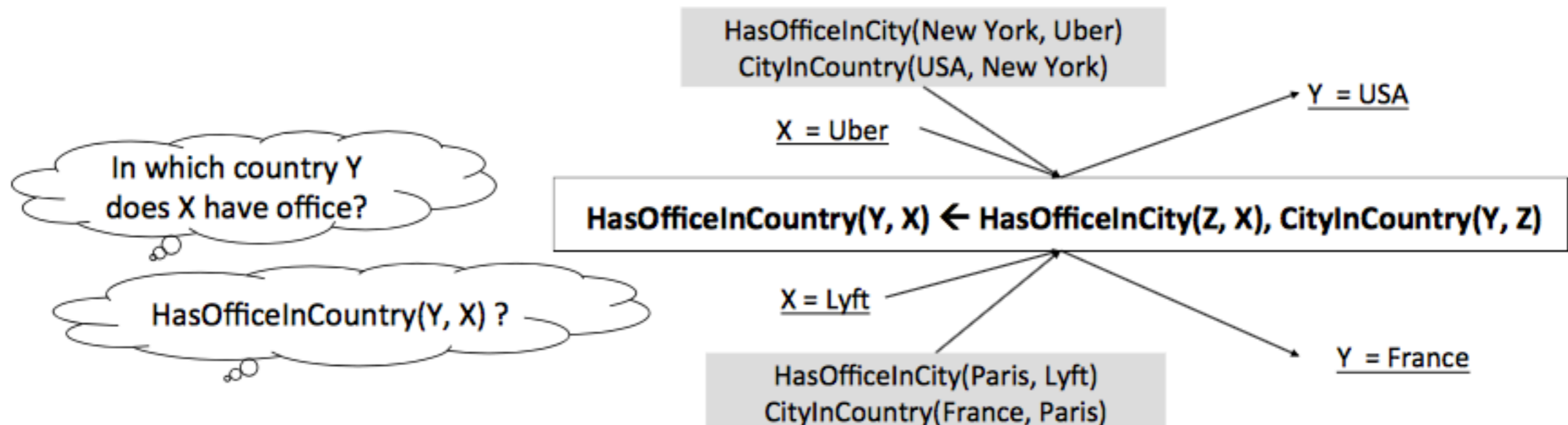
- Traveling over relations might result in error propagation
- Simple idea: optimize so that after traveling along a path, we still get the correct entity



# 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 a

$$\sum_l \alpha_l \prod_{\mathbf{k} \in \beta_l} \mathbf{M}_{\mathbf{R}_k}$$

# Using Knowledge Bases to Inform Embeddings

# Lexicon-aware Learning of Word Embeddings (e.g. Yu and Dredze 2014)

- Incorporate knowledge in the training objective for word embeddings
- Similar words should be in close places in the space

# 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

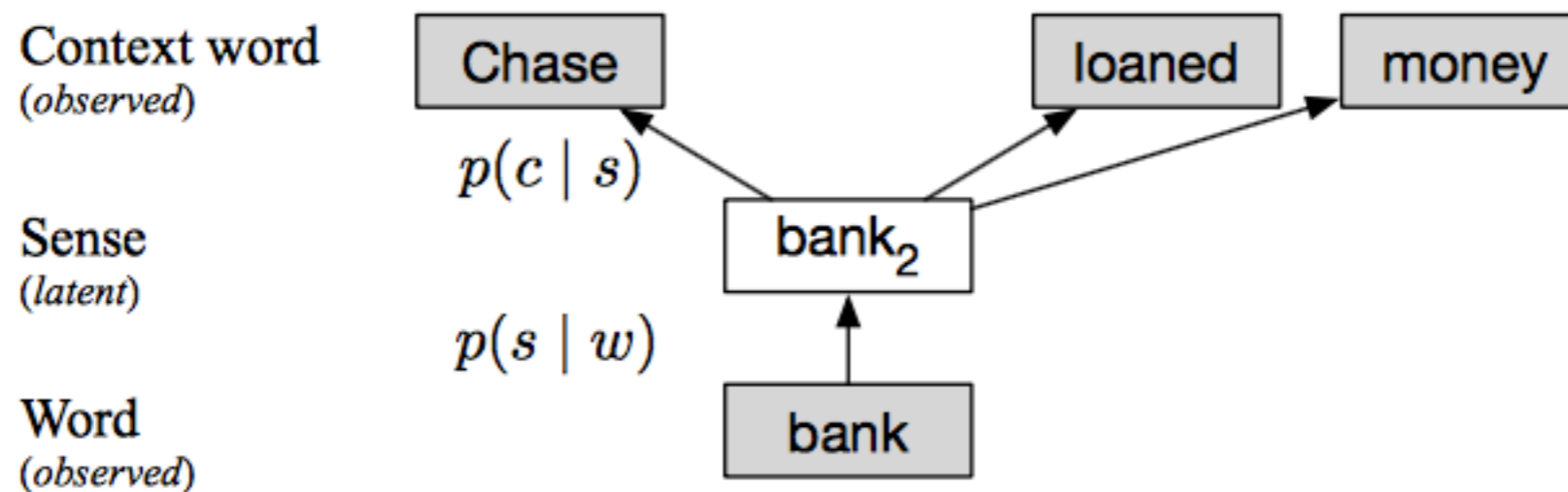
$$\Psi(Q) = \sum_{i=1}^n \left[ \alpha_i \|q_i - \hat{q}_i\|^2 + \sum_{(i,j) \in E} \beta_{ij} \|q_i - q_j\|^2 \right]$$

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



# Multi-sense Embedding w/ Lexicons (Jauhar et al. 2015)

- Create model with latent sense
- Sense can be optimized using EM or hard EM (select the most probable)



Questions?