Scalable Construction and Reasoning of Massive Knowledge Bases

Xiang Ren[1] **Nanyun Peng**[1] **William Yang Wang**[2]

University of Southern California[1]

University of California, Santa Barbara[2]

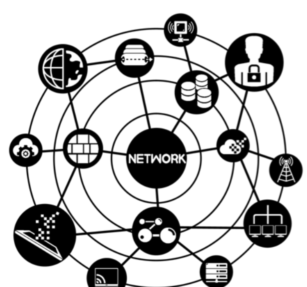
Turning Unstructured Text Data into Structures

Unstructured Text Data

(account for ~80% of all data in organizations)

Knowledge & Insights

Structures



STORE		
Store_key	City	Region
1	New York	East
2	Chicago	Central
3	Atlanta	East
4	Los Angeles	West
5	San Francisco	West
6	Philadelphia	East
.	.	.
.	.	.

PRODUCT		
Product_key	Description	Brand
1	Beautiful Girls	MKF Studios
2	Toy Story	Wolf
3	Sense and Sensibility	Parabuster Inc.
4	Holiday of the Year	Wolf
5	Pulp Fiction	MKF Studios
6	The Juror	MKF Studios
7	From Dusk Till Dawn	Parabuster Inc.
8	Hellraiser: Bloodline	Big Studios
.	.	.
.	.	.

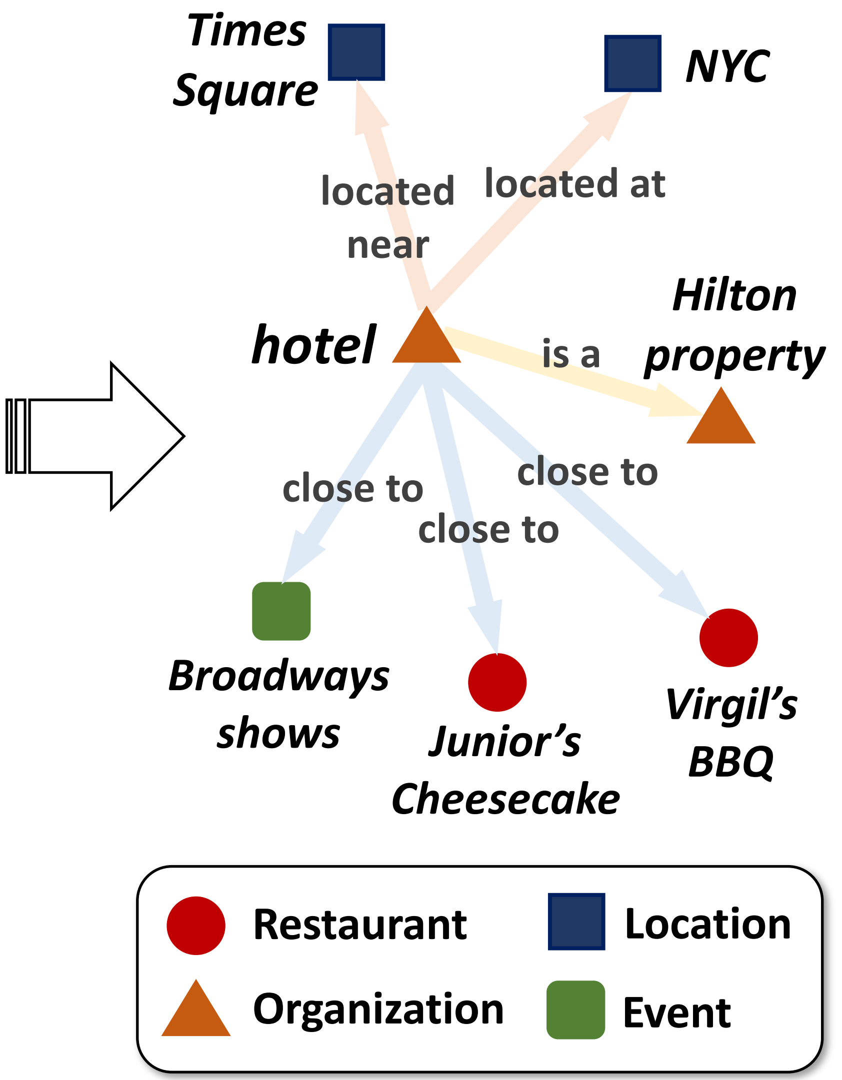
SALES_FACT				
Store_key	Product_key	Sales	Cost	Profit
1	6	2.39	1.15	1.24
1	2	16.7	6.91	9.79
2	7	7.16	2.75	4.40
3	2	4.77	1.84	2.93
5	3	11.93	4.59	7.34
5	1	14.31	5.51	8.80
.	.	.	.	.
.	.	.	.	.

(Chakraborty, 2016)

Reading the reviews: From Text to Structured Facts

This **hotel** is my favorite **Hilton property** in **NYC**! It is located right on 42nd street near **Times Square**, it is close to all subways, **Broadways shows**, and next to great restaurants like **Junior's Cheesecake**, **Virgil's BBQ** and many others.

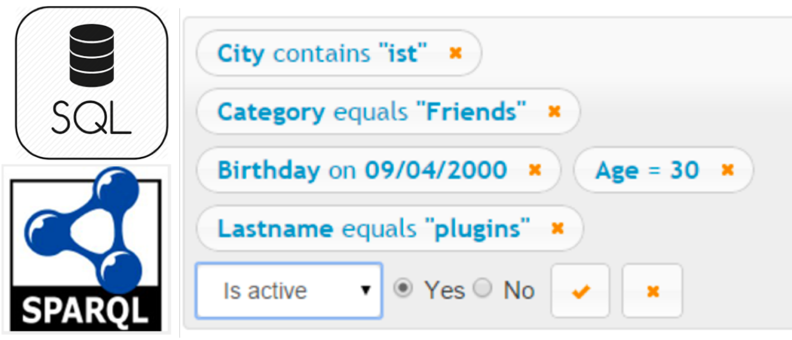
-- *TripAdvisor*



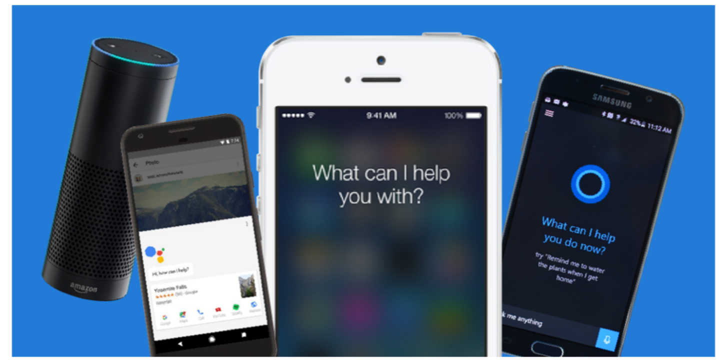
Structured Facts
1. "Typed" entities
2. "Typed" relationships

Why Text to Structures?

Structured Search & Exploration



Dialog Systems

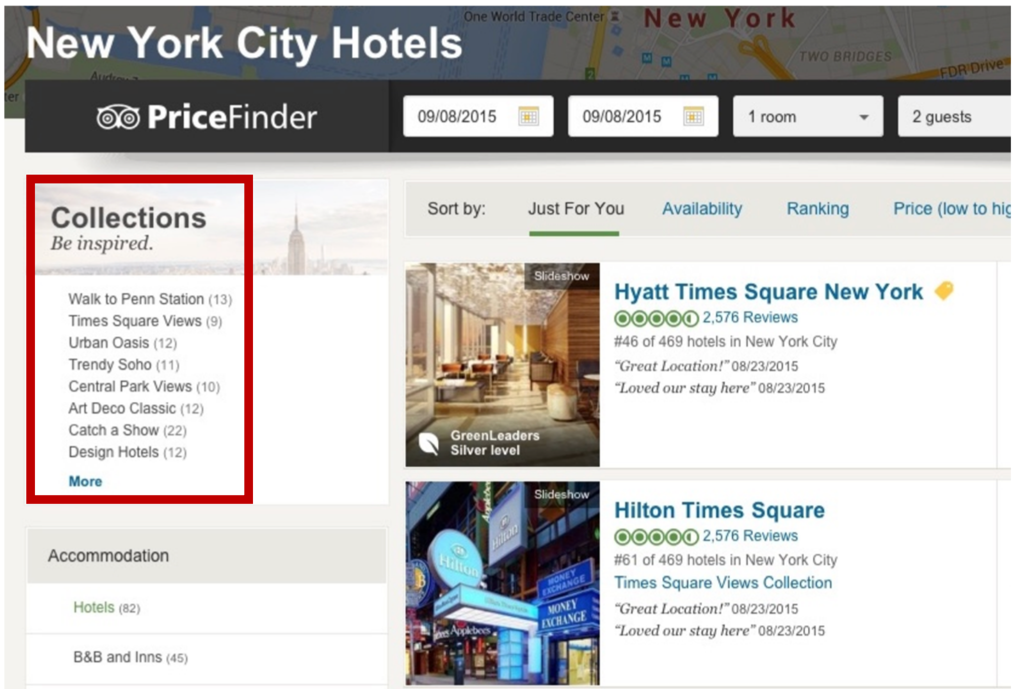


Question Answering

Scientific Inference

A Product Use Case: Finding "Interesting Hotel Collections"

Technology Transfer to TripAdvisor



Grouping hotels based on structured facts extracted from the review text

Features for "**Catch a Show**" collection

1	broadway shows
2	beacon theater
3	broadway dance center
4	broadway plays
5	david letterman show
6	radio city music hall
7	theatre shows

Features for "**Near The High Line**" collection

1	high line park
2	chelsea market
3	highline walkway
4	elevated park
5	meatpacking district
6	west side
7	old railway

http://engineering.tripadvisor.com/using-nlp-to-find-interesting-collections-of-hotels/

A Scientific Use Case: Precision Medicine

Molecular tumor board

www.ucsf.edu/news/2014/11/120451/bridging-gap-precision-medicine

Problem: Hard to scale

U.S. 2016: 1.7 million new cases, 600K deaths

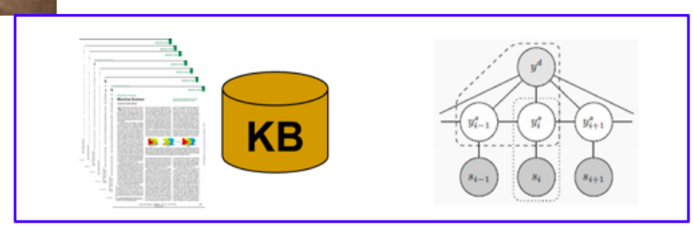
902 cancer hospitals

Memorial Sloan Kettering

Sequence: Tens of thousand

Board can review: A few hundred

Machine Reading



Predict Drug Combo

Azure

Better Structured Search with Reasoning Capabilities

Text to Structures: **Applications**

Technology Transfer

ARL
Microsoft
NIH
MAYO CLINIC
bing
yelp
tripadvisor

Intelligent Personal Assistant

Cortana. Siri amazon echo Google now Facebook M

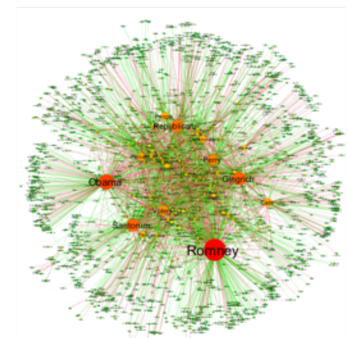
Online Education

coursera
UDACITY Learn. Think. Do.
edX
canvas NETWORK
NovoED
iversity
OPEN 2 STUDY
Future Learn

Medical records
Scientific papers
Clinical reports
...

Healthcare

Social media posts
Web blogs
News articles
...

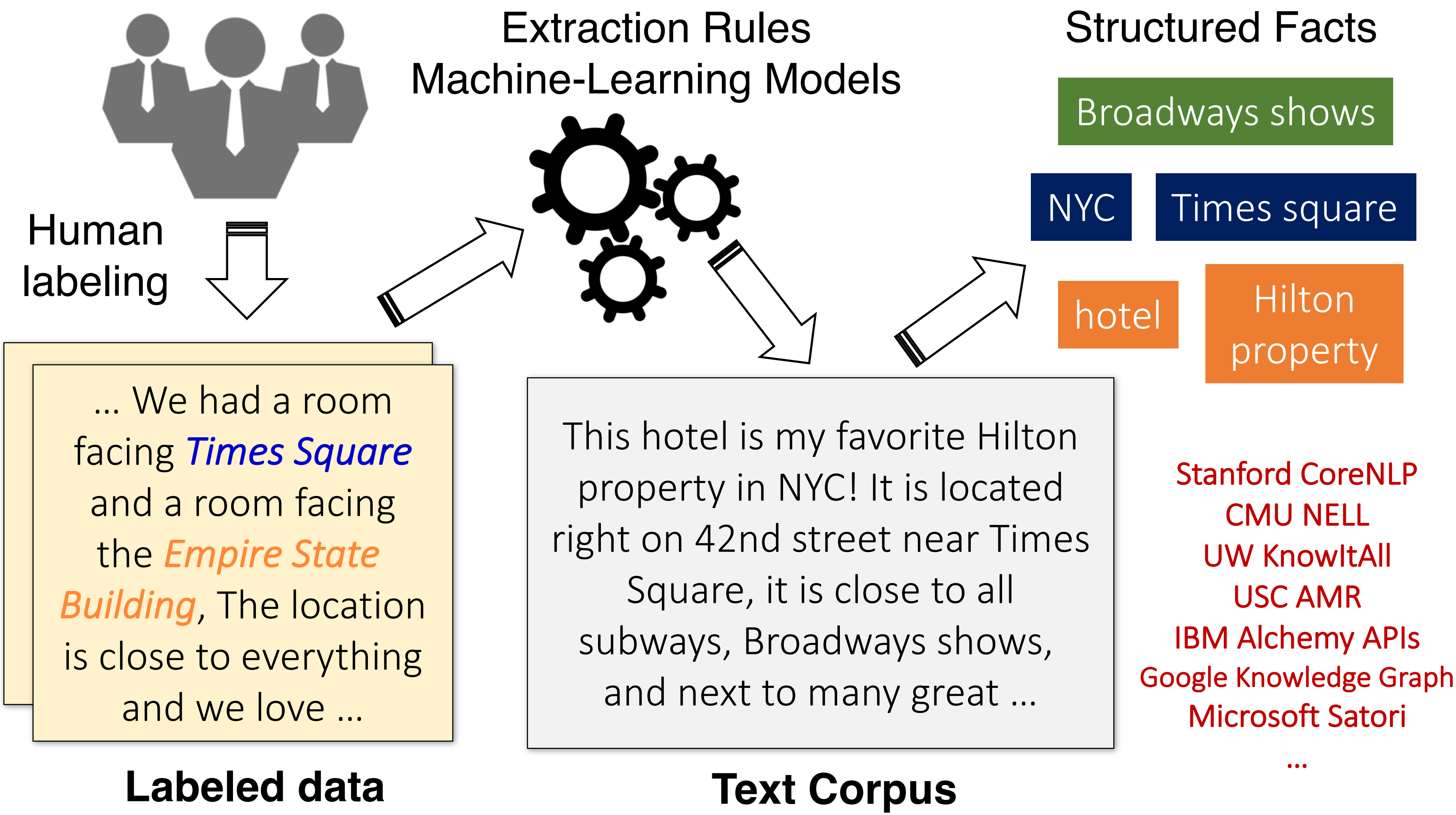


Computational Social Sciences

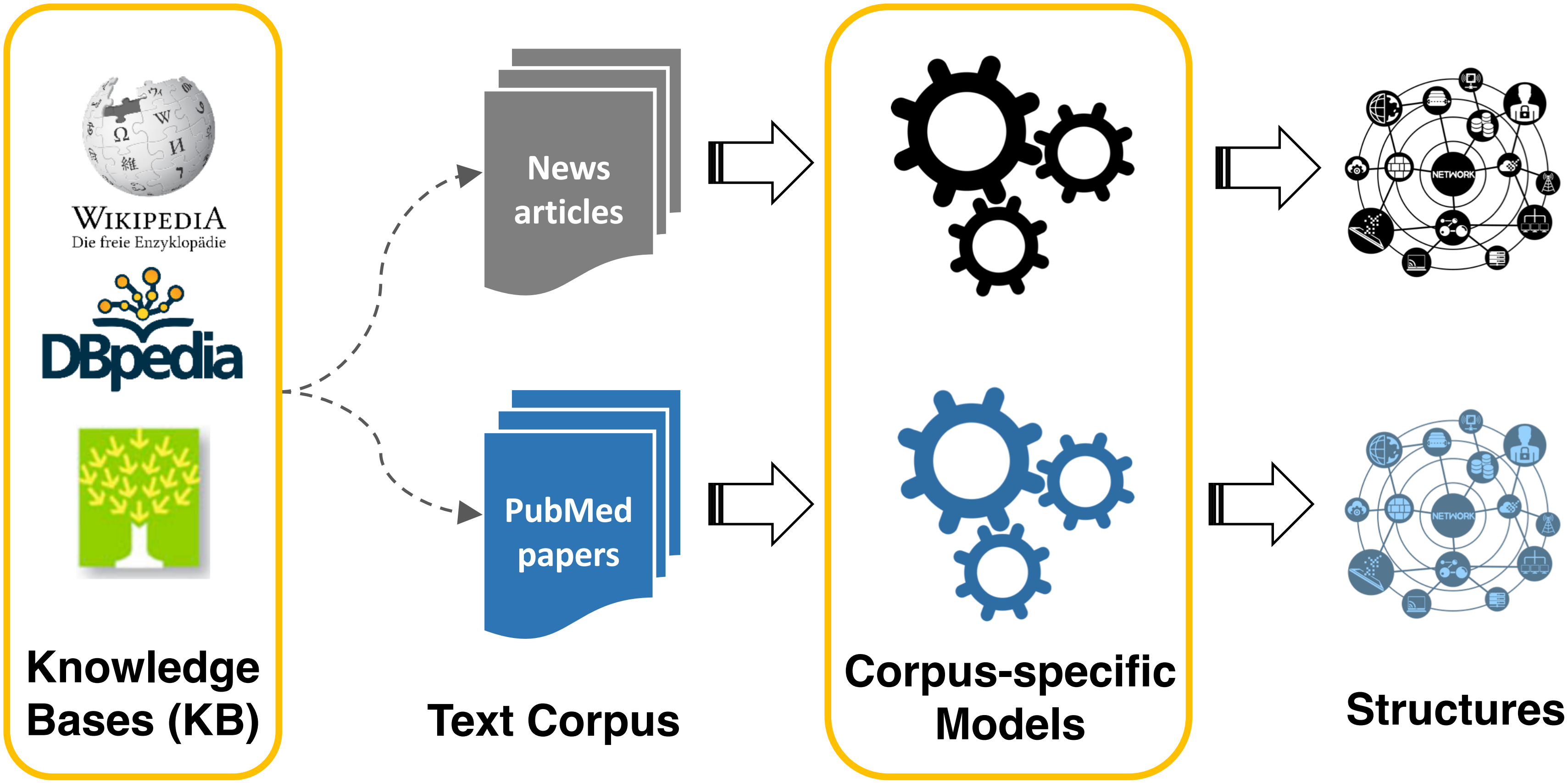
Corporate reports
News streams
Customer reviews
...

Business Intelligence

Prior Art: Extracting Structures with Repeated Human Effort

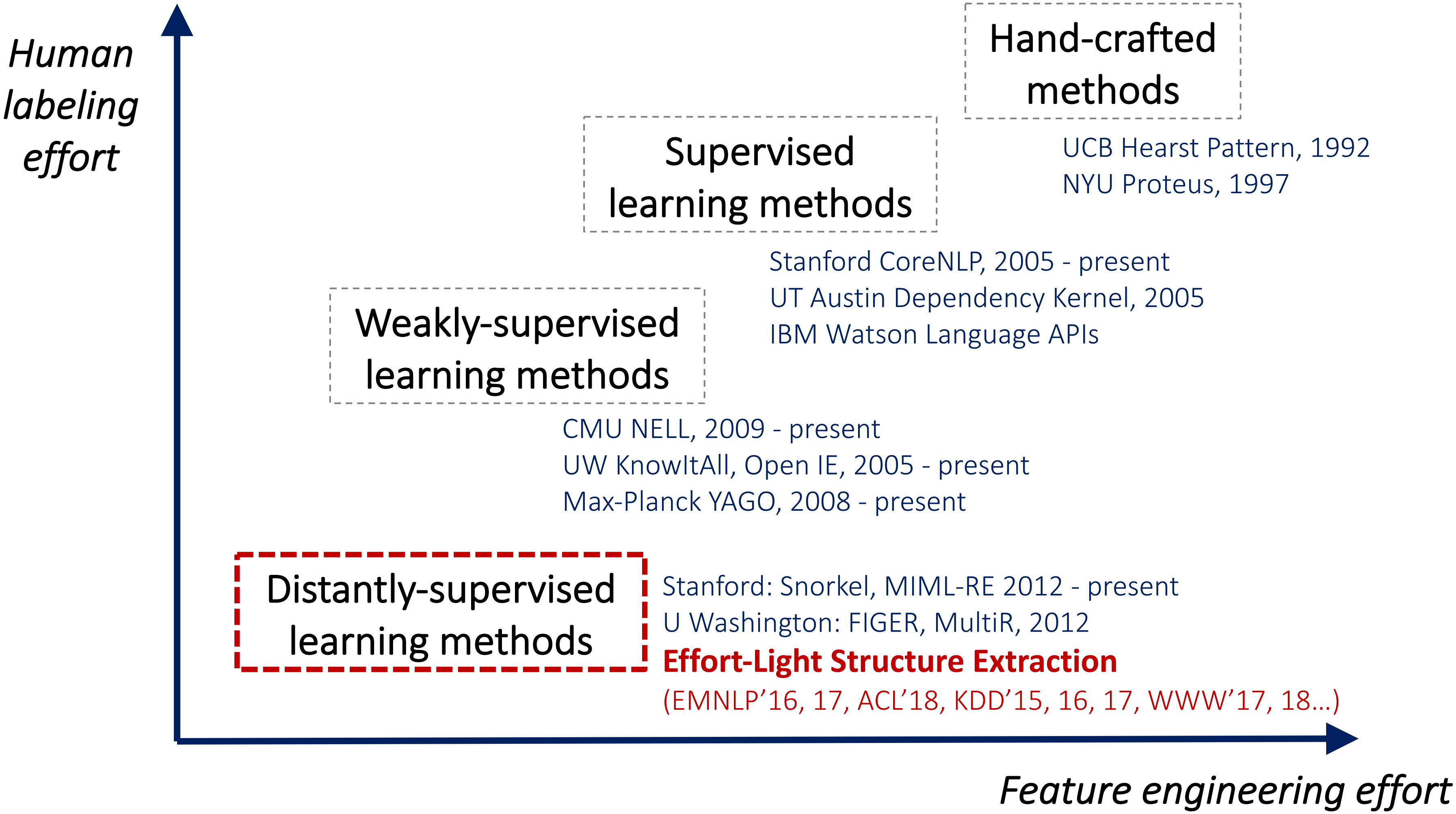


Effort-Light Structure Extraction



- Enables ***quick*** development of applications over various corpora
- Extracts ***complex*** structures without introducing human error

Effort–Light Structure Extraction : **Where Are We?**

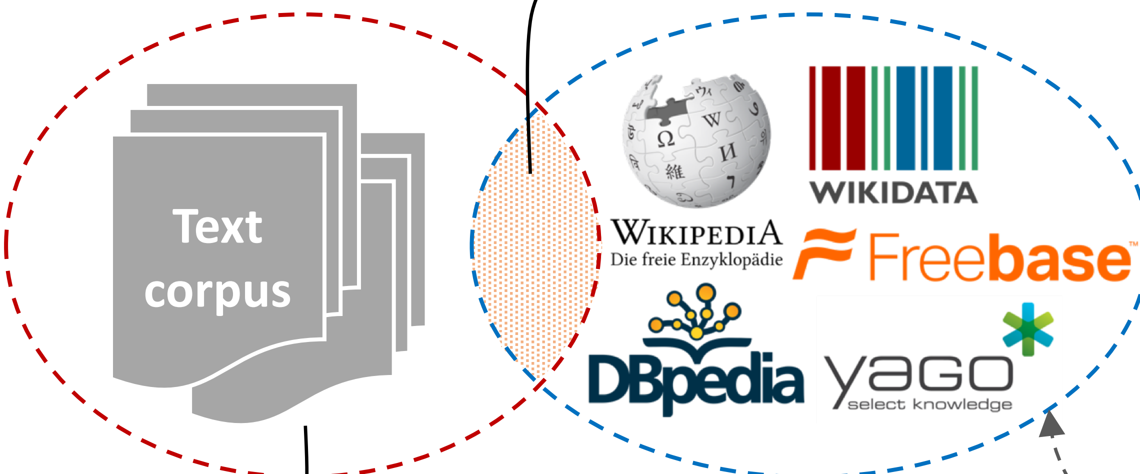


“Distant” Supervision: **What Is It?**

“Matchable” structures: entity names, entity types, typed relationships ...

Freely available!

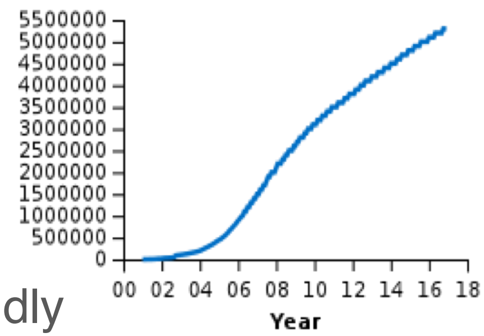
- Common knowledge
- Life sciences
- Art …



Knowledge Bases

“Un-matchable”

Number of Wikipedia articles



Rapidly growing!

Human crowds

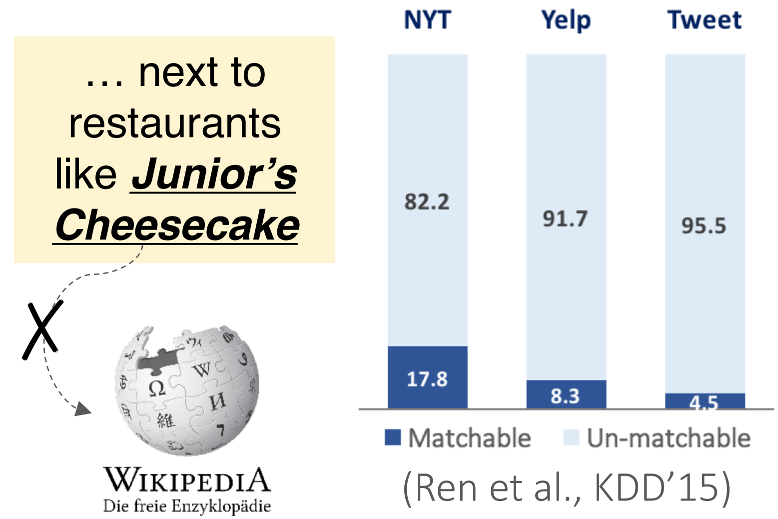
(Mintz et al., 2009) , (Riedek et al., 2010), (Lin et al., 2012), (Ling et al., 2012), (Surdeanu et al., 2012), (Xu et al., 2013), (Nagesh et al., 2014), …

https://en.wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia

Learning with Distant Supervision: **Challenges**

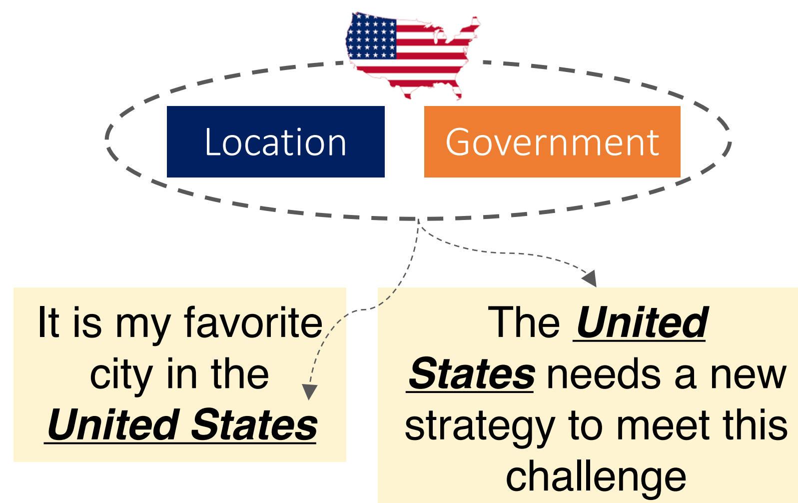
1. Sparsity of “Matchable”

- Incomplete knowledge bases
- Low-confidence matching

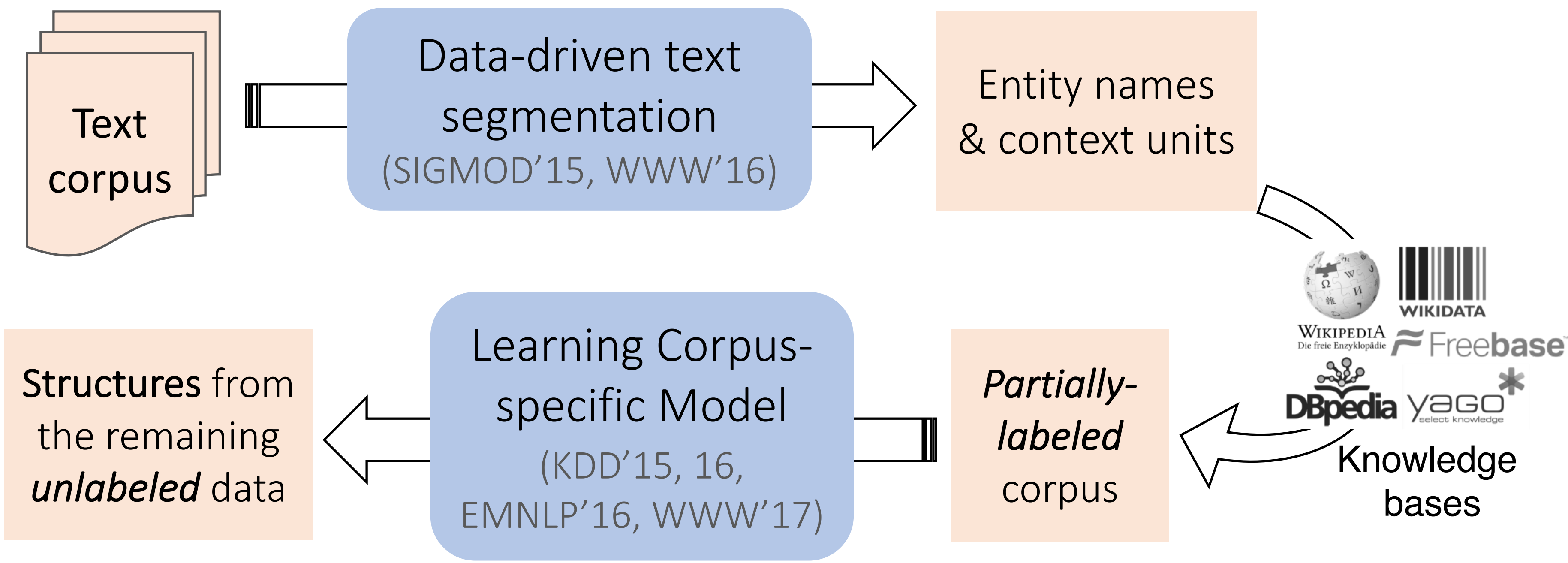


2. Accuracy of “Expansion”

- For “matchable”: *Are all the labels assigned accurately?*
- For “un-matchable”: *How to perform inference accurately?*



Effort-Light StructMine: **Methodology**



IE

IE

IE

Structured Knowledge			
Entity	Entity	Entity	Relation
T790M	EGFR	gefitinib	Resist
Obama	U.S.		President_of
…	…		…

IE

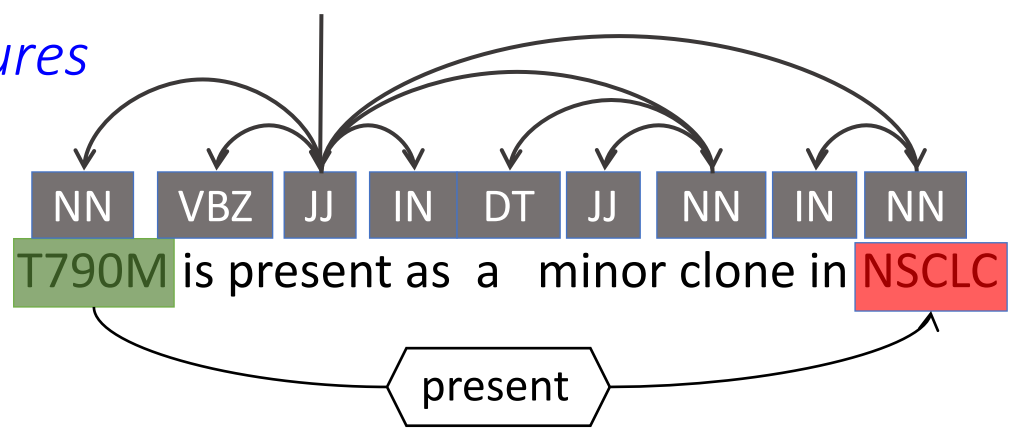
IE

Challenges of Obtaining Training Data

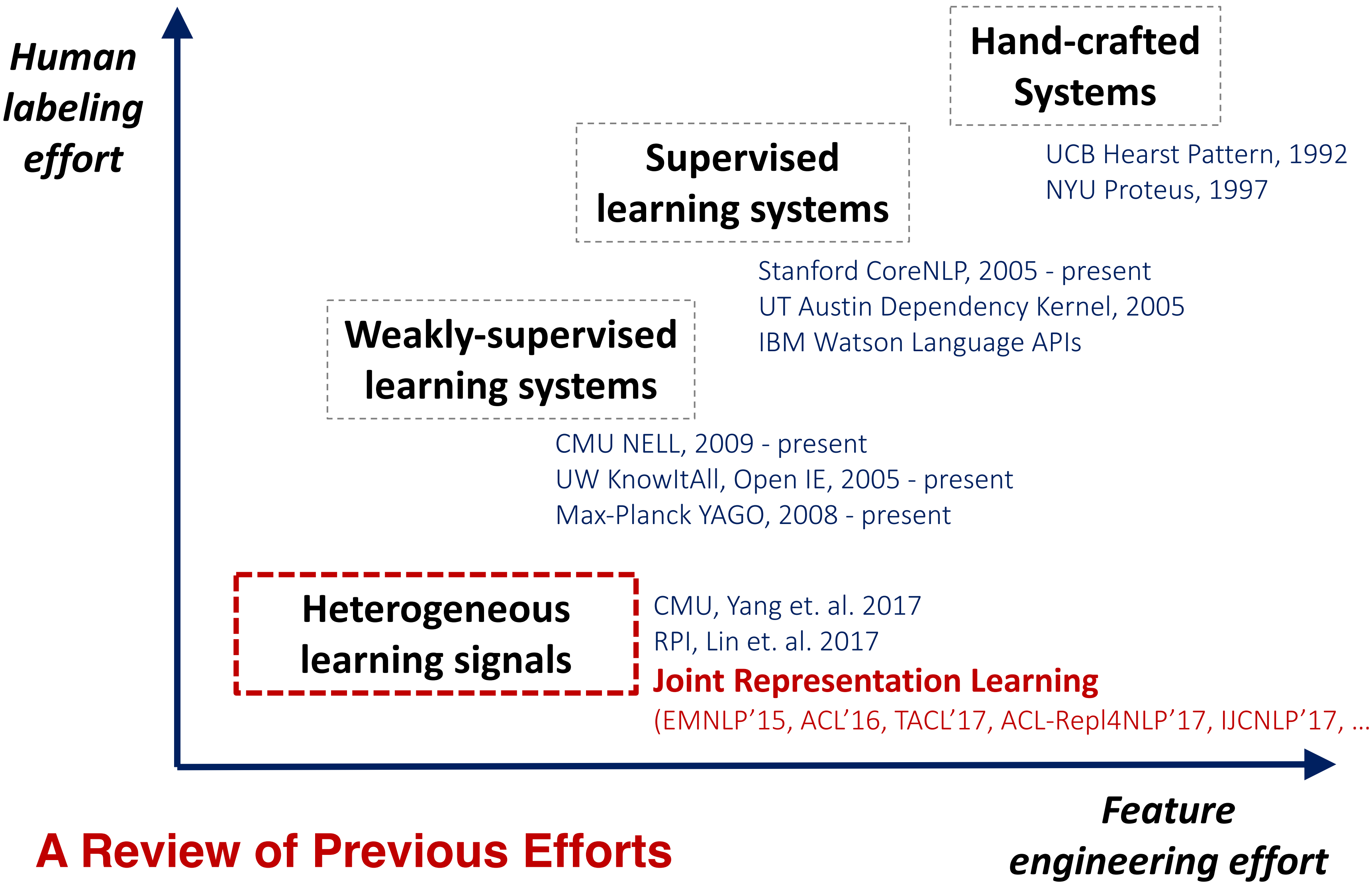
- Constructing data sets is labor intensive
- Many different
  - Languages
  - Domains
  - Modalities
  - …

Joint representation learning

- Learning comprehensive representations from *heterogeneous sources.*
  - *unlabeled data*
  - annotations for *related tasks, domains, languages*.
- Encoding structured knowledge to learn robust representations and make *holistic decisions*.
  - *linguistic structures*

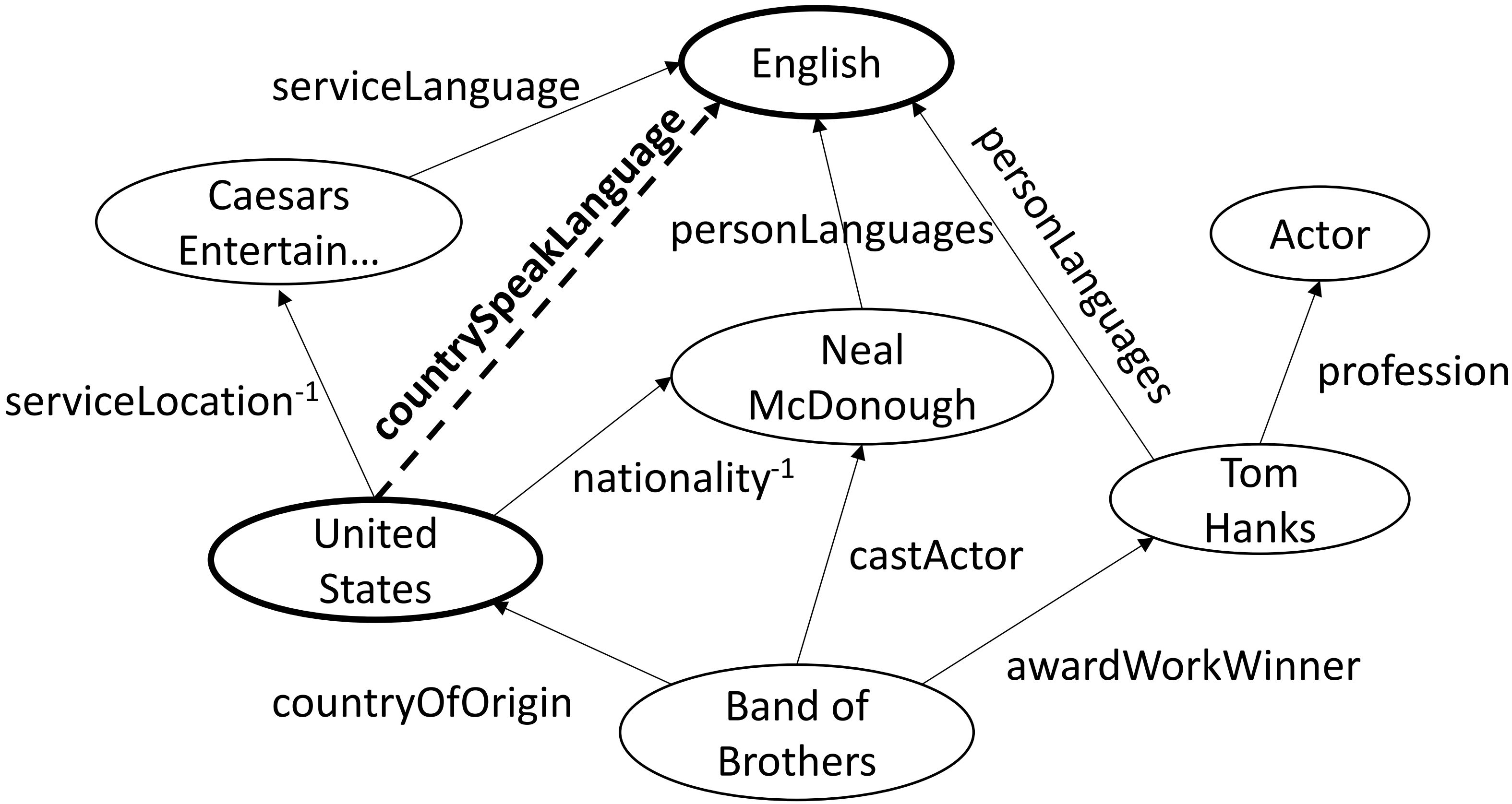


Low-resource IE: Another Way to Reduce Human Effort



A Review of Previous Efforts

Knowledge Bases are Highly Incomplete



Query Start Node: “United States” *Query End Node*: “English”
Query: ?(United States, English)

Knowledge Base Reasoning

- **Question: can we infer missing links based on background KB?**

- **Path-based methods**
  - Path-Ranking Algorithm (PRA), Lao et al. 2011
  - RNN + PRA, Neelakantan et al, 2015
  - Chains of Reasoning, Das et al, 2017

- **Embedding-based methods**
  - RESCAL, Nickel et al., 2011
  - TransE, Bordes et al, 2013
  - TransR/CTransR, Lin et al, 2015

- **Integrating Path and Embedding-Based Methods**
  - DeepPath, Xiong et al, 2017
  - MINERVA, Das et al, 2018
  - DIVA, Chen et al., 2018

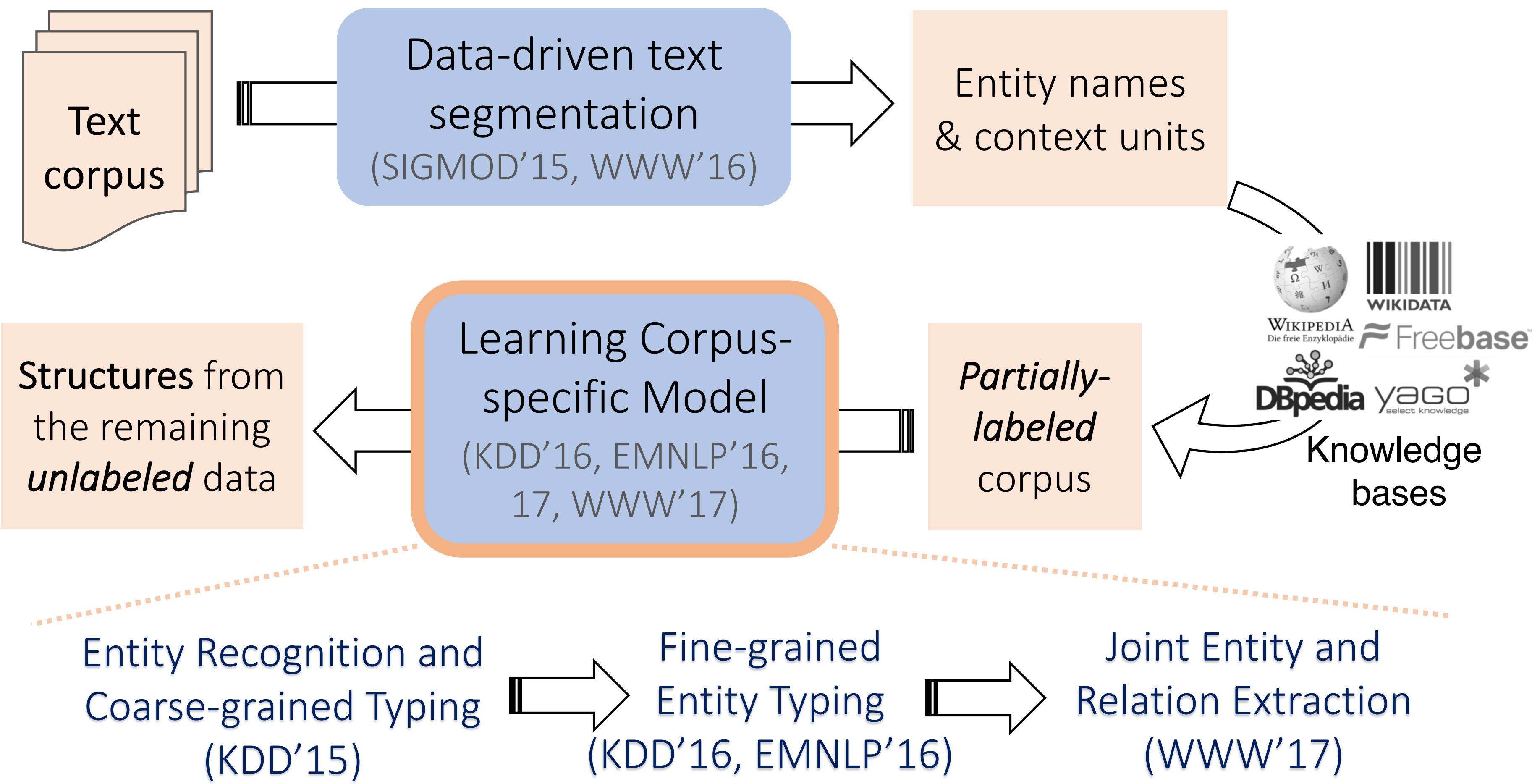
Tutorial Outline

- Introduction
- Part I: Effort–Light Structure Extraction
  - Tea break at 10:00am
- Part II: Low-resource IE
- Part III: Knowledge Base Reasoning
- Summary & Future Directions

Scalable Construction and Reasoning of Massive Knowledge Bases

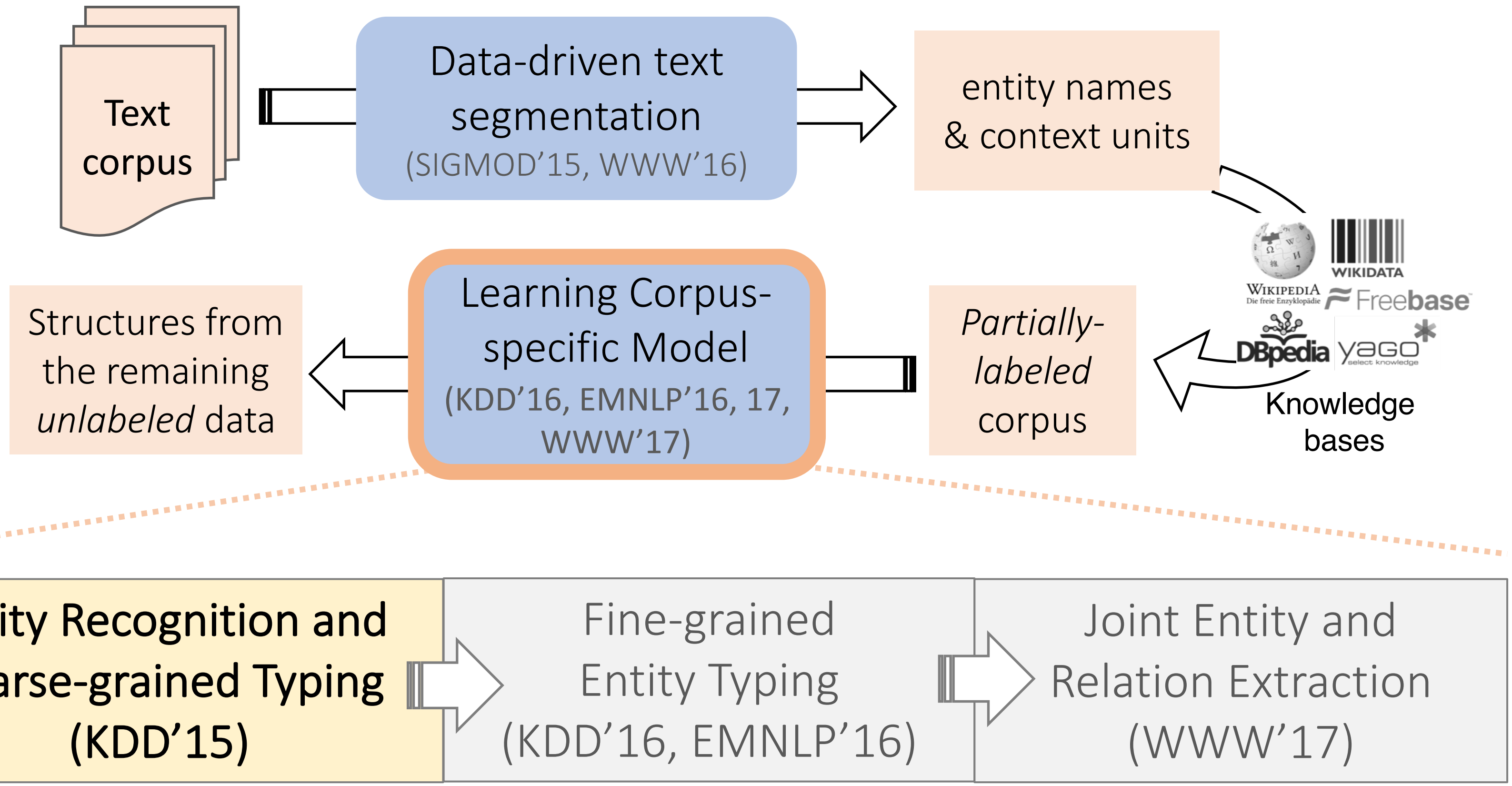
Part I: Effort-Light Structure Extraction

Framework Overview



Corpus to Structured Network: The Roadmap

Corpus to Structured Network: **The Roadmap**



Recognizing Entities of Target Types in Text

The best BBQ I've tasted in Phoenix! I had the pulled pork sandwich with coleslaw and baked beans for lunch. The owner is very nice. ...

The best *BBQ* I've tasted in *Phoenix* ! I had the *pulled pork sandwich* with *coleslaw* and *baked beans* for lunch. The owner is very nice. ...

food

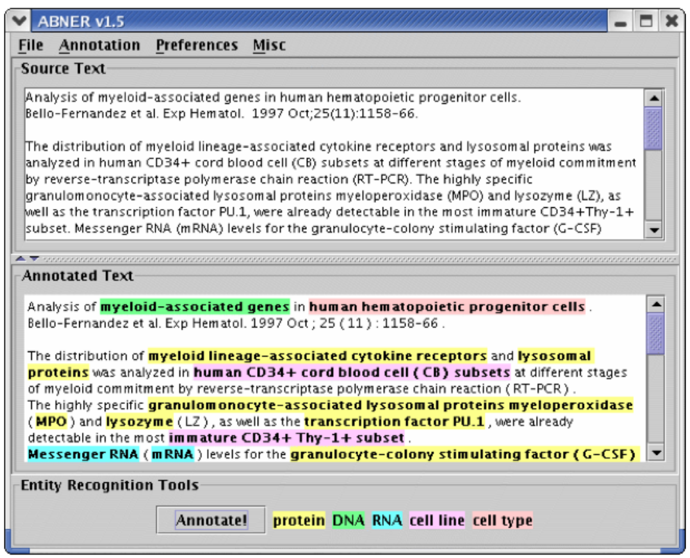
location

person

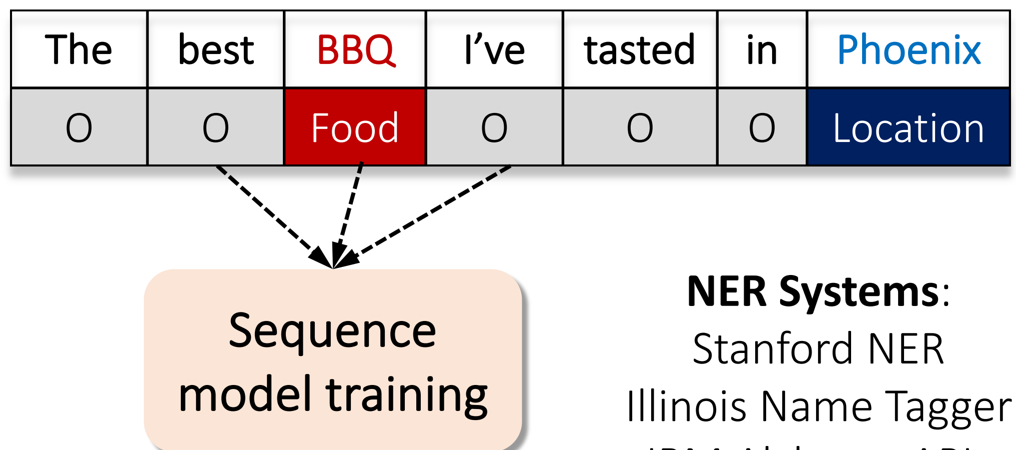
tripadvisor

Traditional Named Entity Recognition (NER) Systems

- Heavy reliance on corpus-specific human labeling
- Training sequence models is slow



A manual annotation interface

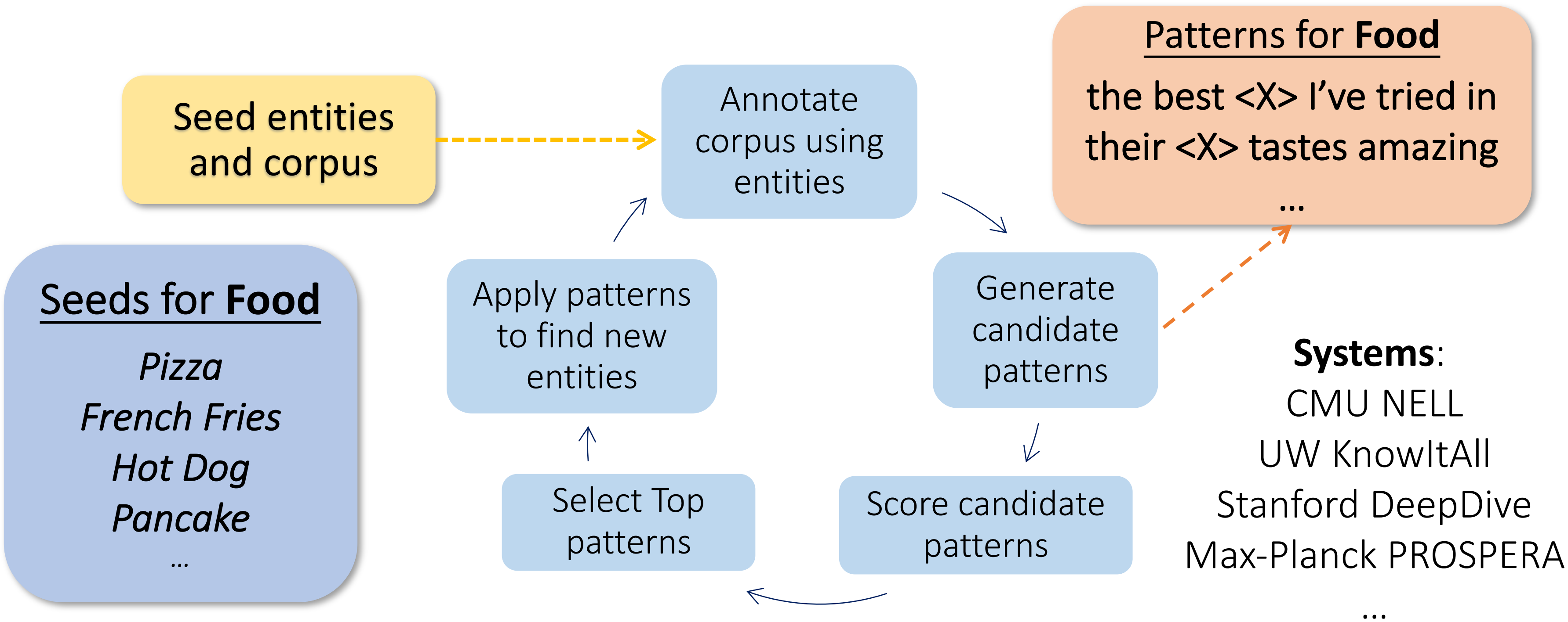


NER Systems:
Stanford NER
Illinois Name Tagger
IBM Alchemy APIs
…

e.g., (McMallum & Li, 2003), (Finkel et al.,2005), (Ratinov & Roth, 2009), …

Weak-Supervision Systems: Pattern-Based Bootstrapping

- Requires manual seed selection & mid-point checking

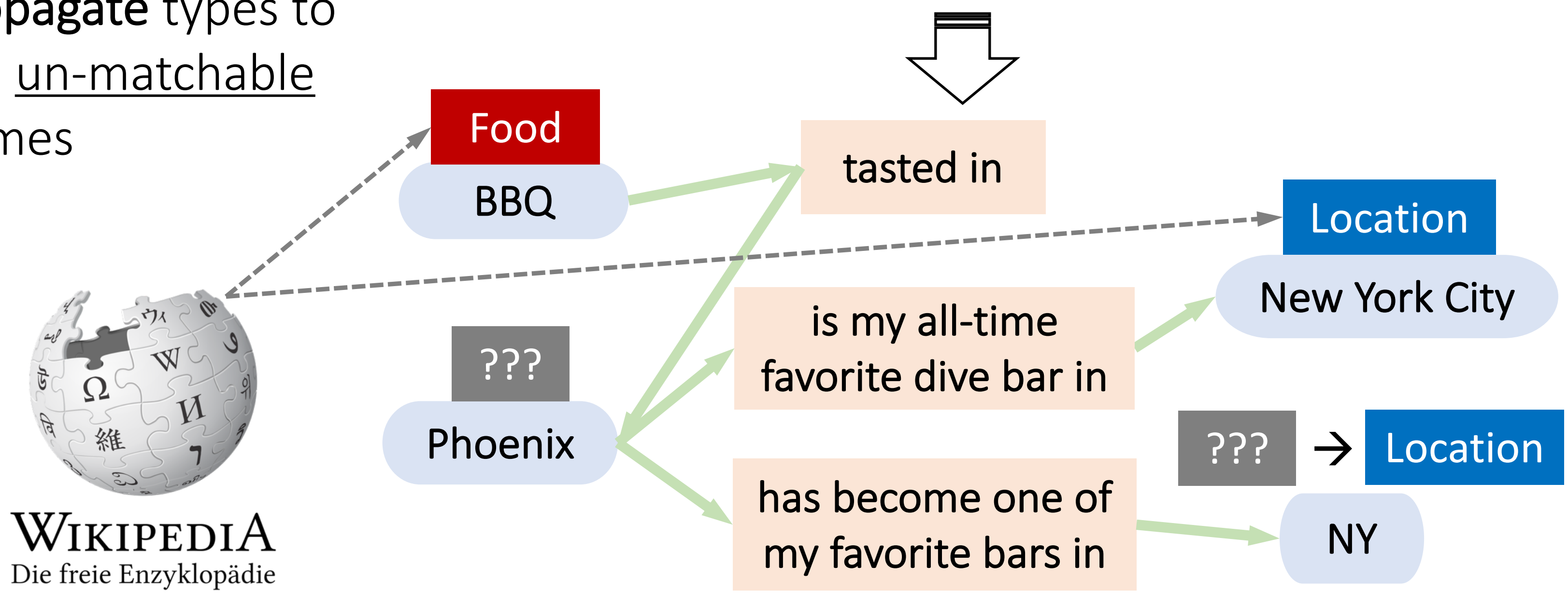


e.g., (Etzioni et al., 2005), (Talukdar et al., 2010), (Gupta et al., 2014), (Mitchell et al., 2015), ...

Leveraging Distant Supervision

1. **Detect** entity names from text
2. **Match** name strings to KB entities
3. **Propagate** types to the un-matchable names

ID	Sentence
S1	*Phoenix* is my all-time favorite dive bar in *New York City* .
S2	The best *BBQ* I've tasted in *Phoenix.*
S3	*Phoenix* has become one of my favorite bars in *NY* .



(Lin et al., 2012), (Ling et al., 2012), (Nakashole et al., 2013)

Current Distant Supervision: Limitation

1. Context-agnostic type prediction
    - Predict types for each mention regardless of context
2. Sparsity of contextual bridges

ID	Sentence
S1	*Phoenix* is my all-time favorite dive bar in *New York City* .
S2	The best *BBQ* I've tasted in *Phoenix*.
S3	*Phoenix* has become one of my favorite bars in *NY* .

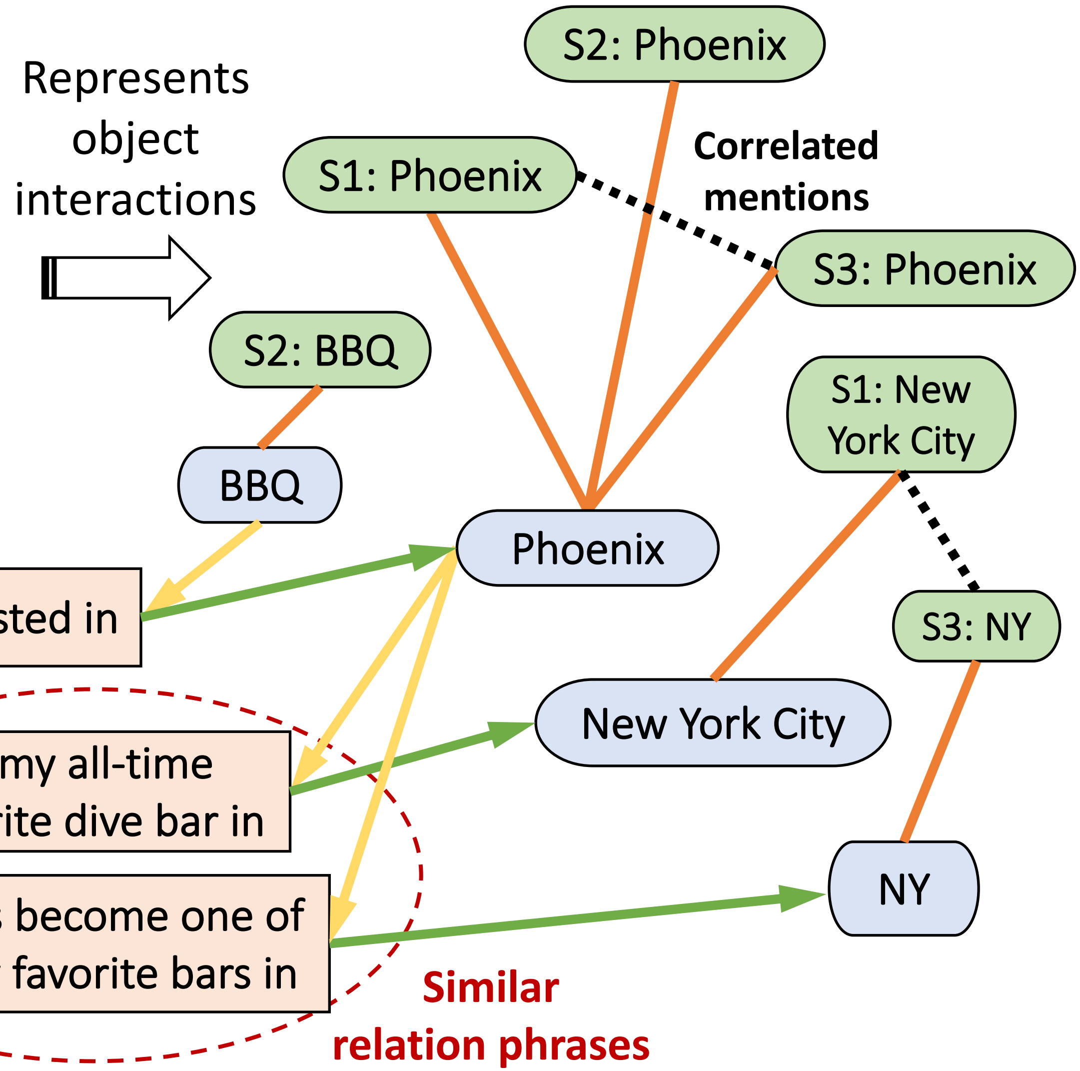
Current Distant Supervision: Limitation

1. Context-agnostic type prediction
2. Sparsity of contextual bridges
   - Some **relational phrases** are **infrequent** in the corpus
     → ineffective type propagation

ID	Sentence
S1	*Phoenix* is my all-time favorite dive bar in *New York City* .
S3	*Phoenix* has become one of my favorite bars in *NY* .

The **ClusType** Approach (KDD'15)

ID	Segmented Sentences
S1	*Phoenix* is my all-time favorite dive bar in *New York City* .
S2	The best *BBQ* I've tasted in *Phoenix*.
S3	*Phoenix* has become one of my favorite bars in *NY* .



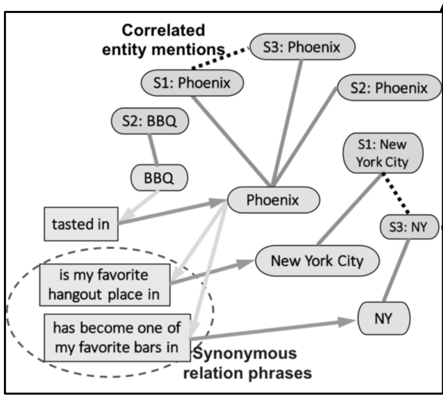
Putting two sub-tasks together:

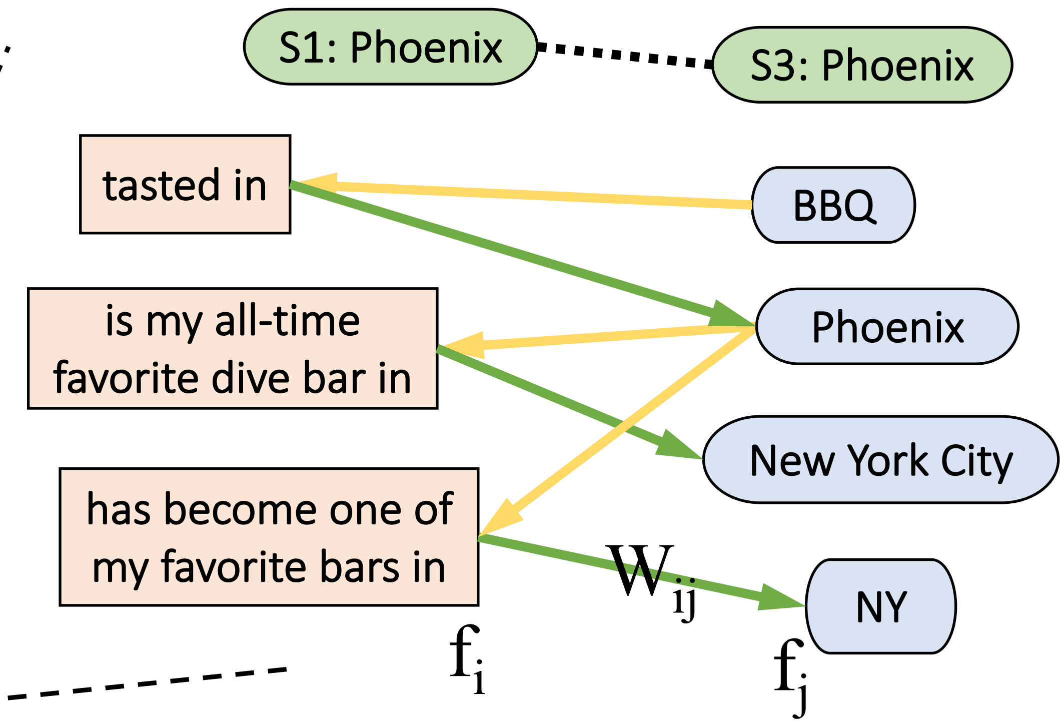
1. Type label propagation
2. Relation phrase clustering

Type Propagation in **ClusType**

Smoothness Assumption

If two objects are similar according to the graph, then their type labels should be also similar





Edge weight / object similarity

Vector of scores for single label on nodes

Measure of Non-Smoothness

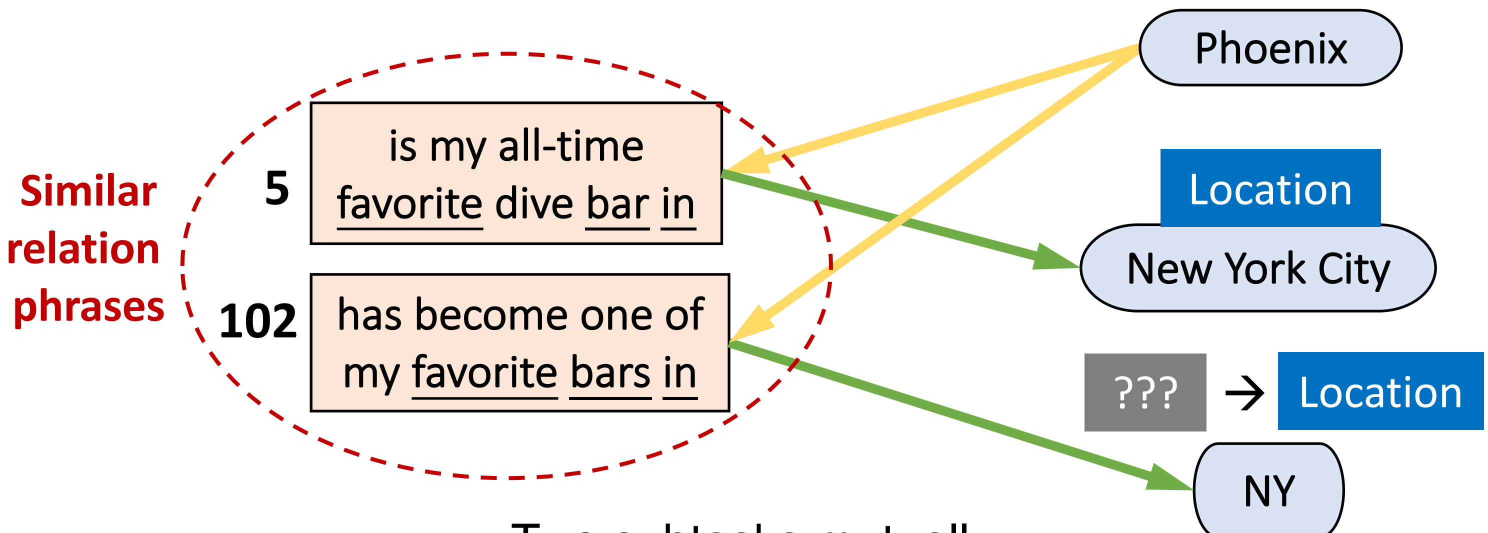
$$f^T L f = \sum_{i,j} W_{ij}(f_i - f_j)^2$$

(Belkin & Partha, NIPS'01), (Ren et al., KDD'15)

Relation Phrase Clustering in **ClusType**

- Two relation phrases should be grouped together if:
    1. Similar string
    2. Similar context
    3. Similar types for entity arguments

“Multi-view” clustering



Two subtasks mutually enhance each other

(Ren et al., KDD’15)

ClusType: Comparing with State-of-the-Art Systems (F1 Score)

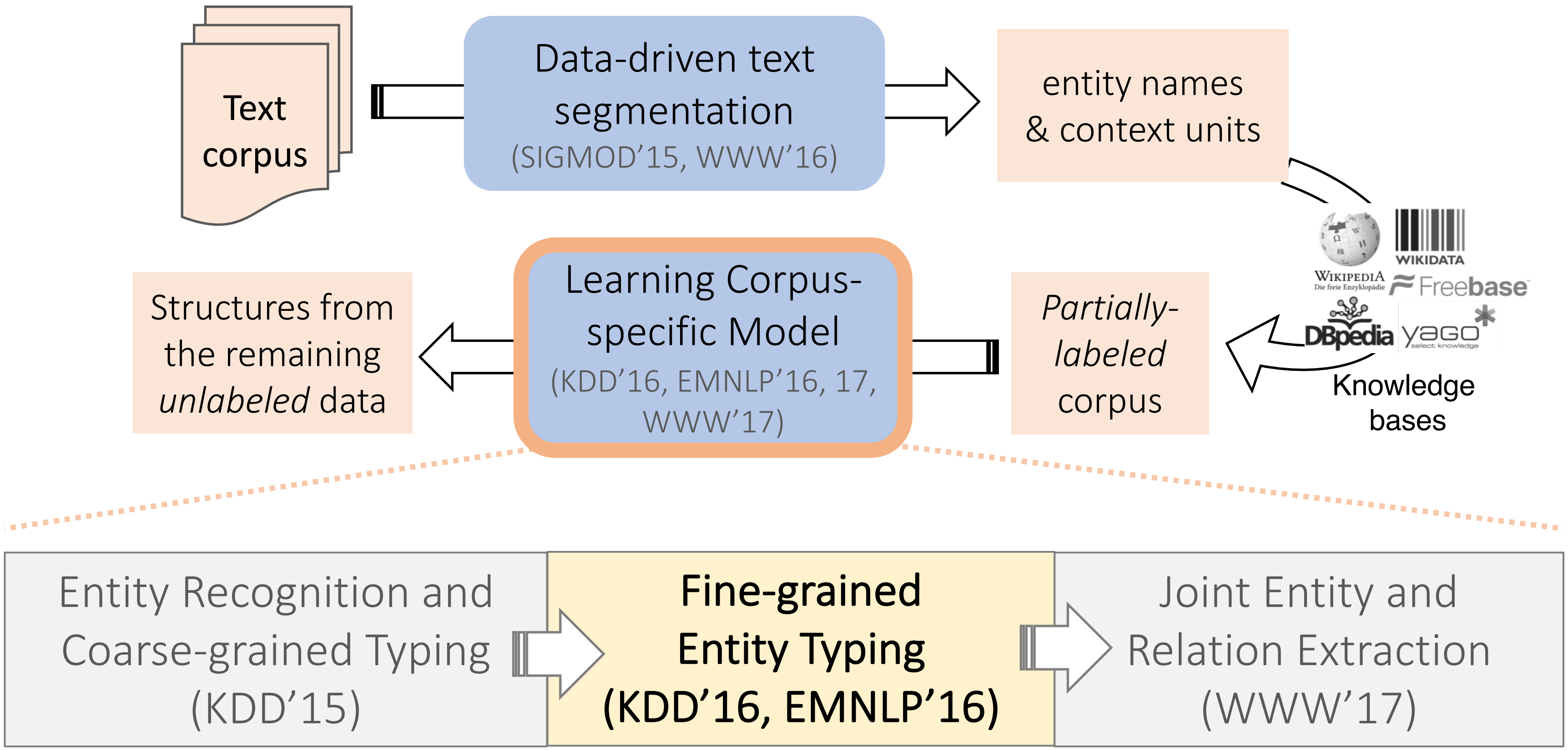
	Methods	NYT	Yelp	Tweet
Bootstrapping	**Pattern** (Stanford, CONLL'14)	0.301	0.199	0.223
	SemTagger (U Utah, ACL'10)	0.407	0.296	0.236
Label propagation	**NNPLB** (UW, EMNLP'12)	0.637	0.511	0.246
	APOLLO (THU, CIKM'12)	0.795	0.283	0.188
Classifier with linguistic features	**FIGER** (UW, AAAI'12)	0.881	0.198	0.308
	ClusType (KDD'15)	**0.939**	**0.808**	**0.451**

- vs. **bootstrapping**: context-aware prediction on "un-matchable"
- **vs. label propagation**: group similar relation phrases
- **vs. FIGER**: no reliance on complex feature engineering

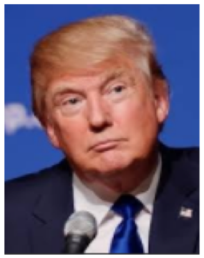
NYT: 118k news articles (1k manually labeled for evaluation); **Yelp**: 230k business reviews (2.5k reviews are manually labeled for evaluation); **Tweet**: 302 tweets (3k tweets are manually labeled for evaluation)

Precision (P) = $\frac{\#Correctly\text{−}typed\ mentions}{\#System\text{−}recognized\ mentions}$, Recall ($R$) = $\frac{\#Correctly\text{−}typed\ mentions}{\#ground\text{−}truth\ mentions}$, F1 score = $\frac{2(P \times R)}{(P+R)}$

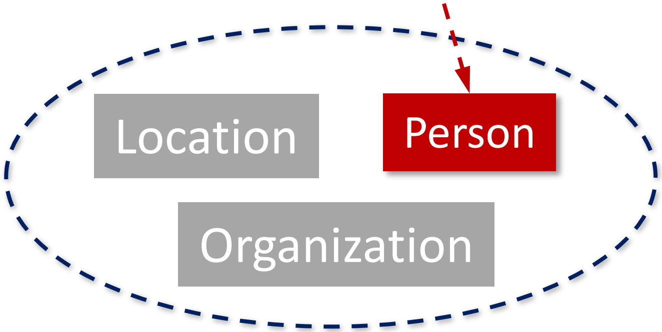
Corpus to Structured Network: **The Roadmap**



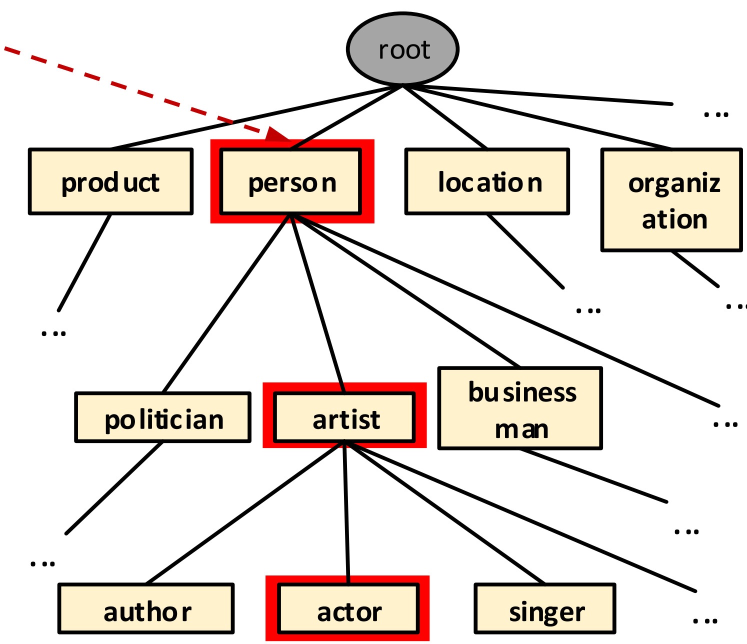
From Coarse-Grained Typing to **Fine-Grained Entity Typing**



ID	Sentence
S1	*Donald Trump* spent 14 television seasons presiding over a game show, NBC's The Apprentice.



A few common types



A type hierarchy with 100+ types
(from knowledge base)

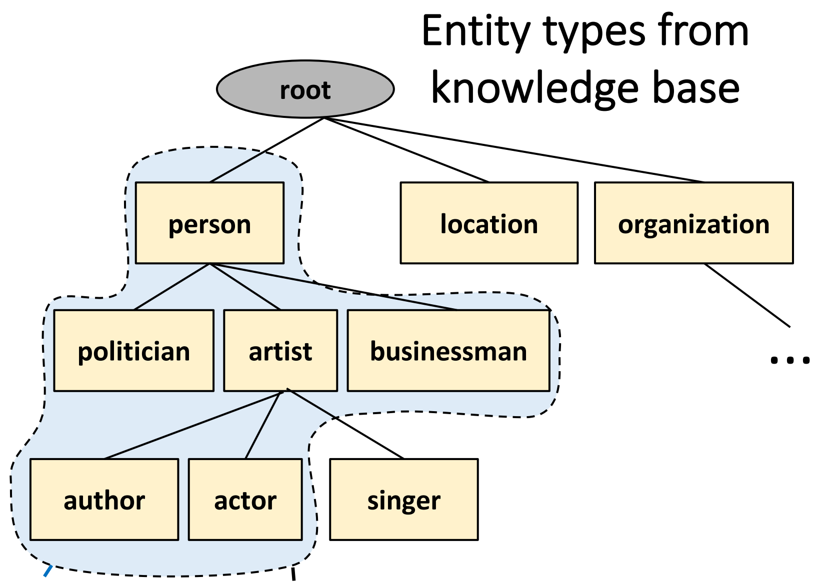
(Ling et al., 2012), (Nakashole et al., 2013), (Yogatama et al., 2015)

Current Distant Supervision: **Context-Agnostic Labeling**

- Inaccurate labels in **training data**
- **Prior work**: <u>all</u> labels are "<u>perfect</u>"

ID	Sentence
S1	*Donald Trump* spent 14 television seasons presiding over a game show, NBC's The Apprentice

S1: *Donald Trump*
Entity Types: person, artist, actor, ~~author~~, ~~businessman~~, ~~politician~~



Entity:
Donald Trump

Modeling Clean and Noisy Mentions Separately

For a clean mention, its "*positive types*" should be **ranked higher** than all its "*negative types*"

S_i: *Ted Cruz*
Types in KB: person, politician

ID	Noisy Entity Mention
S1	***Donald Trump*** spent 14 television seasons presiding over a game show, NBC's The Apprentice

S1: *Donald Trump*
Types in KB: person, artist, actor, author, businessman, politician

Types ranked

(+) actor
(-) singer
(-) coach
(-) doctor
(-) location
(-) organization

"**Best**" candidate type

(+) actor 0.88
(+) artist 0.74
(+) person 0.55
(+) author 0.41
(+) politician 0.33
(+) business 0.31

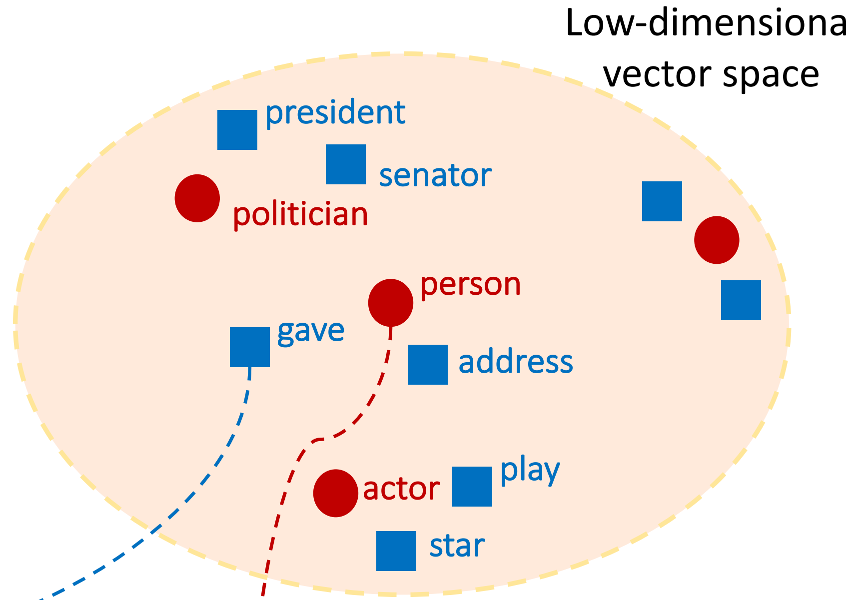
For a noisy mention, its "*best candidate type*" should be **ranked higher** than all its "*non-candidate types*"

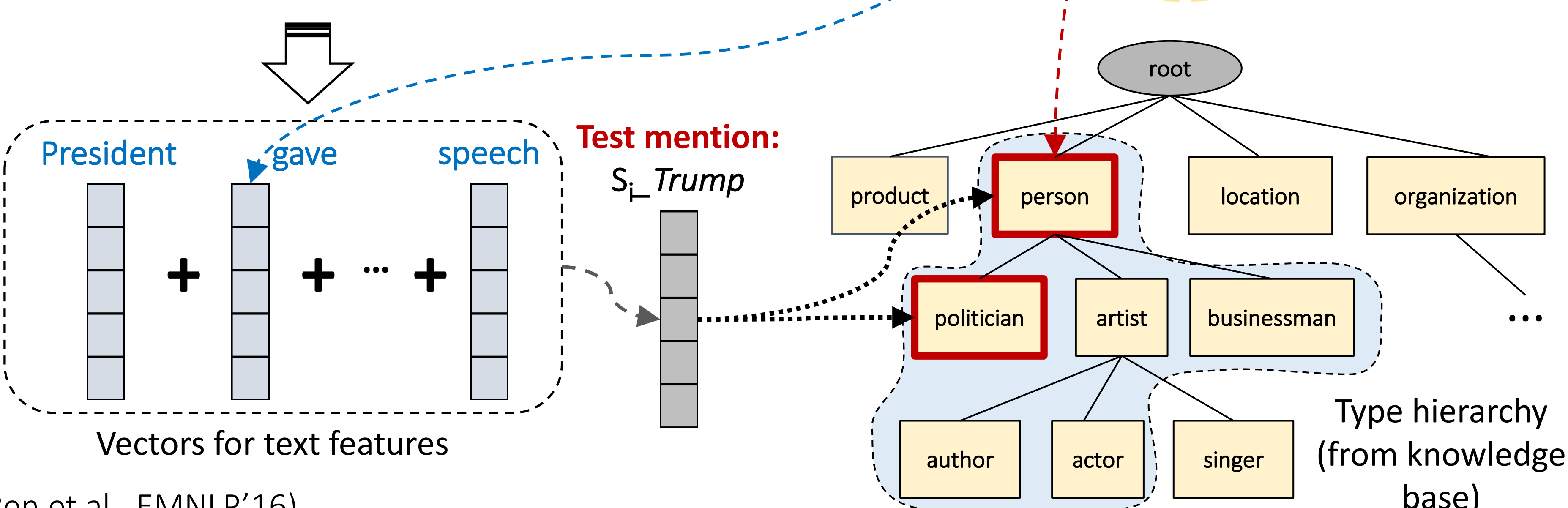
(Ren et al., EMNLP'16, KDD'16)

Hierarchical Type Inference

- Top-down nearest neighbor search in the given type hierarchy

ID	Sentence
S_i	President ***Trump*** gave an all-hands address to troops at the U.S. Central Command headquarters





(Ren et al., EMNLP'16)

Partial **L**abel **E**mbedding (KDD'16)

Extract Text Features

↓

"Label Noise Reduction" with **PLE**

↓

Train Classifiers on De-noised Data

↓

Prediction on New Data

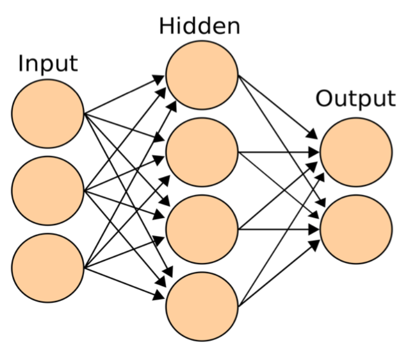
ID	Sentence
S1	*Donald Trump* spent 14 television seasons presiding over a game show, NBC's The Apprentice

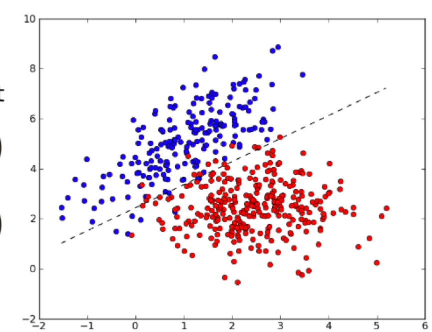
Text features: TOKEN_Donald, CONTEXT: television, CONTEXT: season, TOKEN_trump, SHAPE: AA

S1: *Donald Trump*

Entity Types: person, artist, actor, ~~author~~, ~~businessman~~, ~~politician~~

"De-noised" labeled data



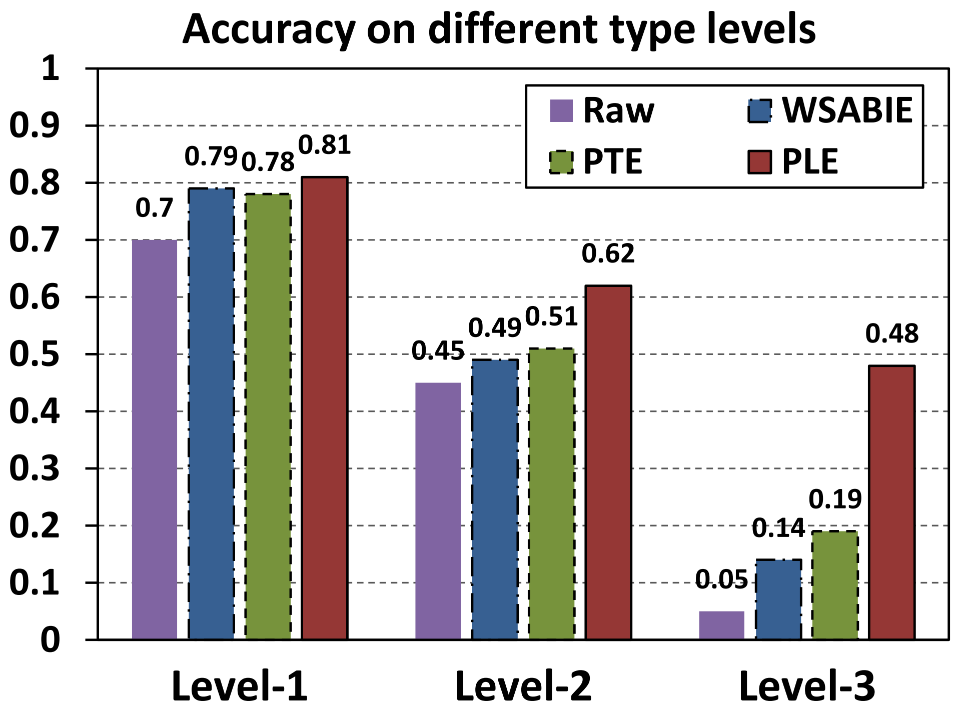


More effective classifiers

(Ren et al., KDD'16)

Performance of Fine-Grained Entity Typing

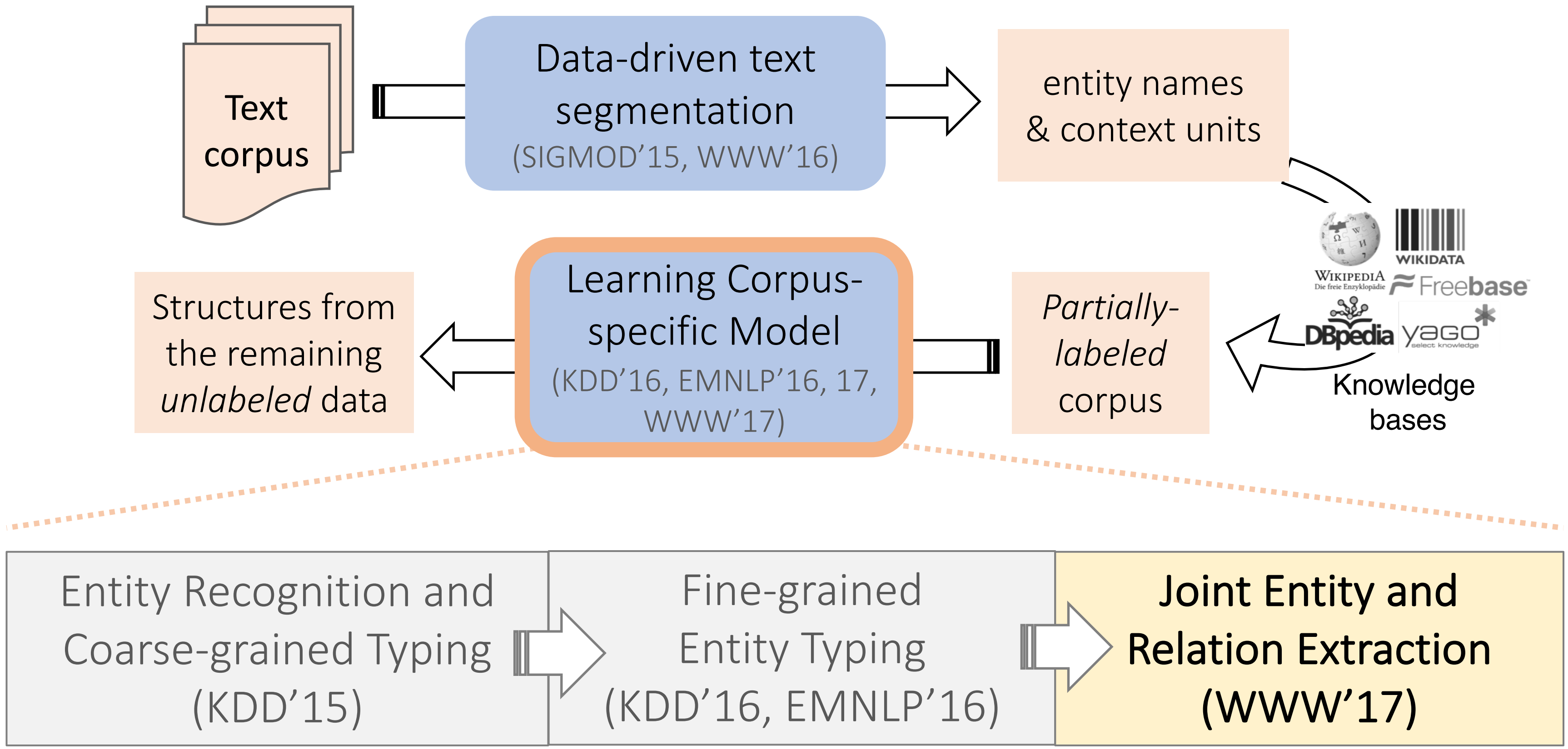
$$\text{Accuracy} = \frac{\#\ \textit{mentions with all types correctly predicted}}{\#\ \textit{mentions in the test set}}$$



- Raw: candidate types from distant supervision
- WSABIE (Google, ACL'15): joint feature and type embedding
- Predictive Text Embedding (MSR, WWW'15): joint mention, feature and type embedding
  - Both WASBIE and PTE suffer from "noisy" training labels
- PLE (KDD'16): partial-label loss for context-aware labeling

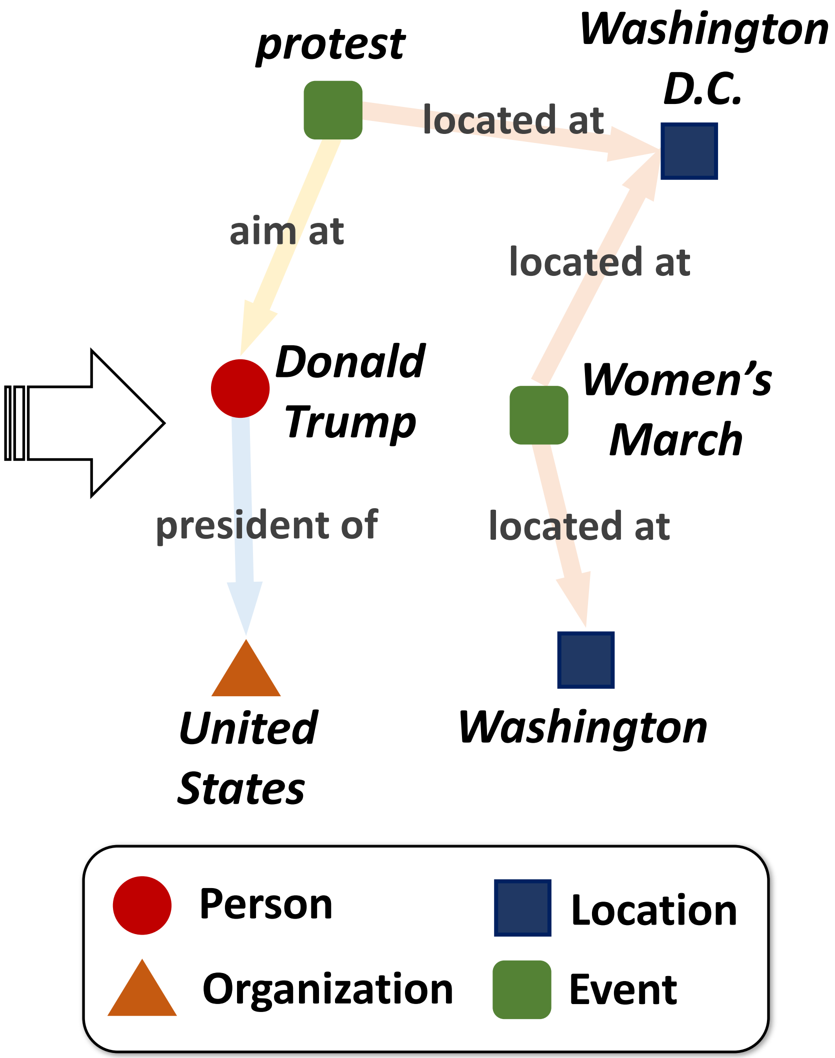
OntoNotes public dataset (Weischedel et al. 2011, Gillick et al., 2014): 13,109 news articles, 77 annotated documents, 89 entity types

Corpus to Structured Network: **The Roadmap**



Joint Extraction of Typed Entities and Relations

The **Women's March** was a worldwide **protest** on January 21, 2017. The **protest** was aimed at **Donald Trump**, the recently inaugurated president of the **United States**. The first **protest** was planned in **Washington, D.C.**, and was known as the **Women's March** on **Washington**.



Prior Work: Relation Extraction (RE)

Substantial task-specific human annotation → ***No task-specific human annotation***

Supervised RE systems ⇨ Pattern-based bootstrapping RE systems ⇨ Distantly-supervised RE systems (cont.)

Supervised RE systems
- Hard to be ported to deal with different kinds of corpora

Pattern-based bootstrapping RE systems
- Focus on “explicit” relation mentions
- “Semantic drift”

Distantly-supervised RE systems (cont.)
- Error propagation
- Noisy candidate type labels

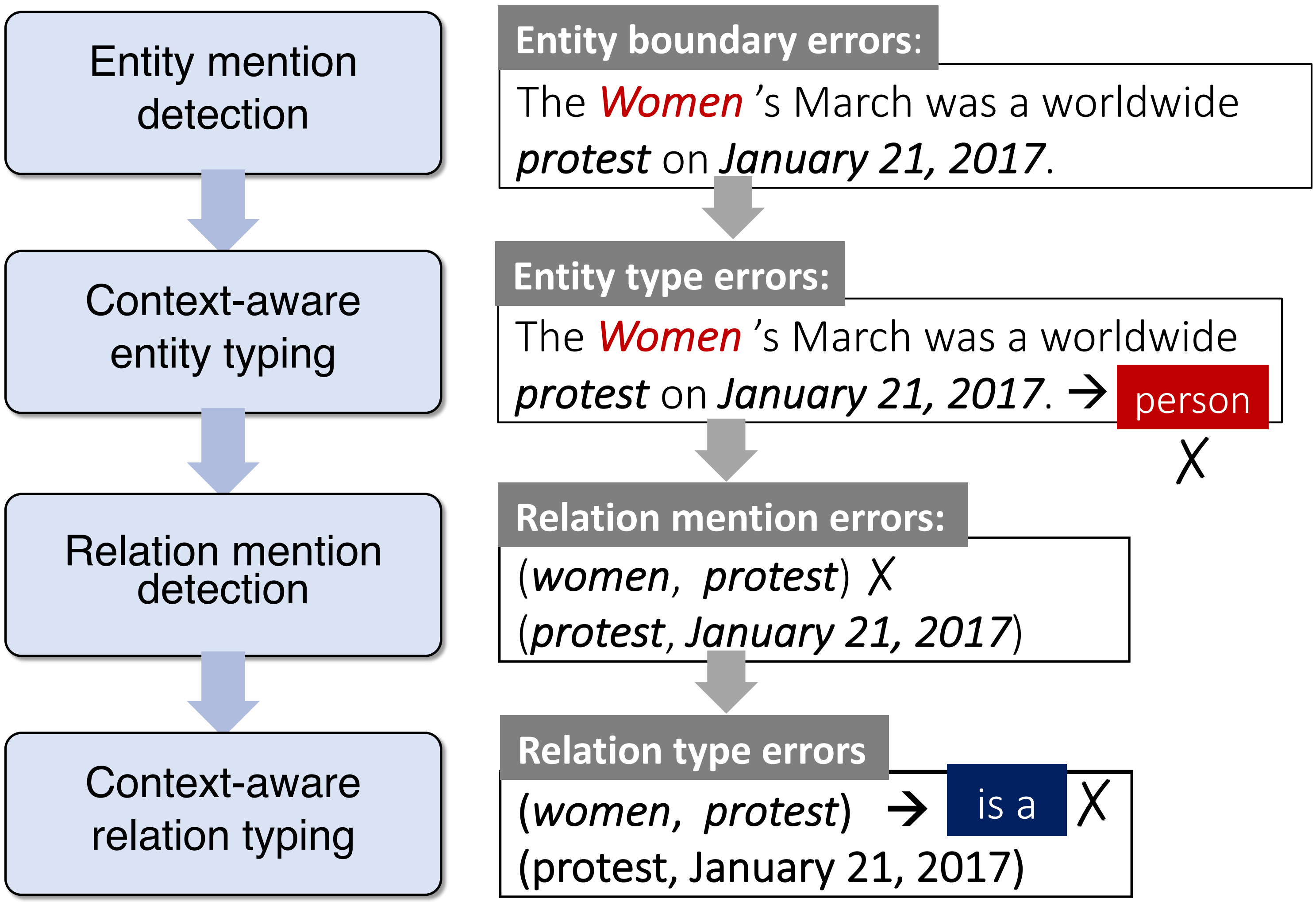
Mintz et al. *Distant supervision for relation extraction without labeled data*. ACL, 2009.

Etzioni et al. *Web-scale information extraction in knowitall.* WWW, 2004.

Surdeanu et al. *Multi-instance multi-label learning for relation extraction*. EMNLP, 2012.

Prior Work: An "Incremental" System Pipeline

Error propagation cascading down the pipeline



(Mintz et al., 2009), (Riedel et al., 2010), (Hoffmann et al., 2011), (Surdeanu et al., 2012), (Nagesh et al., 2014), …

The **CoType** Approach (WWW'17)

Entity mention detection

Context-aware entity typing

Relation mention detection

Context-aware relation typing

1. Data-driven detection of entity and relation mentions

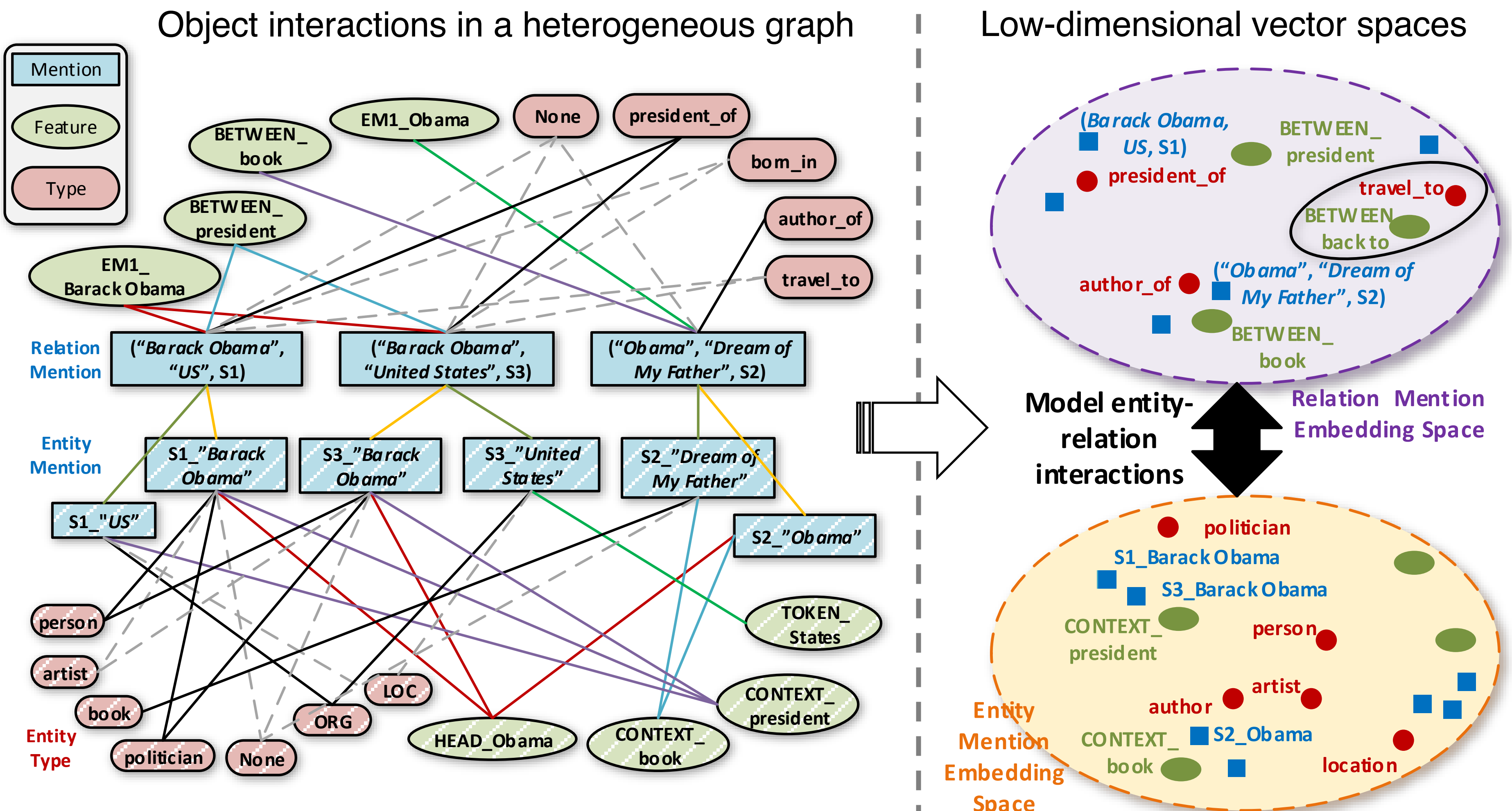
- Data-driven text segmentation
- Syntactic pattern learning from KBs

2. Joint typing of entity and relation mentions

- **Context-aware** type modeling
- Model **entity-relation interactions**

(Ren et al. WWW'17)

CoType: Co-Embedding for Typing Entities and Relations



(Ren et al. WWW'17)

Modeling Entity-Relation Interactions

Object “Translating” Assumption

For a relation mention **z** between entity arguments **m1** and **m2**:

vec(m1) ≈ **vec**(m2) + **vec**(z)

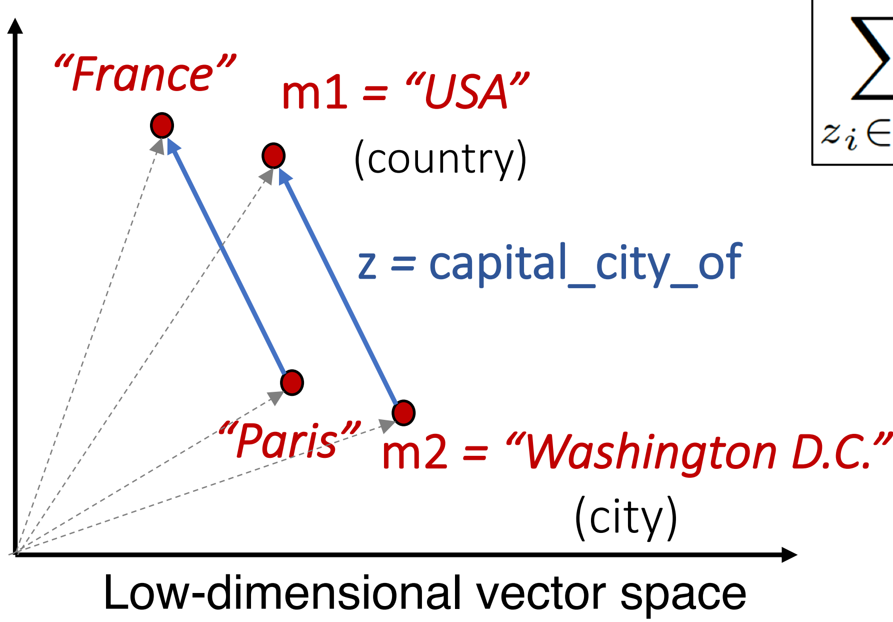
Error on a relation triple (z, m1, m2):

$$\tau(z) = \|\mathbf{m}_1 + \mathbf{z} - \mathbf{m}_2\|_2^2$$

$$\sum_{z_i \in \mathcal{Z}_L} \sum_{v=1}^{V} \max\left\{0, 1 + \tau(z_i) - \tau(z_v)\right\}$$

positive relation triple

negative relation triple



(Bordes, NIPS’13), (Ren et al., WWW’17)

Reducing Error Propagation: A Joint Optimization Framework

Modeling **entity-relation interactions**

$$O_{ZM} = \sum_{z_i \in \mathcal{Z}_L} \sum_{v=1}^{V} \max\left\{0, 1 + \tau(z_i) - \tau(z_v)\right\}$$

$$\min\ \mathcal{O} = \boxed{\mathcal{O}_M} + \boxed{\mathcal{O}_Z} + \boxed{\mathcal{O}_{ZM}}$$

$$O_Z = \mathcal{L}_{ZF} + \sum_{i=1}^{N_L} \ell_i + \frac{\lambda}{2} \sum_{i=1}^{N_L} \|\mathbf{z}_i\|_2^2 + \frac{\lambda}{2} \sum_{k=1}^{K_r} \|\mathbf{r}_k\|_2^2$$

Modeling types of **relation mentions**

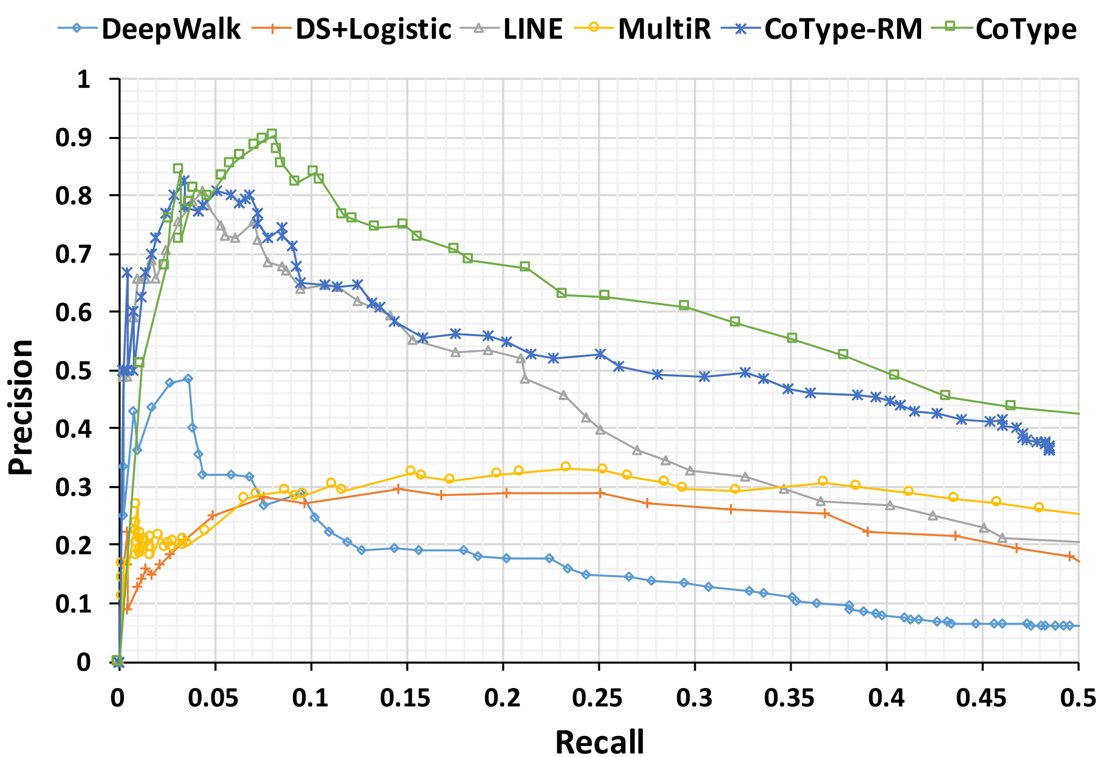
$$O_M = \mathcal{L}_{MF} + \sum_{i=1}^{N_L'} \ell_i' + \frac{\lambda}{2} \sum_{i=1}^{N_L'} \|\mathbf{m}_i\|_2^2 + \frac{\lambda}{2} \sum_{k=1}^{K_y} \|\mathbf{y}_k\|_2^2$$

Modeling types of **entity mentions**

(Ren et al., WWW'17)

CoType: Comparing with State-of-the-Arts RE Systems

- Given candidate relation mentions, predict its relation type if it expresses a relation of interest; otherwise, output “None”



- DS+Logistic (Stanford, ACL’09): logistic classifier on DS
- MultiR (UW, ACL’11): handles inappropriate labels in DS
- DeepWalk (StonyBrook, KDD’14): homogeneous graph embedding
- LINE (MSR, WWW’15): joint feature & type embedding
- CoType-RM (WWW’17): only models relation mentions
- CoType (WWW’17): models entity-relation interactions

NYT public dataset (Riedel et al. 2010, Hoffmann et al., 2011): 1.18M sentences in the corpus, 395 manually annotated sentences for evaluation, 24 relation types

An Application to Life Sciences

LifeNet:

BioInfer Network by human labeling (Pyysalo et al., 2007)
Human-created
1,100 sentences
94 protein-protein interactions
2,500 man-hours
2,662 facts

LifeNet by Effort-Light StructMine
Machine-created
4 Million+ PubMed papers
1,000+ entity types 400+ relation types
<1 hour, single machine
10,000x more facts

Link to PubMed papers

(Pyysalo et al., BMC Bioinformatics'07)
(Ren et al., ACL'17 demo)

Performance evaluation on BioInfer:
Relation Classification Accuracy = 61.7%
(11%↑ over the best-performing baseline)

Towards Automated Structure Extraction

End-to-end extraction models
(Zheng et al., ACL'17), (Xu et al., 2017), (Liu et al., 2017)

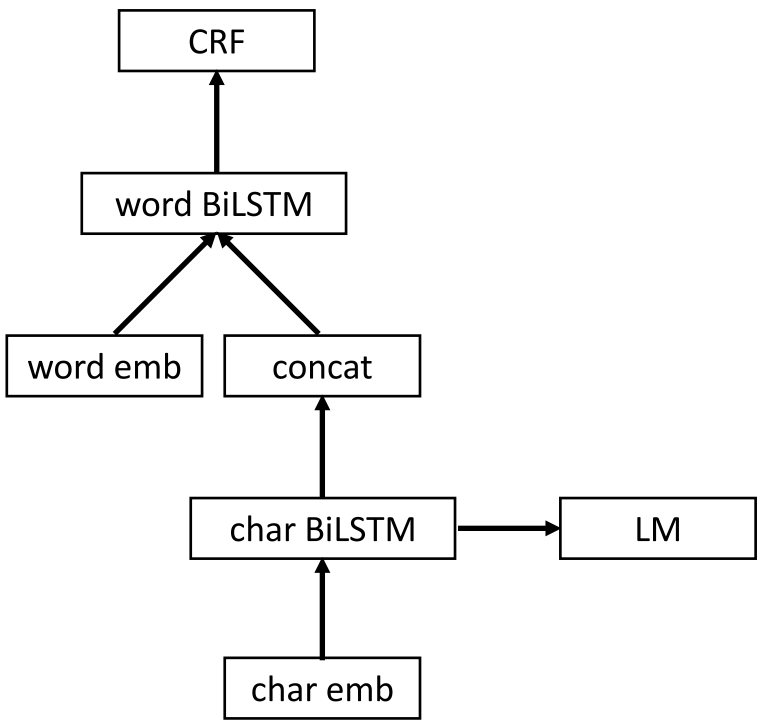
Unifying heterogeneous forms of weak supervisions
(Liu et al., EMNLP'17)

Indirection supervision from auxiliary tasks
(Wu et al., WSDM'18)

Leveraging rich language patterns to facilitate NLU
(Qu et al., 2018), (Liu et al., AAAI'18)

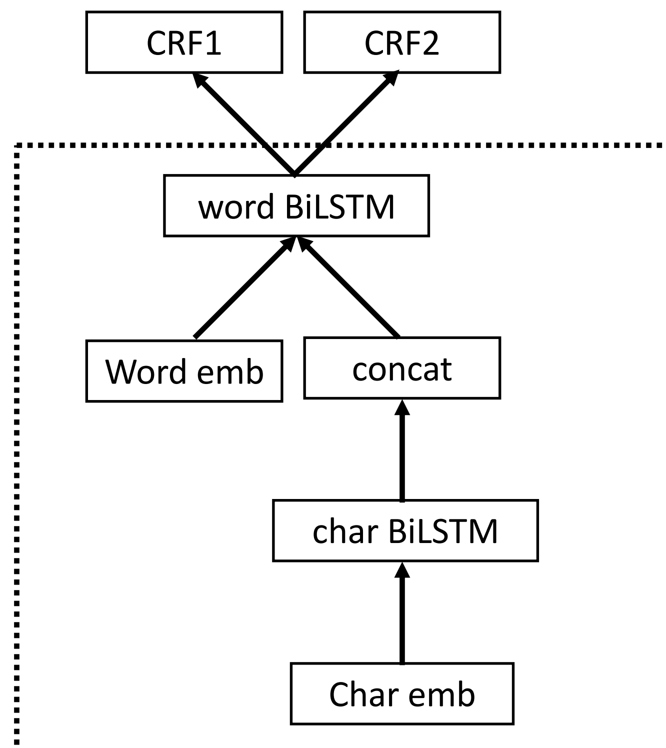
Biomedical Named Entity Recognition by Multi-tasking different datasets

Single-task/dataset learning



(Liu et al., AAAI'18)

Multi-task/dataset learning



(Wang et al., 2018)

State-of-the-art Biomed Entity Tagger

- **One** tagger for **many** biomed entity types (gene, disease, chemical, etc.)
- State-of-the-art performance on several benchmark datasets

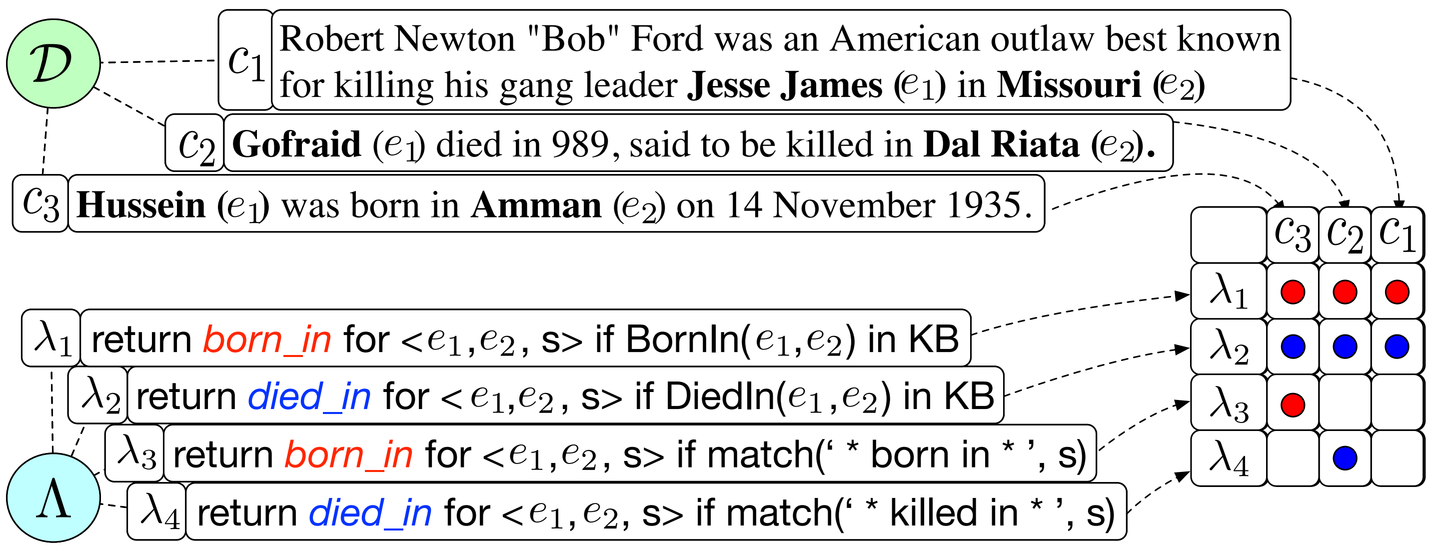
Table 2. Performances of baseline neural network models and the MTM-CW model. Significance test is performed on the F1 values. Bold: best scores, *: significantly worse than the MTM-CW model ($p \leq 0.05$), **: significantly worse than the MTM-CW model ($p \leq 0.01$).

		Dataset Benchmark	Crichton *et al.*	Lample *et al.* Habibi *et al.*	Ma and Hovy	Liu *et al.* STM	MTM-CW
BC2GM (Exact)	Precision	-	-	78.99	83.33	83.07	**83.98**
	Recall	-	-	78.16	81.25	82.02	**82.32**
	F1	-	73.17**	78.57**	82.28**	82.54*	**83.14**
BC2GM (Alternative)	Precision	88.48	-	86.11	83.50	88.21	**89.45**
	Recall	85.97	-	86.96	87.13	87.43	**88.67**
	F1	87.21**	84.41**	86.53**	85.27**	87.82*	**89.06**
BC4CHEMD	Precision	89.09	-	87.83	**90.59**	89.55	90.51
	Recall	85.75	-	85.45	82.63	84.62	**86.18**
	F1	87.39	83.02**	86.62*	86.43*	87.01*	**88.29**
BC5CDR	Precision	**89.21**	-	86.82	88.24	87.41	87.69
	Recall	84.45	-	86.40	78.79	83.05	**87.17**
	F1	86.76	83.90**	86.61*	83.24**	85.18**	**87.43**
NCBI-Disease	Precision	85.10	-	**86.43**	84.33	84.84	85.00
	Recall	80.80	-	82.92	83.77	85.39	**87.80**
	F1	82.90**	80.37**	84.64**	84.04**	85.10**	**86.37**
JNLPBA	Precision	69.42	-	71.35	**72.88**	72.29	72.72
	Recall	75.99	-	75.74	75.98	77.25	**77.83**
	F1	72.55**	70.09**	73.48**	74.40*	74.69*	**75.19**

(Wang et al., 2018)

Heterogeneous Supervision for Relation Extraction

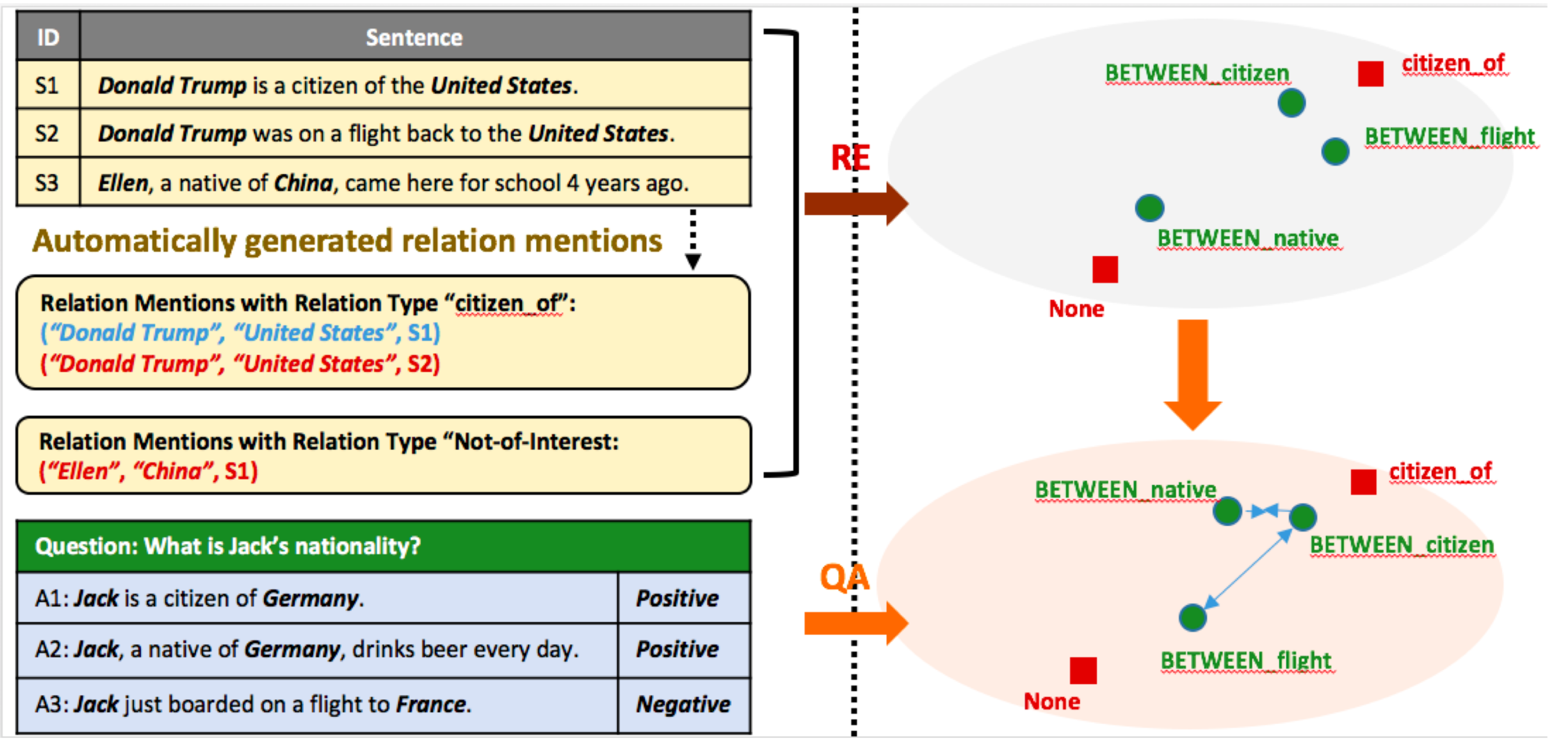
- A principled l framework to **unify** KB-supervision, manual rules, crowd-sourced labels, etc.
- Multiple "**labeling functions**" annotate one instance → resolve conflicts & redundancy → "**expertise**" of each labeling function



(Liu et al., EMNLP'17)

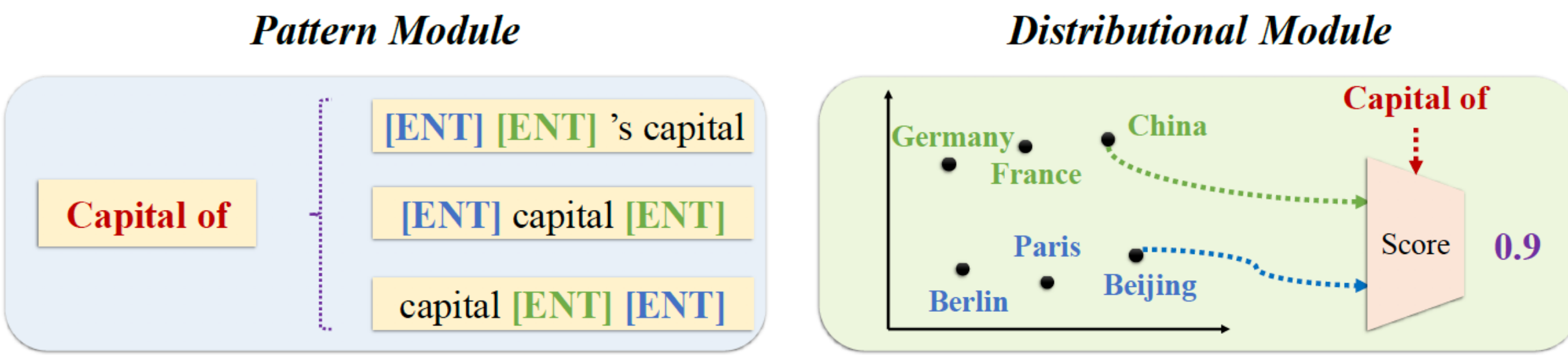
Indirect Supervision for Relation Extraction – using QA Pairs

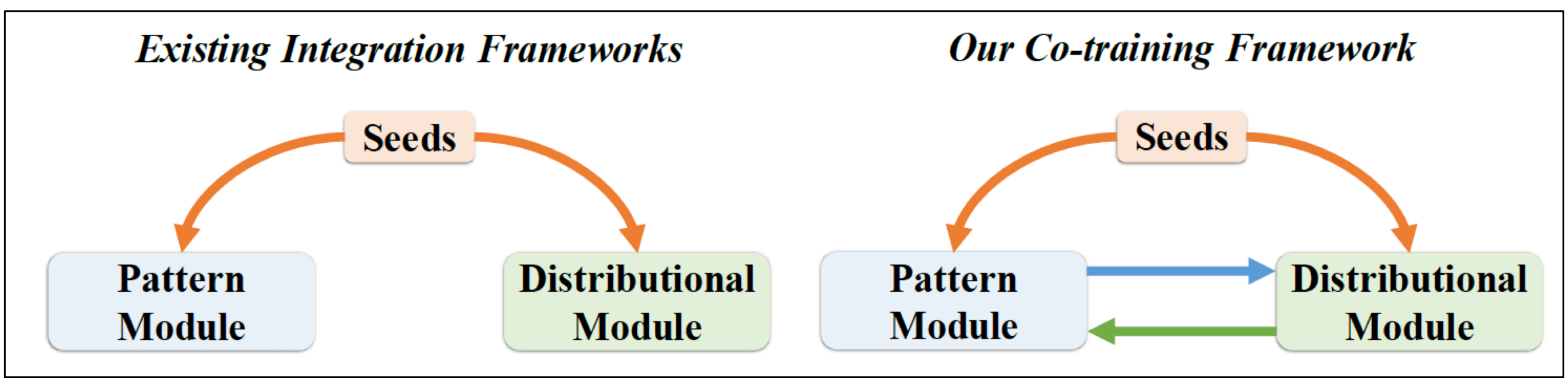
- Questions → positive / negative answers
- pos pairs → similar relation; neg pairs → distinct relations



(Wu et al., WSDM'18)

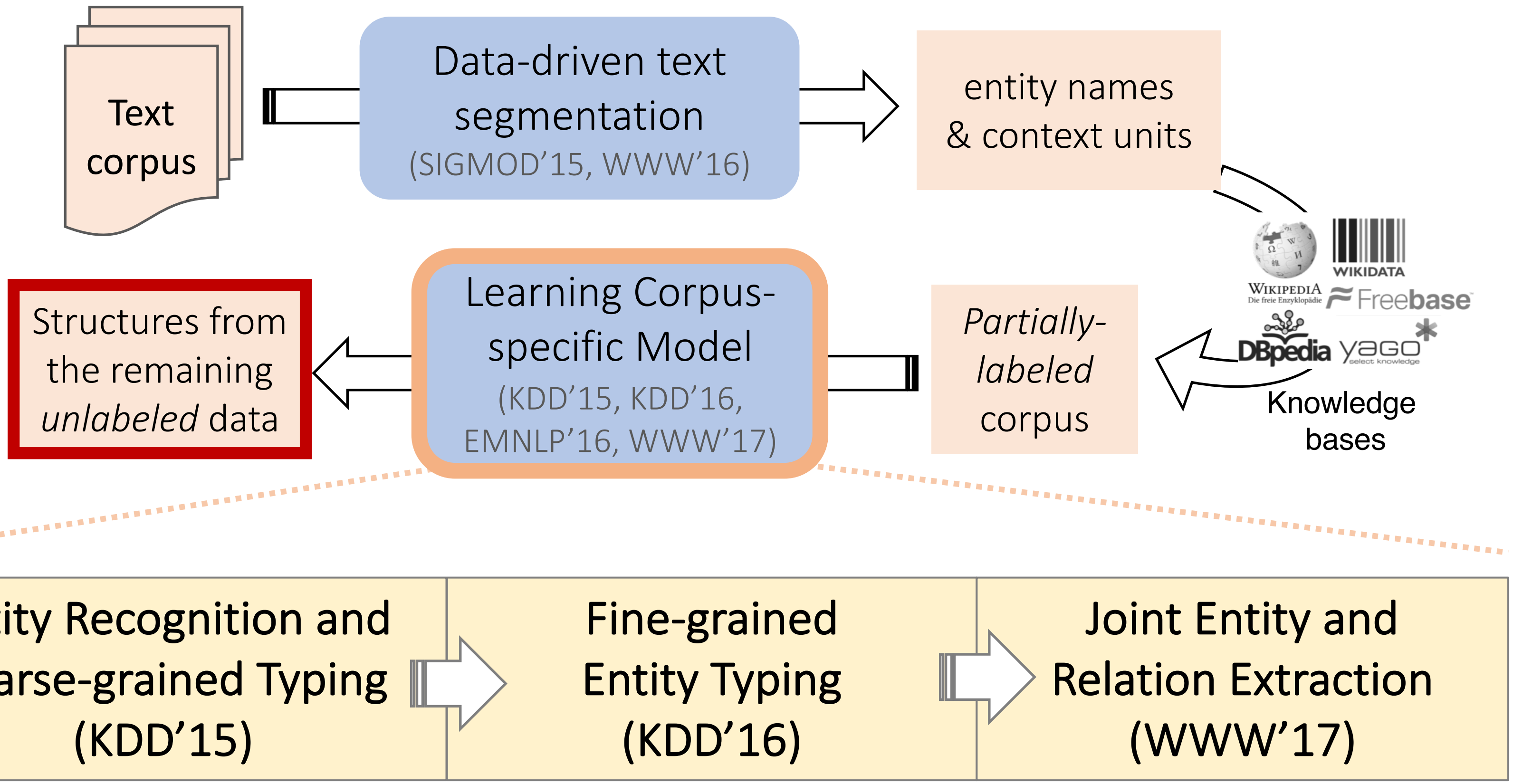
Pattern-enhanced Distributional Representation Learning





(Qu et al., WWW'18)

Corpus to Structured Network: **The Roadmap**



References I

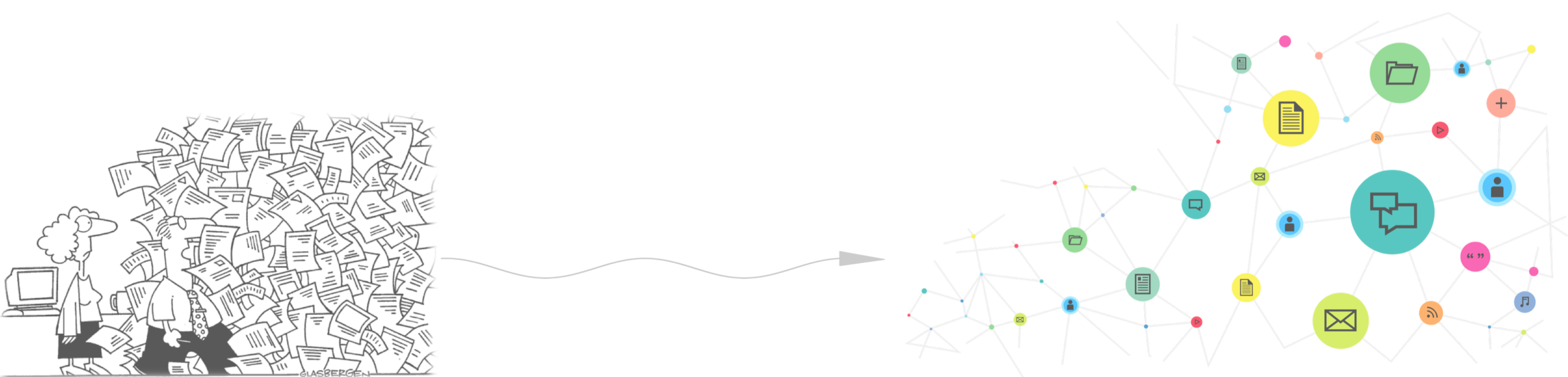
- Xiang Ren, Zeqiu Wu, Wenqi He, Meng Qu, Clare R. Voss, Heng Ji, Tarek F. Abdelzaher, Jiawei Han. CoType: Joint Extraction of Typed Entities and Relations with Knowledge Bases. WWW, 2017.
- Xiang Ren, Ahmed El-Kishky, Heng Ji, and Jiawei Han. Automatic Entity Recognition and Typing in Massive Text Data (Conference Tutorial). SIGMOD, 2016.
- Xiang Ren*, Wenqi He*, Meng Qu, Lifu Huang, Heng Ji, Jiawei Han. AFET: Automatic Fine-Grained Entity Typing by Hierarchical Partial-Label Embedding. EMNLP, 2016.
- Xiang Ren*, Wenqi He*, Meng Qu, Heng Ji, Clare R. Voss, Jiawei Han. Label Noise Reduction in Entity Typing by Heterogeneous Partial-Label Embedding. KDD, 2016.
- Xiang Ren, Wenqi He, Ahmed El-Kishky, Clare R. Voss, Heng Ji, Meng Qu, Jiawei Han. Entity Typing: A Critical Step for Mining Structures from Massive Unstructured Text (Invited Paper). MLG, 2016.
- Xiang Ren, A. El-Kishky, C. Wang, F. Tao, C. R. Voss, H. Ji, J. Han. ClusType: Effective Entity Recognition and Typing by Relation Phrase-Based Clustering. KDD, 2015.
- Xiang Ren, Tao Cheng. Synonym Discovery for Structured Entities on Heterogeneous Graphs. WWW, 2015.
- Meng Qu, Xiang Ren, Yu Zhang, Jiawei Han. Weakly-supervised Relation Extraction by Pattern-enhanced Embedding Learning. WWW, 2018.
- Ellen Wu, Xiang Ren, Frank Xu, Ji Li, Jiawei Han. Indirect Supervision for Relation Extraction using Question-Answer Pairs
WSDM, 2018.
- Liyuan Liu*, Xiang Ren*, Qi Zhu, Shi Zhi, Huan Gui, Heng Ji, Jiawei Han. Heterogeneous Supervision for Relation Extraction: A Representation Learning Approach. EMNLP, 2017.
- Tarique A. Siddiqui*, Xiang Ren*, Aditya Parameswaran, Jiawei Han. FacetGist: Collective Extraction of Document Facets in Large Technical Corpora. CIKM, 2016.

References II

- Jialu Liu, Jingbo Shang, Chi Wang, **Xiang Ren**, Jiawei Han. Mining Quality Phrases from Massive Text Corpora. SIGMOD, 2015.
- Marina Danilevsky, Chi Wang, Nihit Desai, **Xiang Ren**, Jingyi Guo, and Jiawei Han. Automatic Construction and Ranking of Topical Keyphrases on Collections of Short Documents. SDM, 2014
- **Xiang Ren**, Yuanhua Lv, Kuansan Wang, Jiawei Han. Comparative Document Analysis for Large Text Corpora. WSDM, 2017.
- Jialu Liu, **Xiang Ren**, Jingbo Shang, Taylor Cassidy, Clare R. Voss, Jiawei Han. Representing Documents via Latent Keyphrase Inference. WWW, 2016.
- Hyungsul Kim, **Xiang Ren**, Yizhou Sun, Chi Wang, and Jiawei Han. Semantic Frame-Based Document Representation for Comparable Corpora. ICDM, 2013.
- **Xiang Ren**, J. Liu, X. Yu, U. Khandelwal, Q. Gu, L. Wang, and J. Han. ClusCite: Effective Citation Recommendation by Information Network-Based Clustering. KDD, 2014.
- X. Yu, **Xiang Ren**, Y. Sun, B. Sturt, U. Khandelwal, Q. Gu, B. Norick, and J. Han. Personalized Entity Recommendation: A Heterogeneous Information Network Approach. WSDM 2014a.
- **Xiang Ren**, Yujing Wang, Xiao Yu, Jun Yan, Zheng Chen, Jiawei Han. Heterogeneous Graph-Based Intent Learning from Queries, Web Pages and Wikipedia Concepts. WSDM 2014b.
- X. Yu, **Xiang Ren**, Y. Sun, B. Sturt, U. Khandelwal, Q. Gu, B. Norick, and J. Han. HeteRec: Entity Recommendation in Heterogeneous Information Networks with Implicit User Feedback. RecSys, 2013..
- Xiao Yu, Xiang Ren, Quanquan Gu, Yizhou Sun and Jiawei Han. Collaborative Filtering with Entity Similarity Regularization in Heterogeneous Information Networks. IJCAI-HINA, 2013.

Scalable Construction and Reasoning of Massive Knowledge Bases

Part II: Joint Representation Learning for Low-resource Information Extraction



IE

IE

IE

Structured Knowledge			
Entity	Entity	Entity	Relation
T790M	EGFR	gefitinib	Resist
Obama	U.S.		President_of
…	…		…

IE

IE

Challenges of Obtaining Training Data

- Constructing data sets is labor intensive
- Many different
  - Languages
  - Domains
  - Modalities
  - …

Joint representation learning models for *low-resource* IE.

- Learning comprehensive representations from *heterogeneous sources.*
  - *unlabeled data*
  - annotations for *related tasks, domains and languages*.
- Encoding structured knowledge to learn robust representations and make *holistic decisions*.
  - *linguistic structures*

Named Entity Recognition (NER)

- Identifying entities (in social media domain, usually person, organization, location and GPE) boundaries and their type from the plain text.

成都(GPE.NAM)电信(ORG.NAM)到底有没的时间观念哦，一托再托，日妈(PER.NOM)我们时间就不是时间哇，等了你两天啥子速度。

Chengdu(GPE.NAM) Telecom(ORG.NAM) do you have no concept of time, delay again and again, mother(PER.NOM) (curse word) our time is not time, waited for you for two days what a speed.

Structured Model for NER

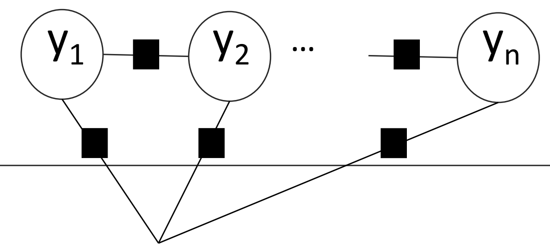
- Sequence Tagging Models:

成都(GPE.NAM)电信(ORG.NAM)
到底有没的时间观念

B I B I O O O O O O O O O O
成都电信到底有没的时间观念哦

Beginning of entities
Inside of entities
Outside of entities

Linear Chain CRF for NER

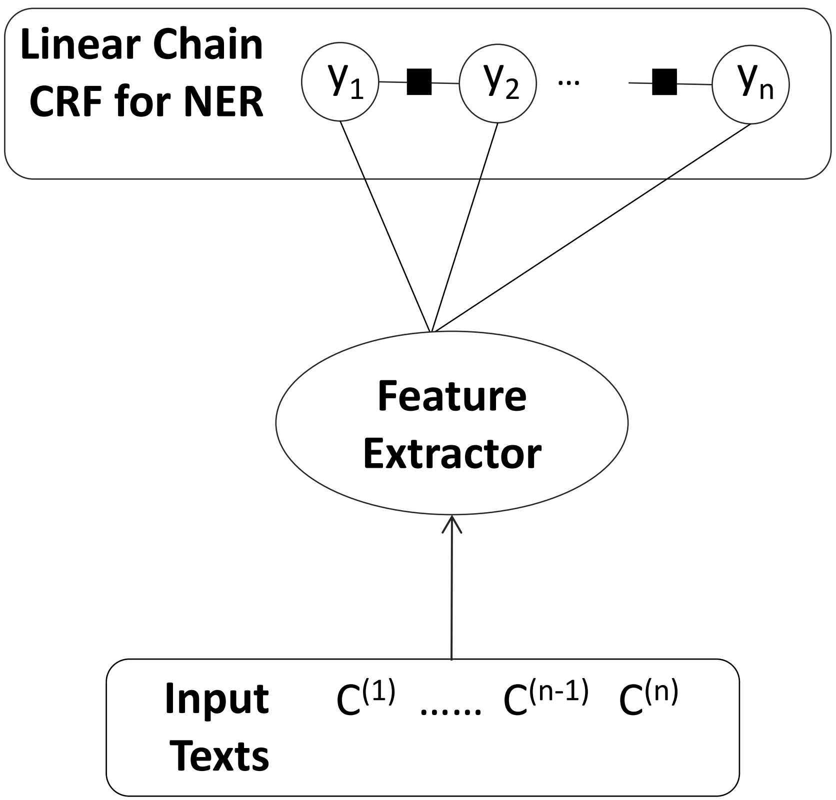


$$P(y\,|\,x)=\frac{1}{Z(x)}\prod_{t=1}^{T}\exp\{\sum_{k=1}^{K}\theta_k f_k(y_t,y_{t-1},x_t)\}$$

make joint decisions over a sequence

McCallum and Li, 2003

Representation Learning for NER

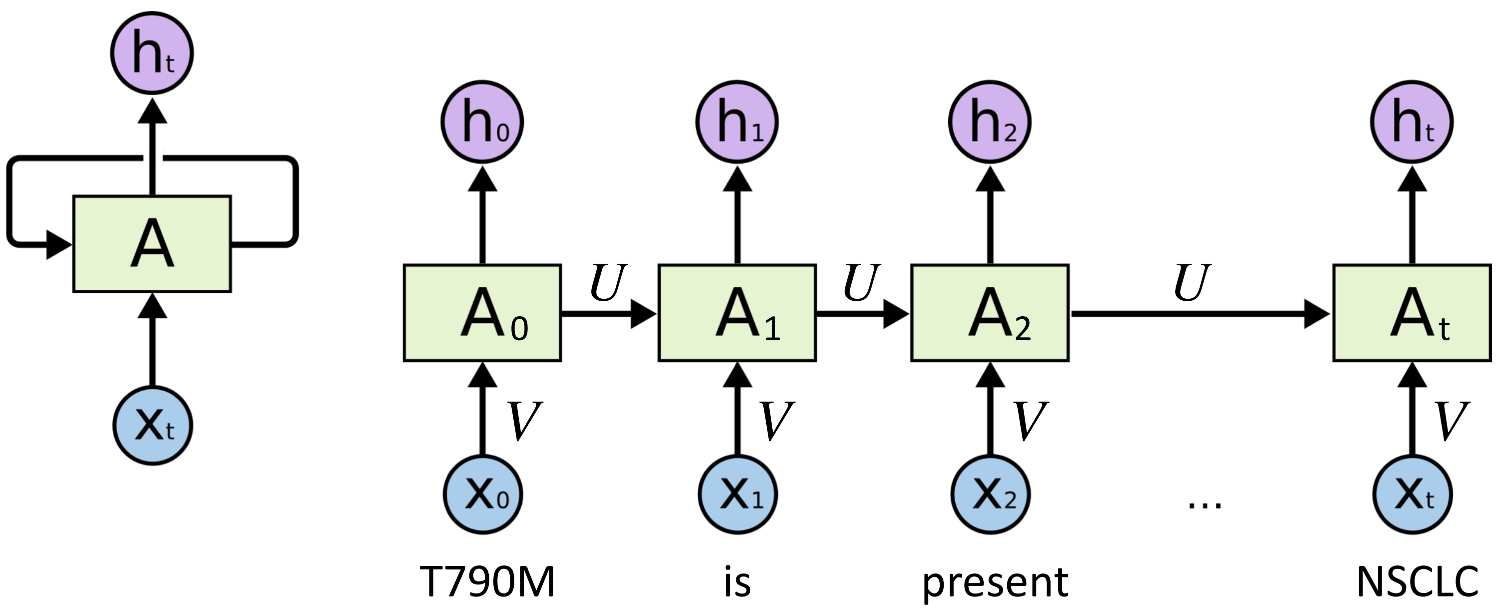


- Recurrent Neural Networks (RNNs) for Representation :
  - Automatically learns data representations for features
  - Model input dependencies.

Huang et. al., 2015 Lample et. al., 2016 Ma and Hovy., 2016

Representation Learning for NER

Recurrent Neural Networks (RNNs)

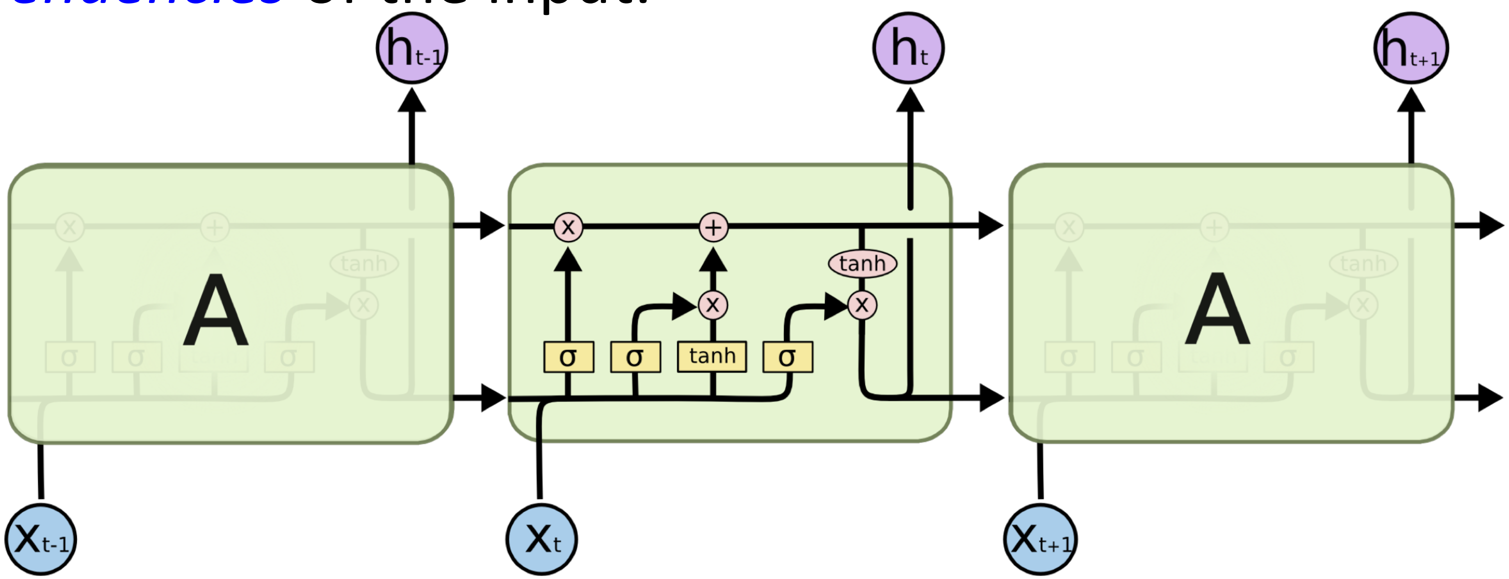


$$A_t = g(Vx_t + UA_{t-1} + c)$$

Very deep neural network, back propagation training

Long-Short Term Memory Networks (LSTMs)

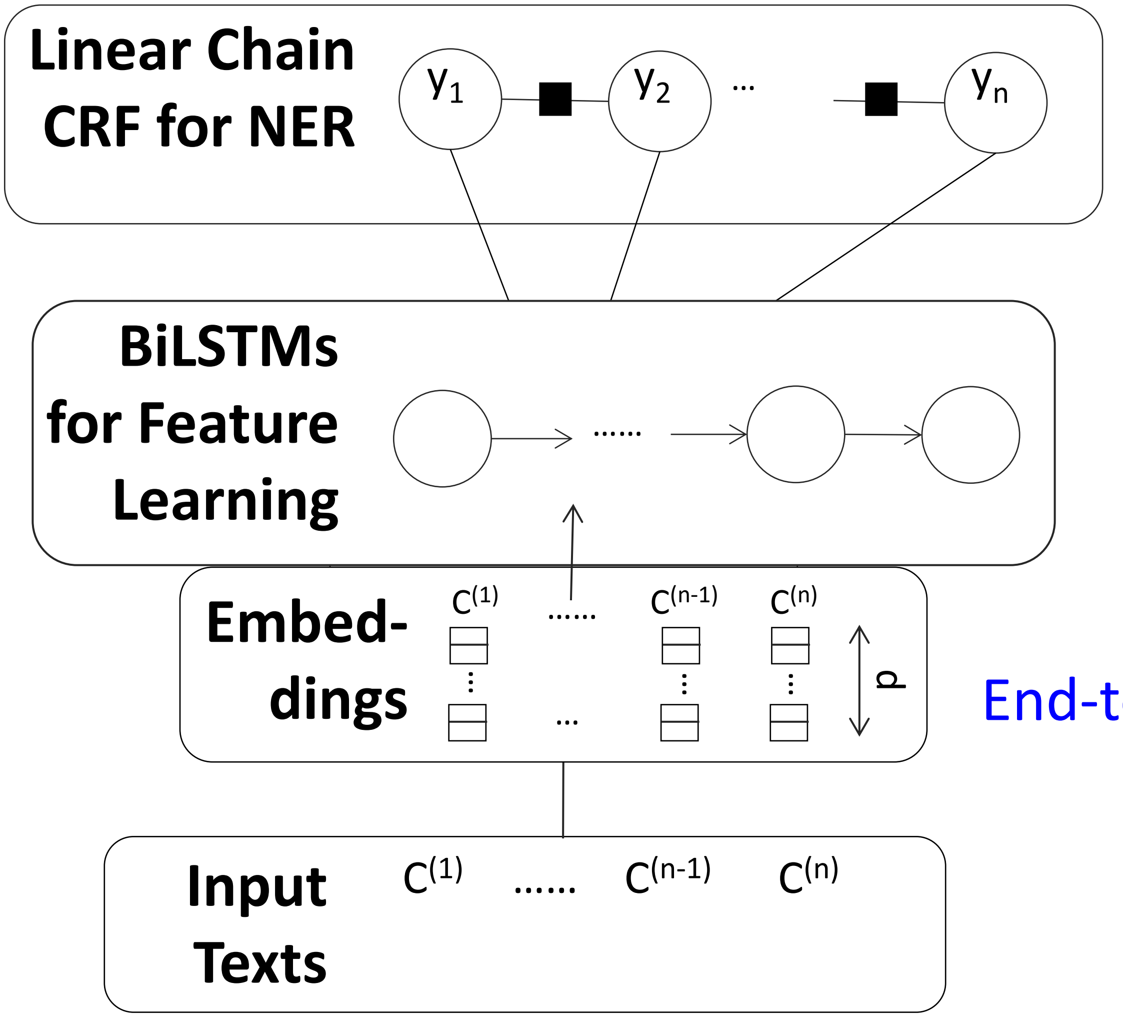
LSTMs are special RNNs that use gates to control the information flow and essentially capture *long-term dependencies* of the input.



Very deep neural network, back propagation training

Picture credit: colah's blog, 2015

Neural Sequence Tagging Models

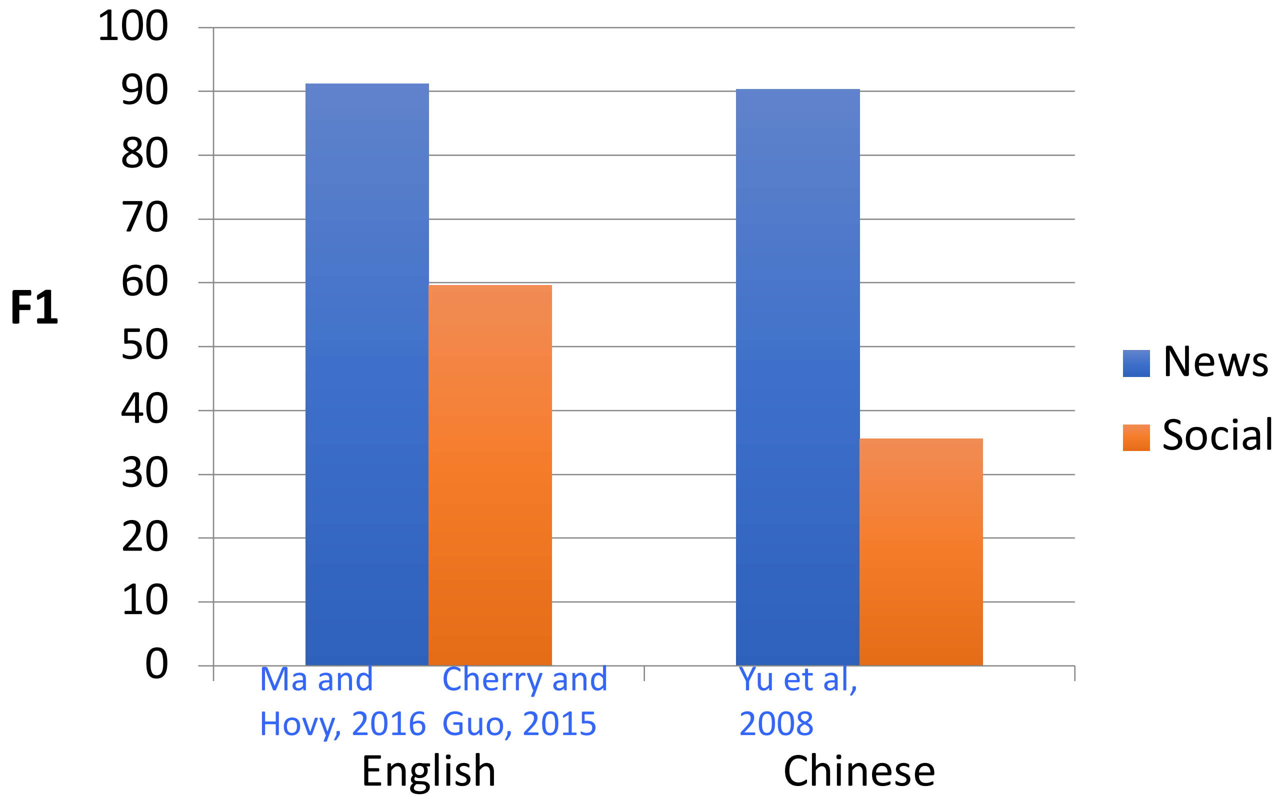


End-to-end training

Huang et. al., 2015 Lample et. al., 2016 Ma and Hovy., 2016

Challenges for low-resource settings

- HUGE gap on social media (noisy) v.s news text:
  - informal language and insufficient annotations.



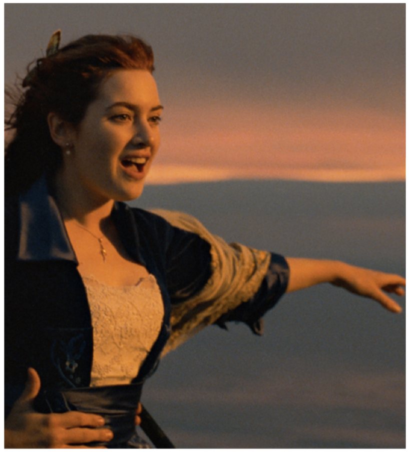
Ideas

- Leverage existing resources to learn representations that generalize across multiple types of data.
  - Multi-task Learning.
  - Domain Adaptation.
  - Cross-lingual Transfer.

Distributional Similarity of Words

Generalizability

- Rose



- Violet

Peng and Dredze, 2015

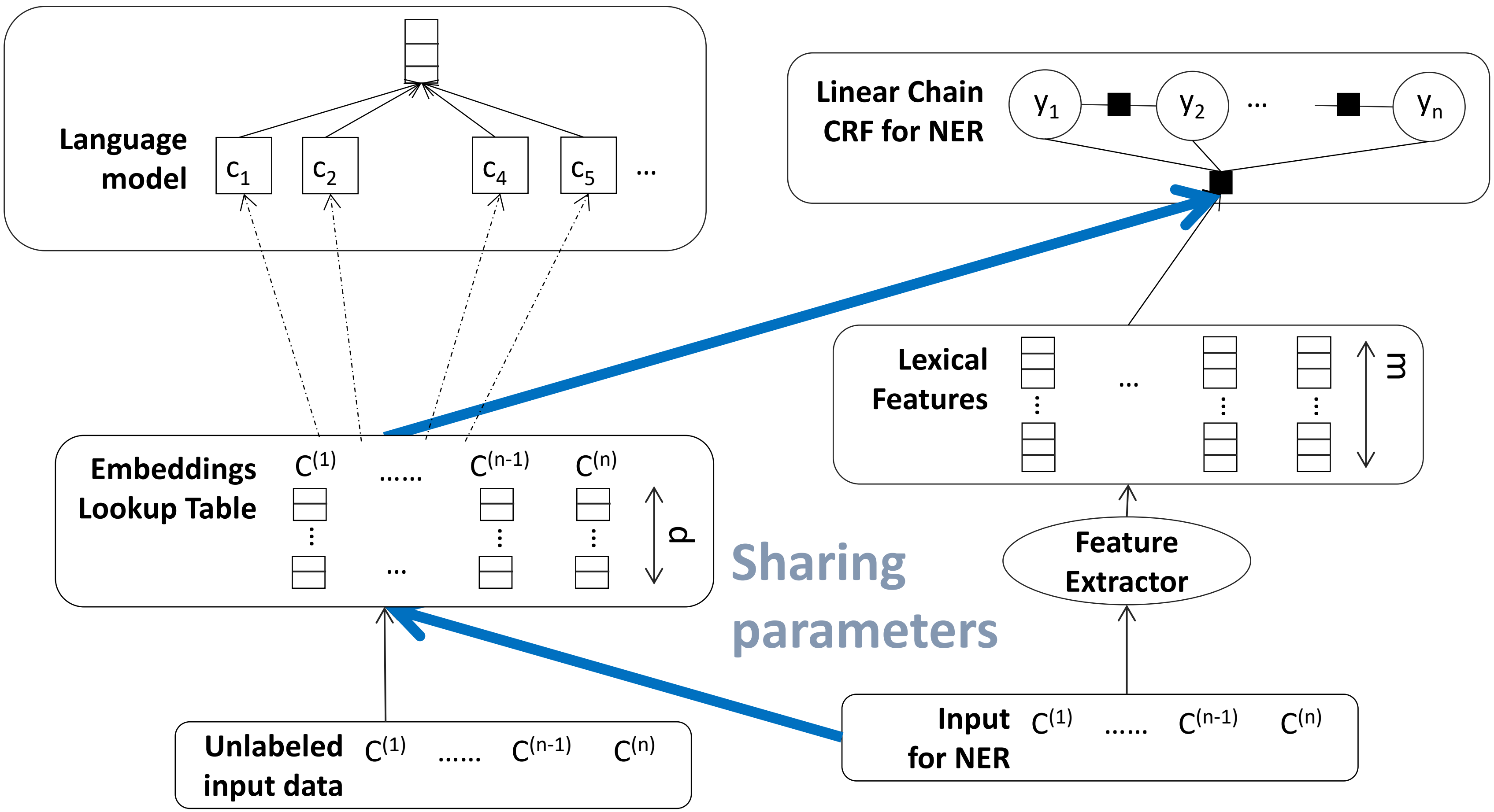
Joint Learning of Word Embeddings and Named Entity Recognition

Model for Learning Word Representations

Model for Named Entity Recognition

Peng and Dredze, 2015

Joint Learning of Word Embeddings and Named Entity Recognition



Peng and Dredze, 2015

Joint Learning of Word Embeddings and Named Entity Recognition

Skip-gram model to learn word representations

Shared Parameters

$$L_u(X, \boxed{e_x}) = \frac{1}{T}\sum_{t=1}^{T}\sum_{-c\le j\le c, j\ne 0}\log p(x_{t+j} \mid x_t)$$

$$p(x_i \mid x_j) = \frac{\exp(e_{xi}^T e_{xj})}{\sum_{i'}\exp(e_{xi'}^T e_{xj})}$$

Log-bilinear CRF model for named entity recognition

$$L_s(Y \mid X; \theta, \boxed{e_x}) = \frac{1}{T}\sum_{t=1}^{T}\log p(y_t \mid x_t)$$

$$P(y \mid x) = \frac{1}{Z(x)}\prod_{t=1}^{T}\exp\{\sum_{k=1}^{K}\theta_k f_k(y_t, y_{t-1}, x_t, e_x)\}$$

2 millions of unannotated weibo message for training

1350 NER annotated weibo message for training

Chinese Word Boundaries

成都(GPE.NAM)电信(ORG.NAM)到底有没的时间观念哦，一托再托，日妈(PER.NOM)我们时间就不是时间哇，等了你两天啥子速度。

Chengdu(GPE.NAM) Telecom(ORG.NAM) do you have no concept of time, delay again and again, mother(PER.NOM) (curse word) our time is not time, waited for you for two days what a speed.

成都(GPE.NAM) / 电信(ORG.NAM) / 到底/ 有/ 没的/ 时间 / 观念/ 哦/ ，/ 一/ 托/ 再/ 托/，/ 日/ 妈(PER.NOM) / 我们/ 时间/ 就/ 不/ 是/ 时间/ 哇/ ，/ 等/ 了/ 你/ 两/ 天/ 啥子/ 速度/ 。/

Peng and Dredze, 2016

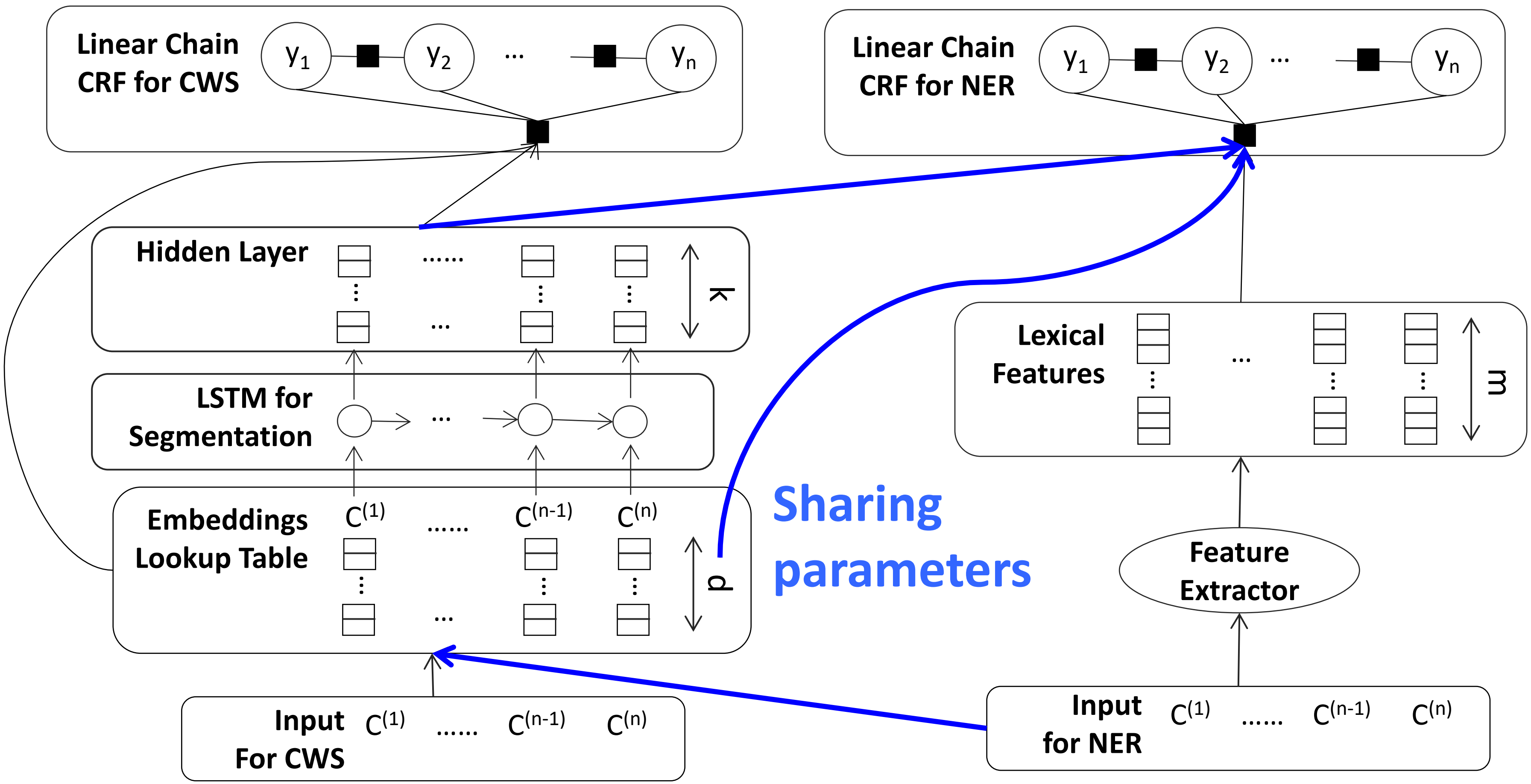
Multi-task Learning of Word Segmentation and Named Entity Recognition

Model for Chinese Word Segmentation

Model for Named Entity Recognition

http://www.cs.jhu.edu/~npeng

Multi-task Learning of Word Segmentation and Named Entity Recognition



Peng and Dredze, 2016

Domains for Languages

McDonald *'s* Seeks Its **Fast-Food** Soul

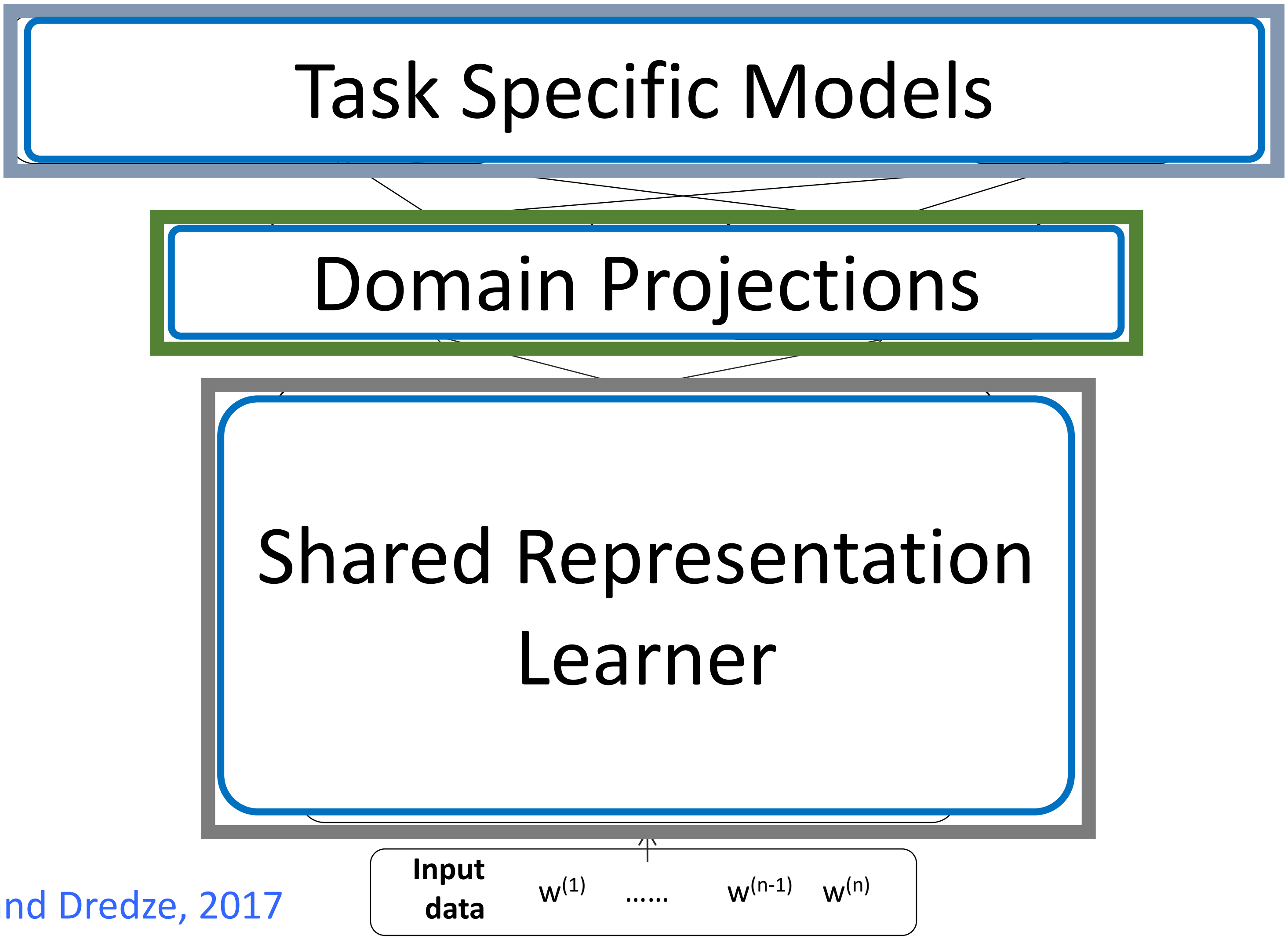
- *NYTimes 3/7/2015*

Nivre and ***McDonald*** (2008) used the output of one dependency **parser** to provide features for another.

- Stacking Dependency Parsers, Martins+ (EMNLP 2008)

Peng and Dredze, 2017

Multi-task Multi-domain Learning



Peng and Dredze, 2017

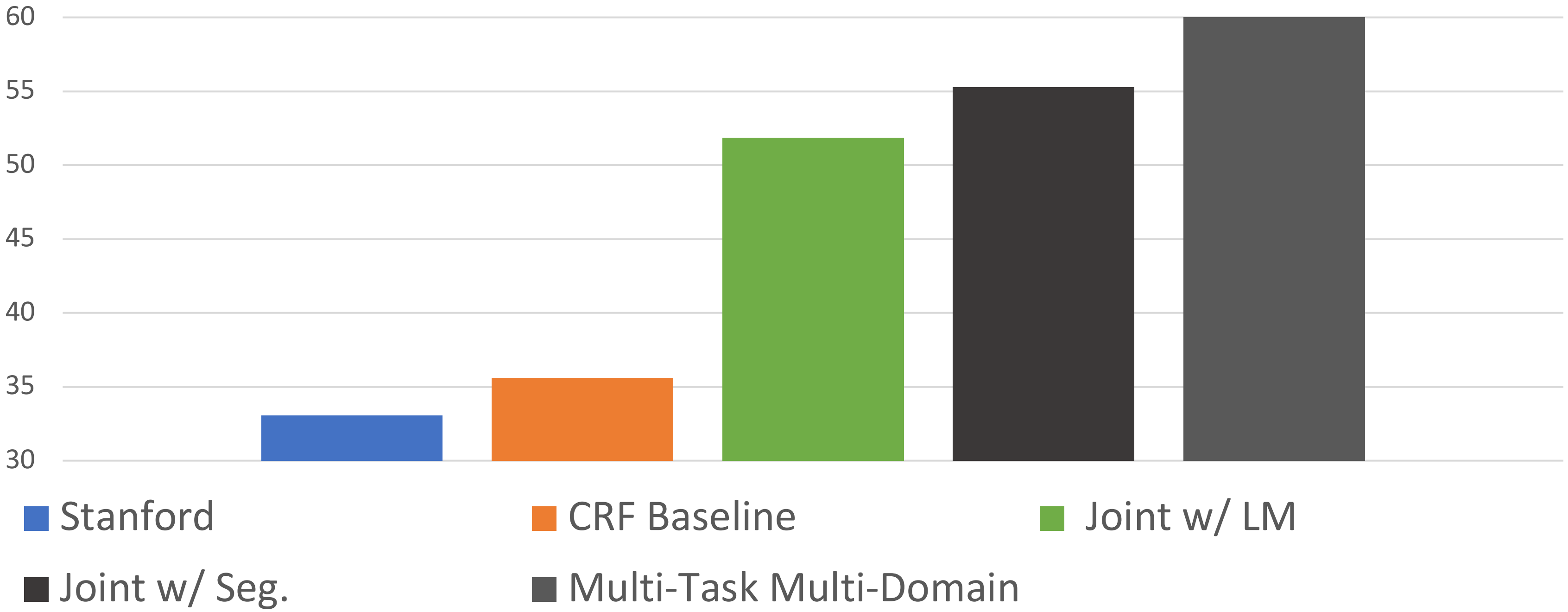
Multi-task Multi-domain learning for sequence tagging

- Domains: news and social media
- Tasks: Chinese word segmentation and NER
- Datasets:

Dataset	#Train	#Dev	#Test
News CWS	39,567	4,396	4,278
News NER	16,814	1,868	4,636
Social CWS	1,600	200	200
Social NER	1,350	270	270

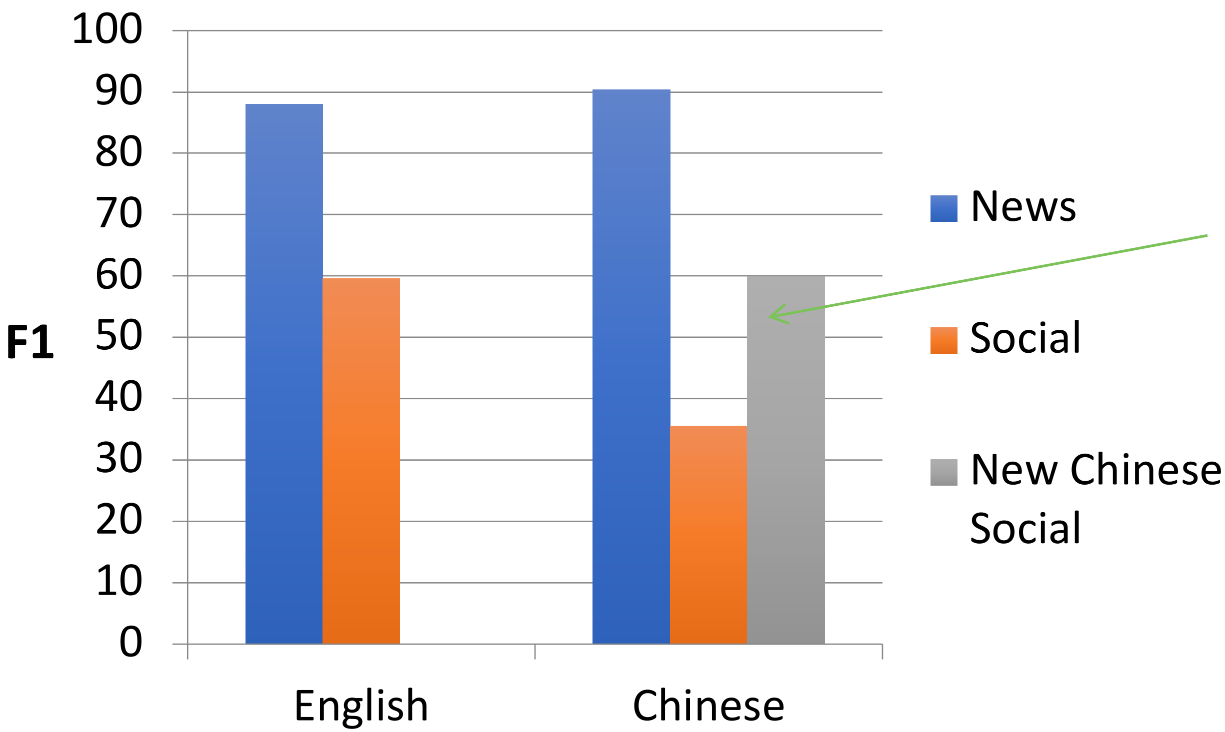
2 millions of unannotated weibo message for training

NER on Chinese Social Media



Named Entity Recognition
on Chinese social media

Closing The Gap



Ours new Chinese Social NER system leveraging heterogeneous sources for representation learning

Peng and Dredze, 2017

How to build NER for a new language using
(1) Comparable Corpora
(2) English NER tagger

維基百科
自由的百科全書

約翰·霍普金斯大學

約翰·霍普金斯大學 是一所主校區位於美國馬里蘭州巴爾的摩市的研究型私立大學。截止至2012年，共有36名校友獲諾貝爾獎[3]。

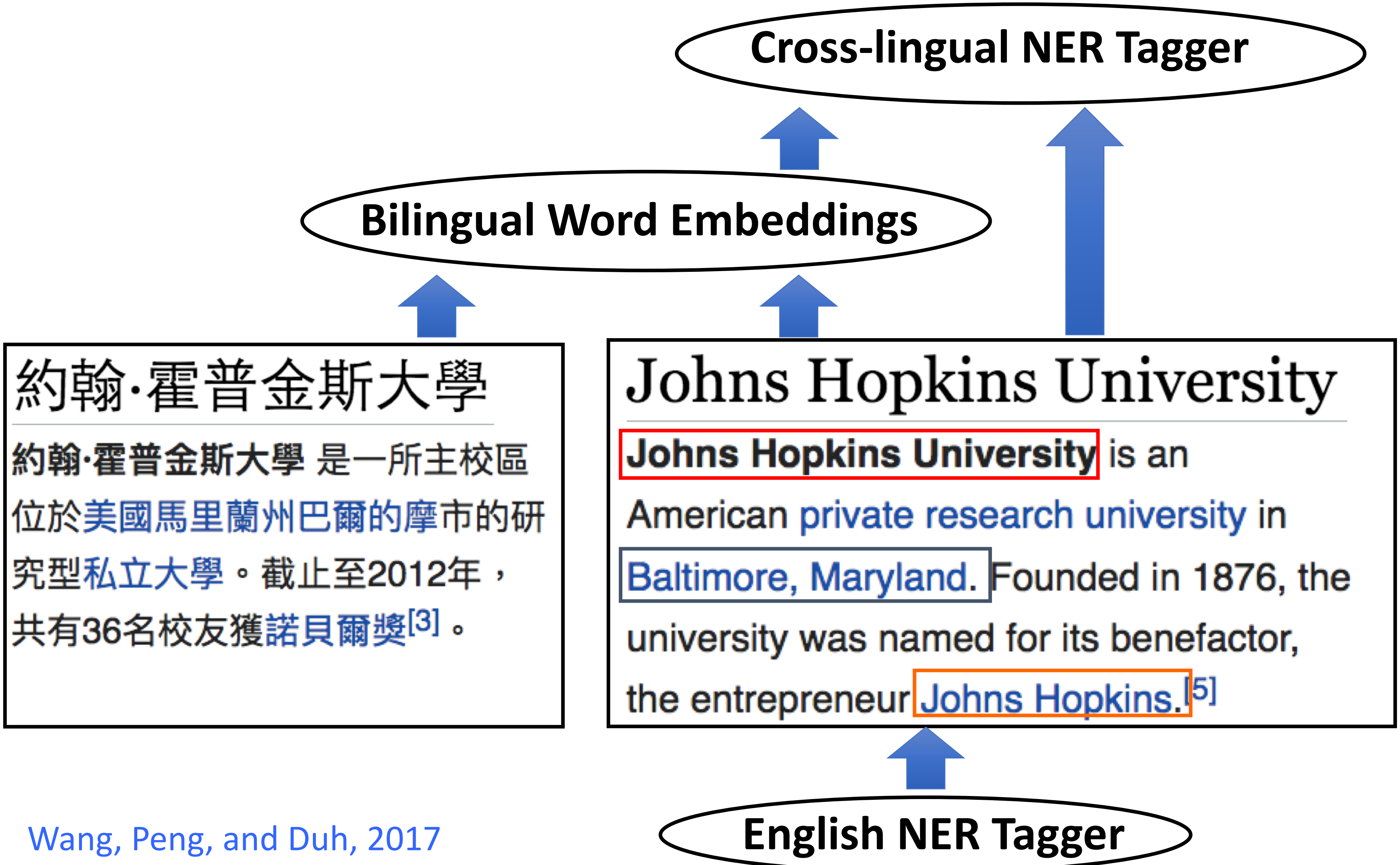
WIKIPEDIA
The Free Encyclopedia

Johns Hopkins University

Johns Hopkins University is an American private research university in Baltimore, Maryland. Founded in 1876, the university was named for its benefactor, the entrepreneur Johns Hopkins.[5]

Wang, Peng, and Duh, 2017

Idea



Wang, Peng, and Duh, 2017

Training Bilingual Word Embeddings

Word2Vec

Mixed-Language Pseudo-Document

Johns 約翰·霍普金斯 Hopkins University 大學 is 是一所 an American 主校區位於 private research university 美國 巴爾的摩 市的研究型 in Baltimore,

約翰·霍普金斯大學

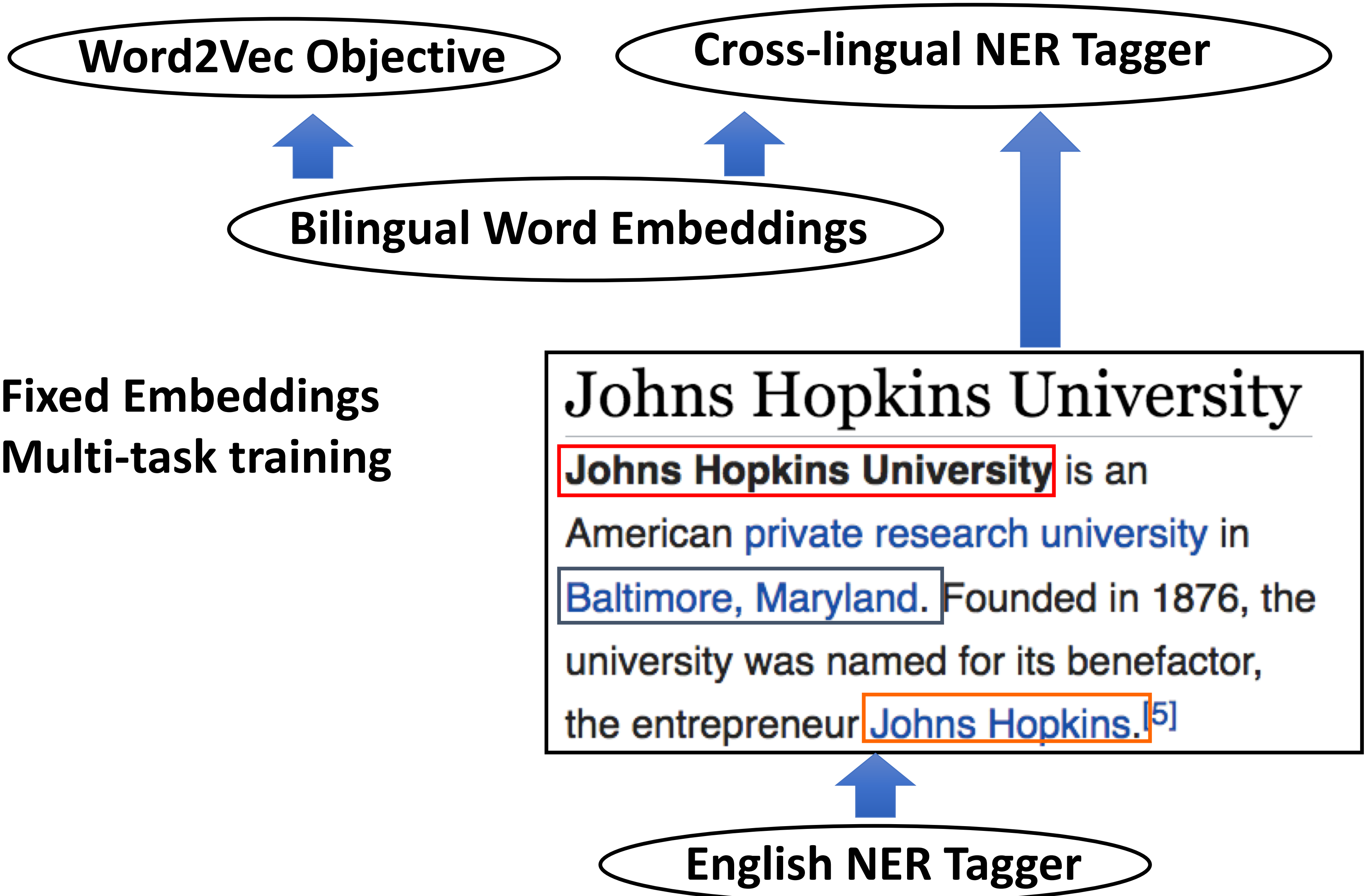
約翰·霍普金斯大學 是一所主校區位於美國馬里蘭州巴爾的摩市的研究型私立大學。截止至2012年，共有36名校友獲諾貝爾獎[3]。

Johns Hopkins University

Johns Hopkins University is an American private research university in Baltimore, Maryland. Founded in 1876, the university was named for its benefactor, the entrepreneur Johns Hopkins.[5]

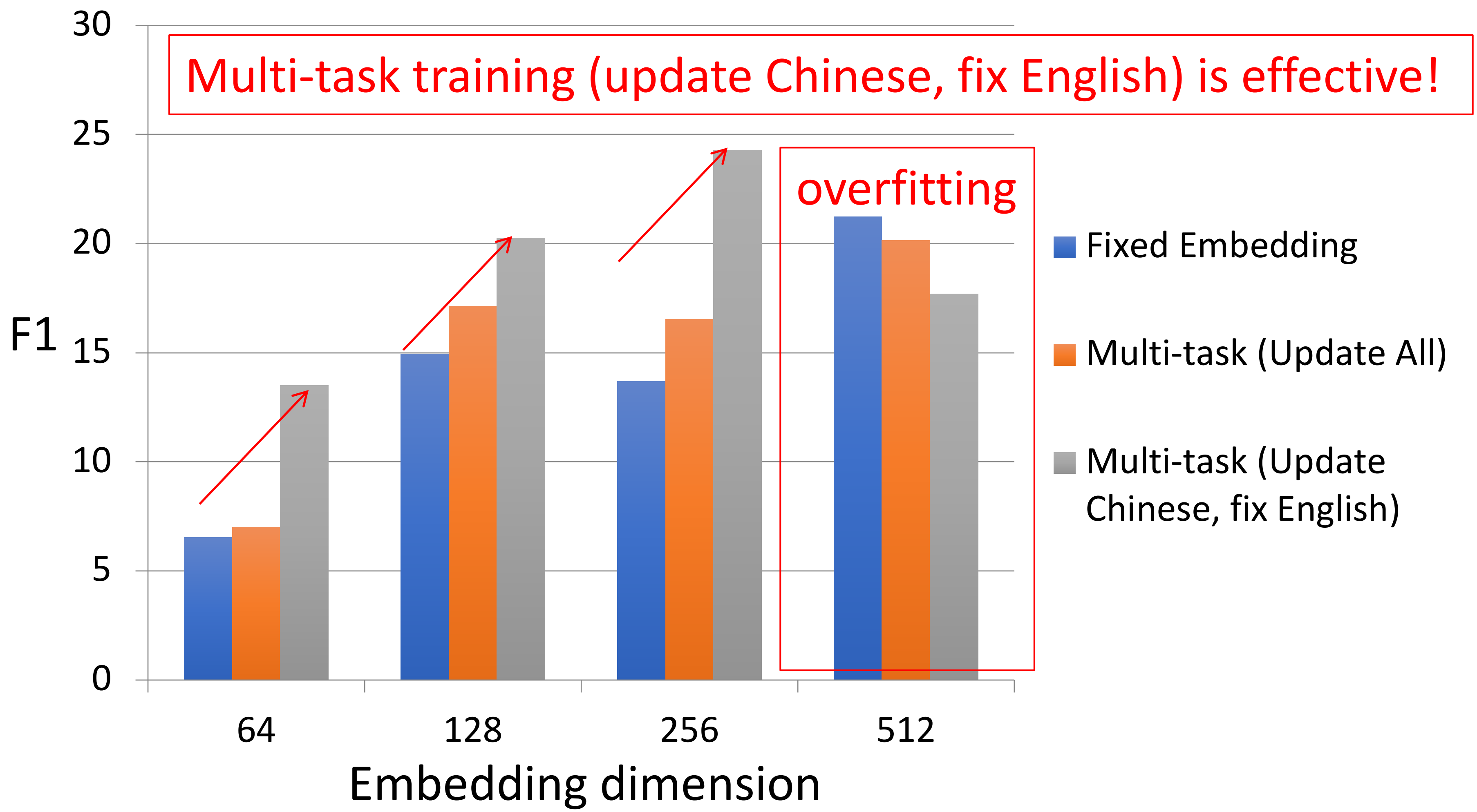
Vulic and Moens, 2015

Training Cross-lingual NER Tagger



1. **Fixed Embeddings**
2. **Multi-task training**

Results (F1 score)



Joint representation learning models for *low-resource* IE.

- Learning comprehensive representations from *heterogeneous sources.*
  - *unlabeled data*
  - annotations for *related tasks, domains and languages*.
- Encoding structured knowledge to learn robust representations and make *holistic decisions*.
  - *linguistic structures*

Cross-Sentence N-ary Relation Extraction

Mutation

T790M is present as a minor clone in NSCLC ,

and may be selected for during therapy .

This mutation has been shown to prevent the

Drug

activation of BIM in response to gefitinib but can

Gene

be overcome by an irreversible inhibitor of EGFR.

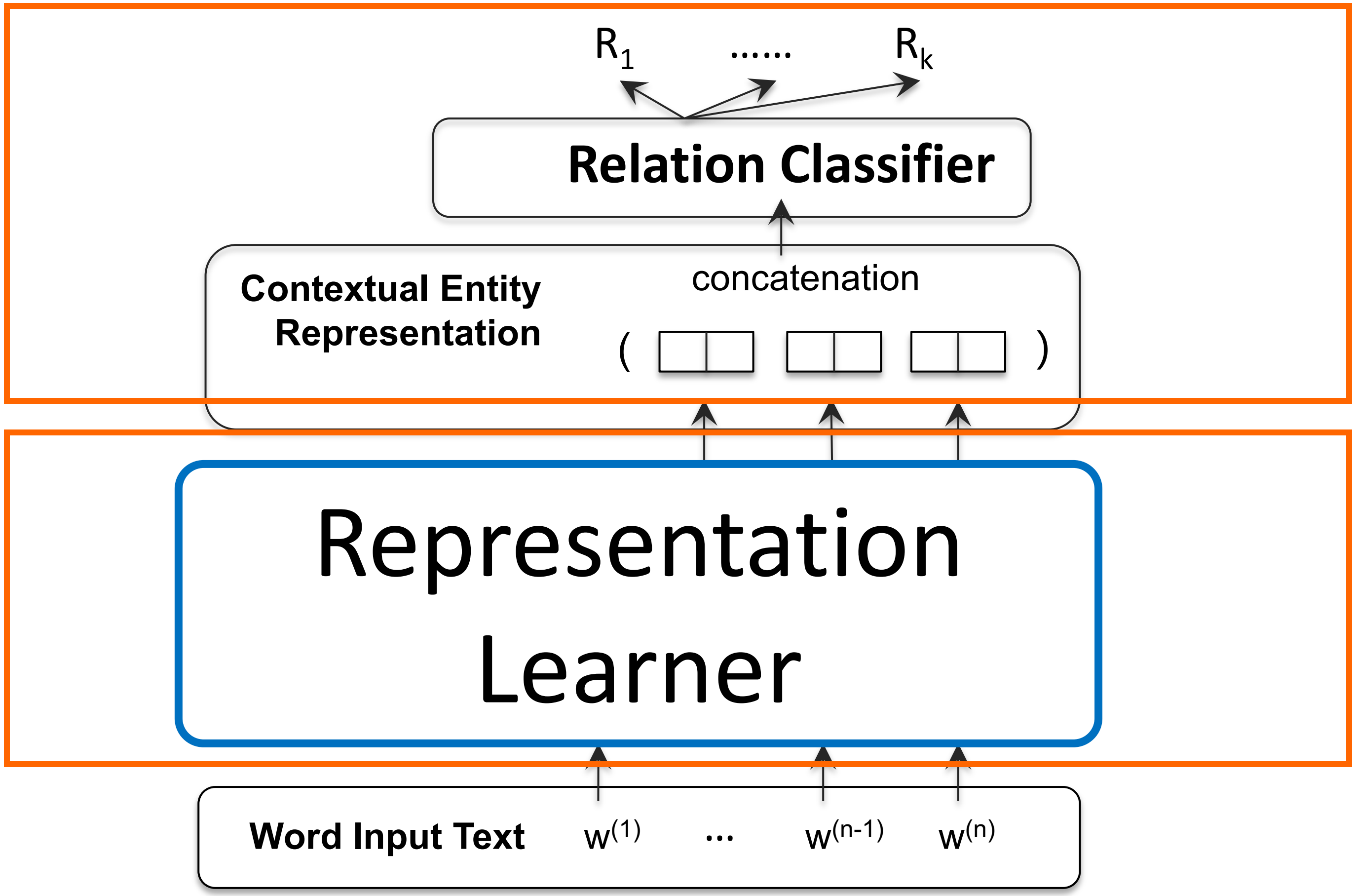
Knowledge Bases for Drug-Gene-Mutation Interaction

- People manually curate drug-gene-mutation interaction databases for precision medicine:
  - Gene Drug Knowledge Database (GDKD) (Dienstmann et al., 2015)
  - Clinical Interpretations of Variants in Cancer (CiViC) (Washington University School of Medicine)

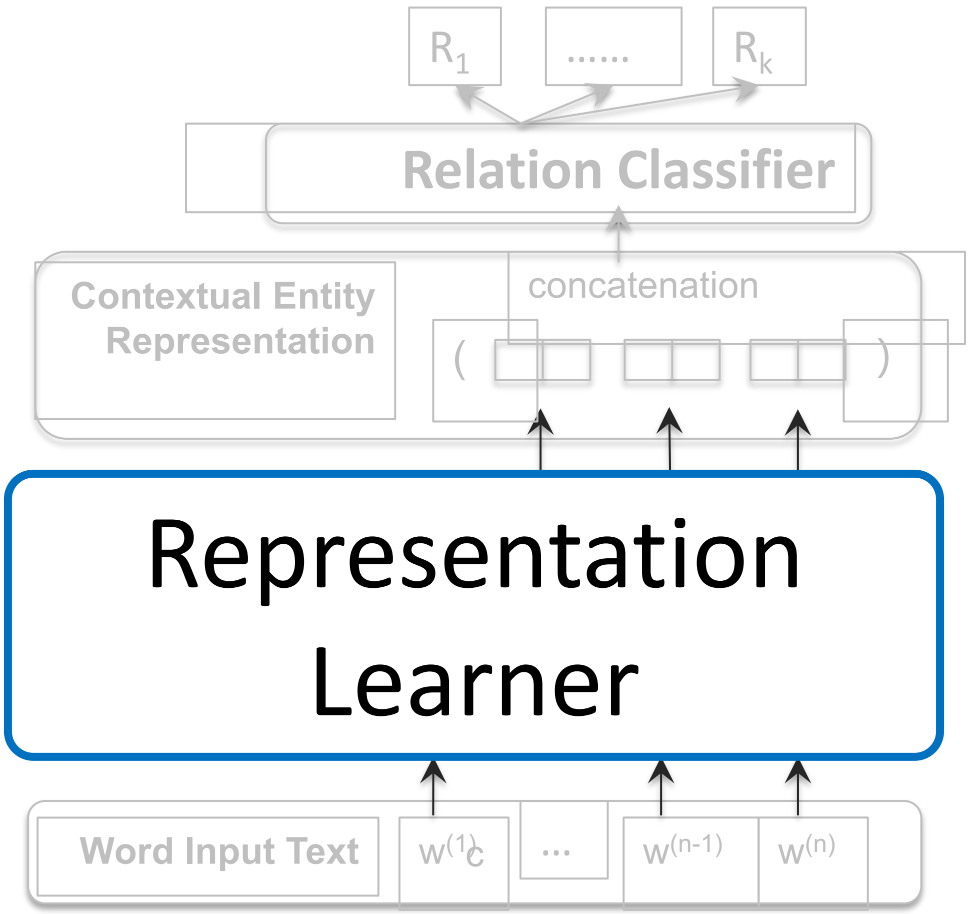
Special Challenges

- N-ary relations:
  - Traditional feature-based classification method usually use features defined on the *shortest syntactic dependency paths* between two entities.
  - Such features are hard to define in the *N*-ary case.
- Cross sentence relations:
  - Traditional features become sparser and learning becomes harder.

A Representation Learning Framework



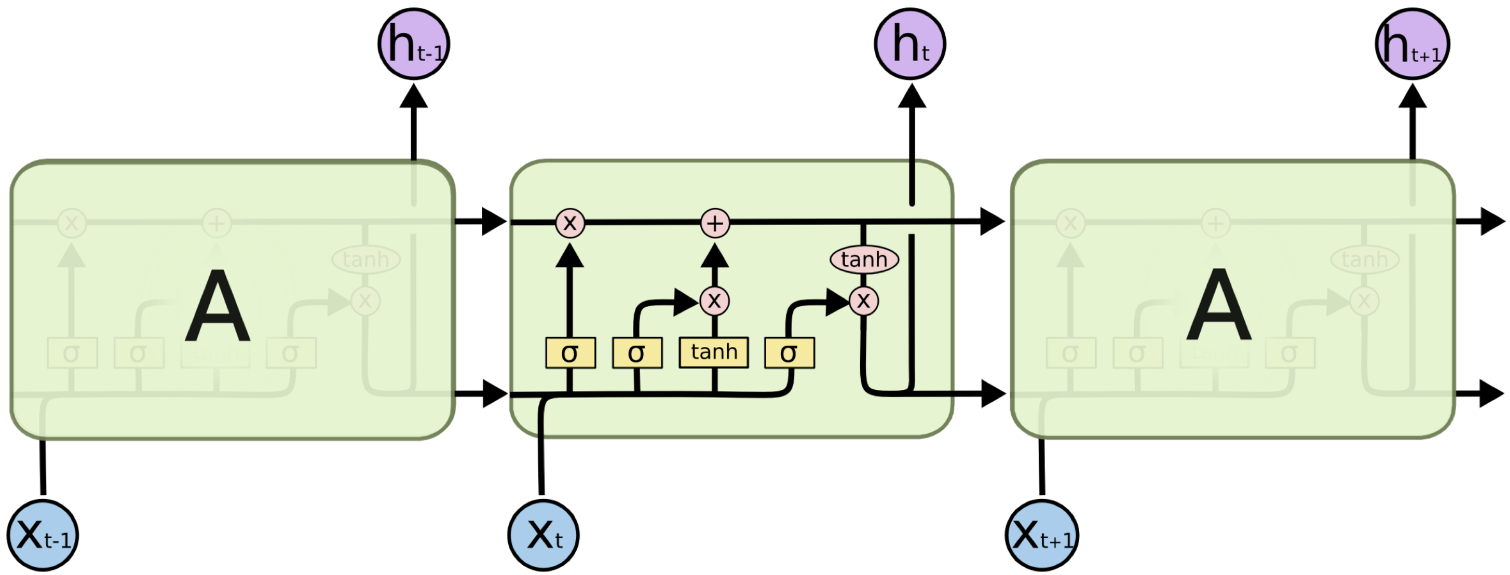
Peng et. al., 2017



R1
......
Rk
Relation Classifier
Contextual Entity Representation
concatenation
(
)
Representation Learner
Word Input Text
w(1)
...
w(n-1)
w(n)

Long-Short Term Memory Networks (LSTMs)

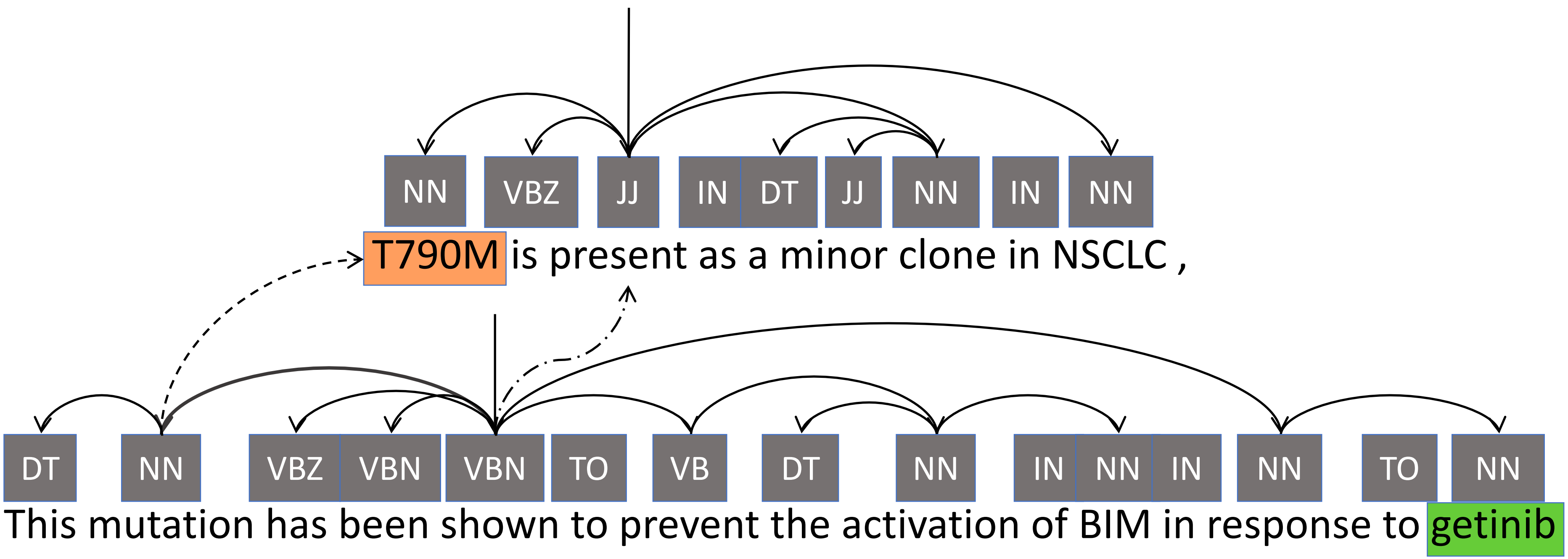
Capture *long-term dependencies* of the input.



However, it still only explicitly models the dependencies between the adjacent inputs.

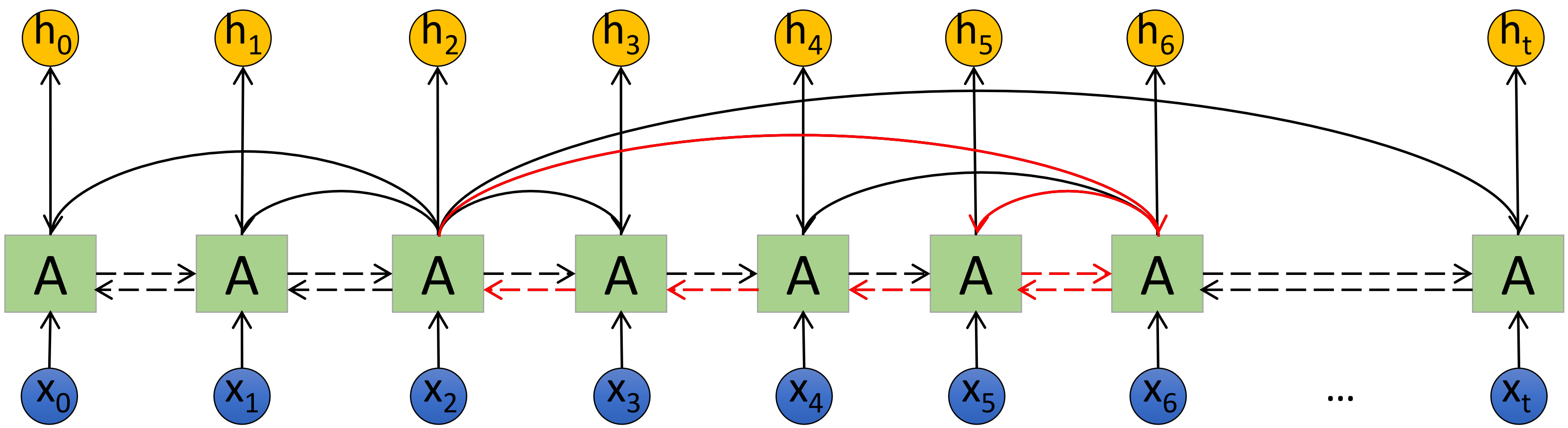
Picture credit: colah's blog, 2015

Linguistics Structures for Input Texts



Peng et. al., 2017

Directed Cyclic Graph



Peng et. al., 2017

Graph Long Short-Term Memory Networks (Graph LSTMs)

- Goals:
  - *different types* of dependencies: adjacency, *syntactic* dependencies, *coreferences*, and *discourse* relations.
  - *long-distance* dependencies: entities span sentences.
- Challenges: how to define a neural architecture over a cyclic graph?

Work beyond Linear-Chain

- NLP: Tree LSTM (Tai et. al. 2015, Miwa and Bansal, 2016)

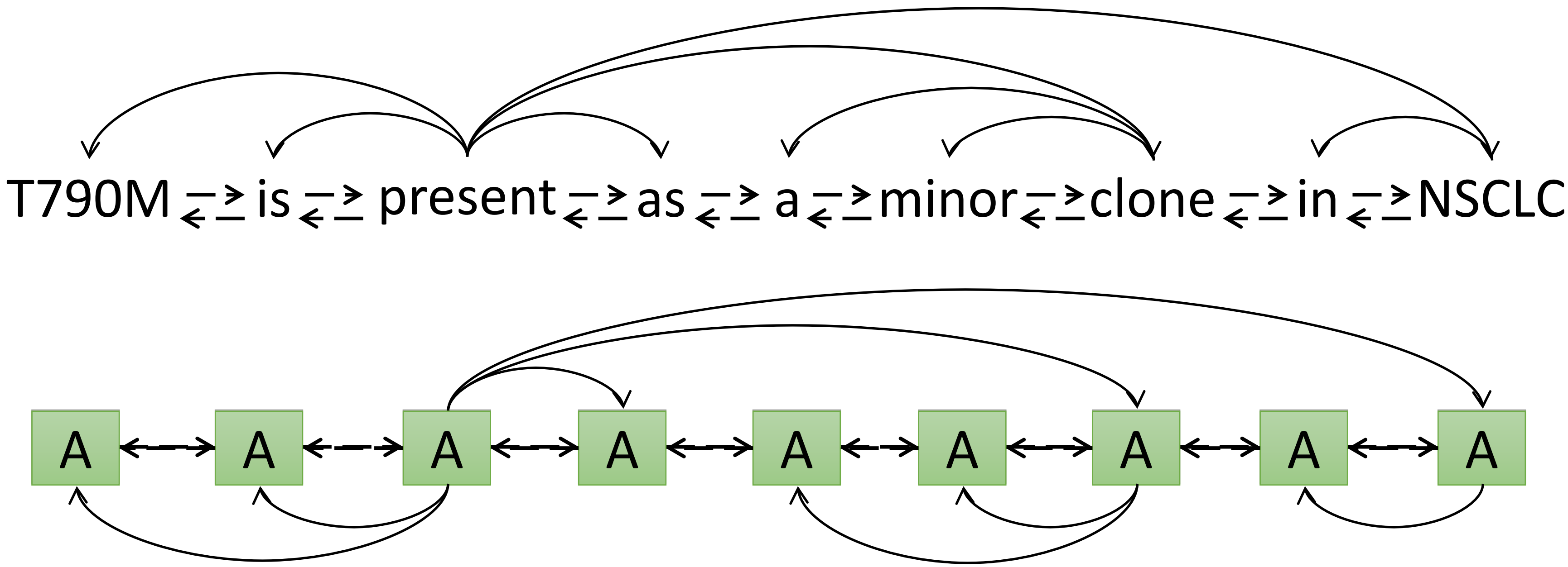
- Programming verification: Gated Graph Neural Network (Li et. al. 2016)
- Graph Convolutional Networks (Kipf and Welling, 2017)

Challenge in Training

- Existing approach
  - Unroll recurrence for a number of steps
  - Analogous to loopy belief propagation (LBP)
- Problems
  - Expensive: Many steps per epoch
  - Information does not propagate from distant nodes

Training Graph LSTMs

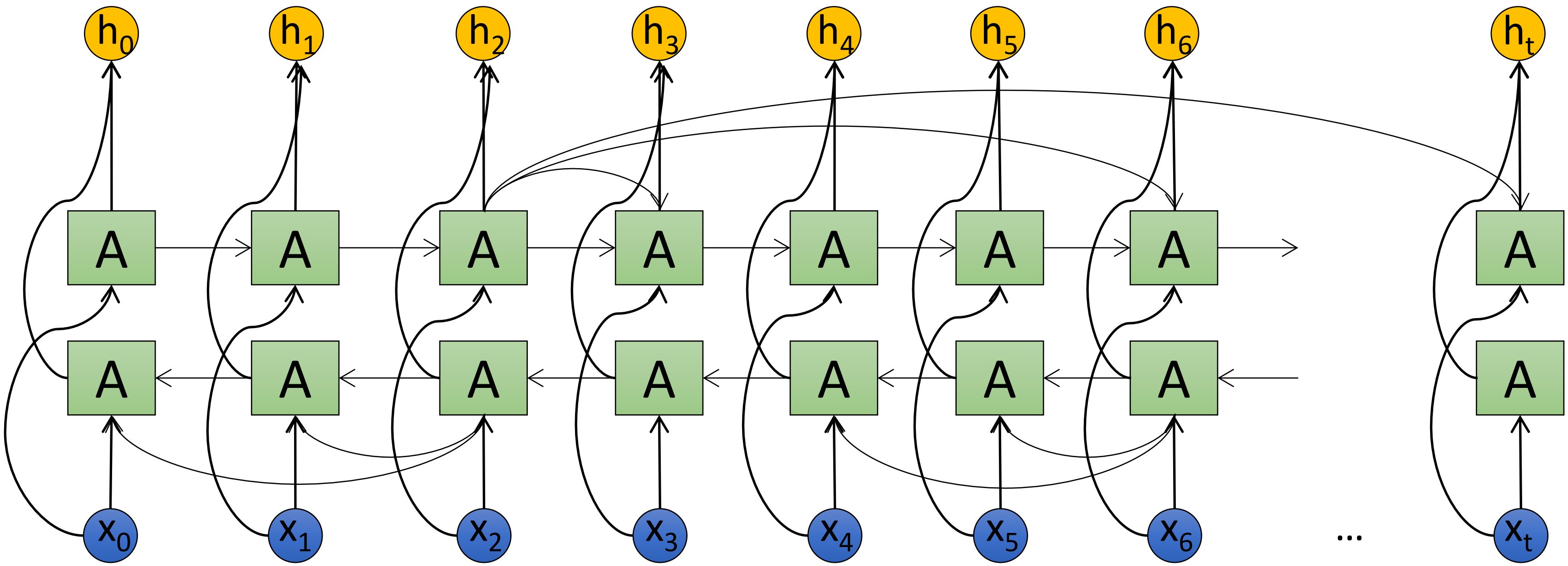
- *Provably*, all directed cyclic graph without self-loop can be decomposed into two DAGs.



Peng et. al., 2017

Training Graph LSTMs

- Approximate a cyclic graph by two directed acyclic graphs (DAGs), and stack the DAGs.



Topological order is well-defined, back propagation training

Chain LSTMs v.s. Graph LSTMs

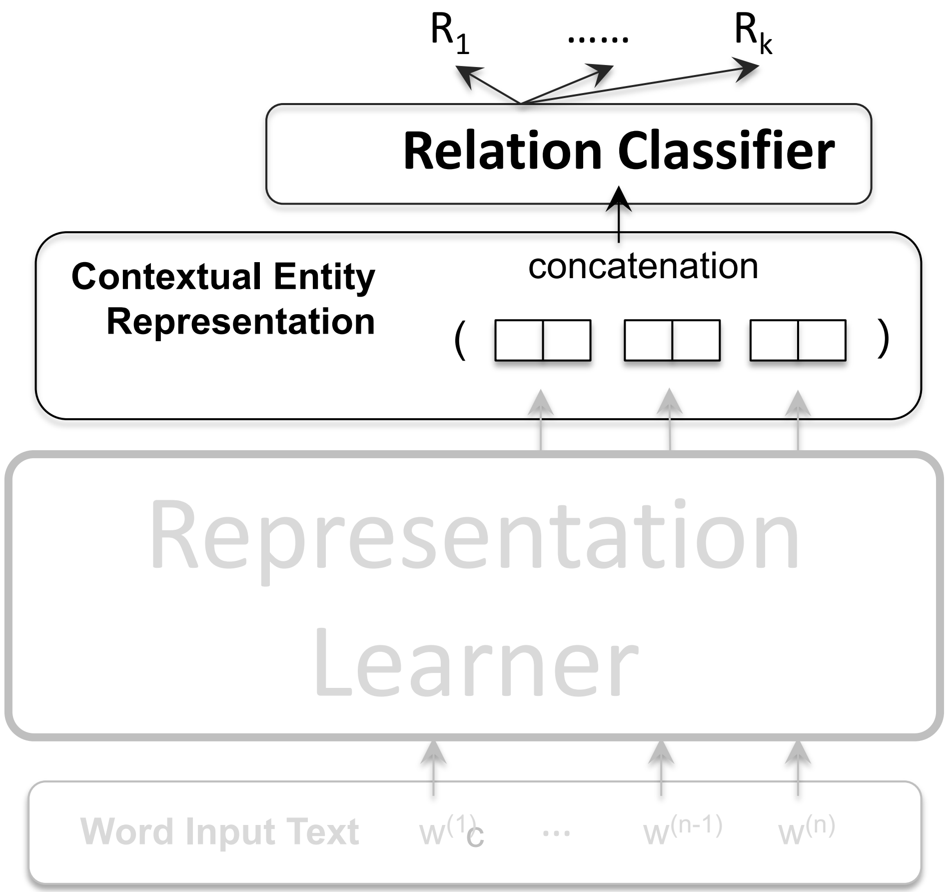
Linear-chain LSTM

$$
\begin{aligned}
i_t &= \sigma(W_i x_t + \boxed{U_i h_{t-1} + b_i}) \\
o_t &= \sigma(W_o x_t + \boxed{U_o h_{t-1} + b_o}) \\
\tilde{c}_t &= \tanh(W_c x_t + \boxed{U_c h_{t-1}} + b_c) \\
\boxed{f_t} &= \sigma(W_f x_t + U_f h_{t-1} + b_f) \\
c_t &= i_t \odot \tilde{c}_t + \boxed{f_t \odot c_{t-1}} \\
h_t &= o_t \odot \tanh(c_t)
\end{aligned}
$$

Graph LSTM (one DAG)

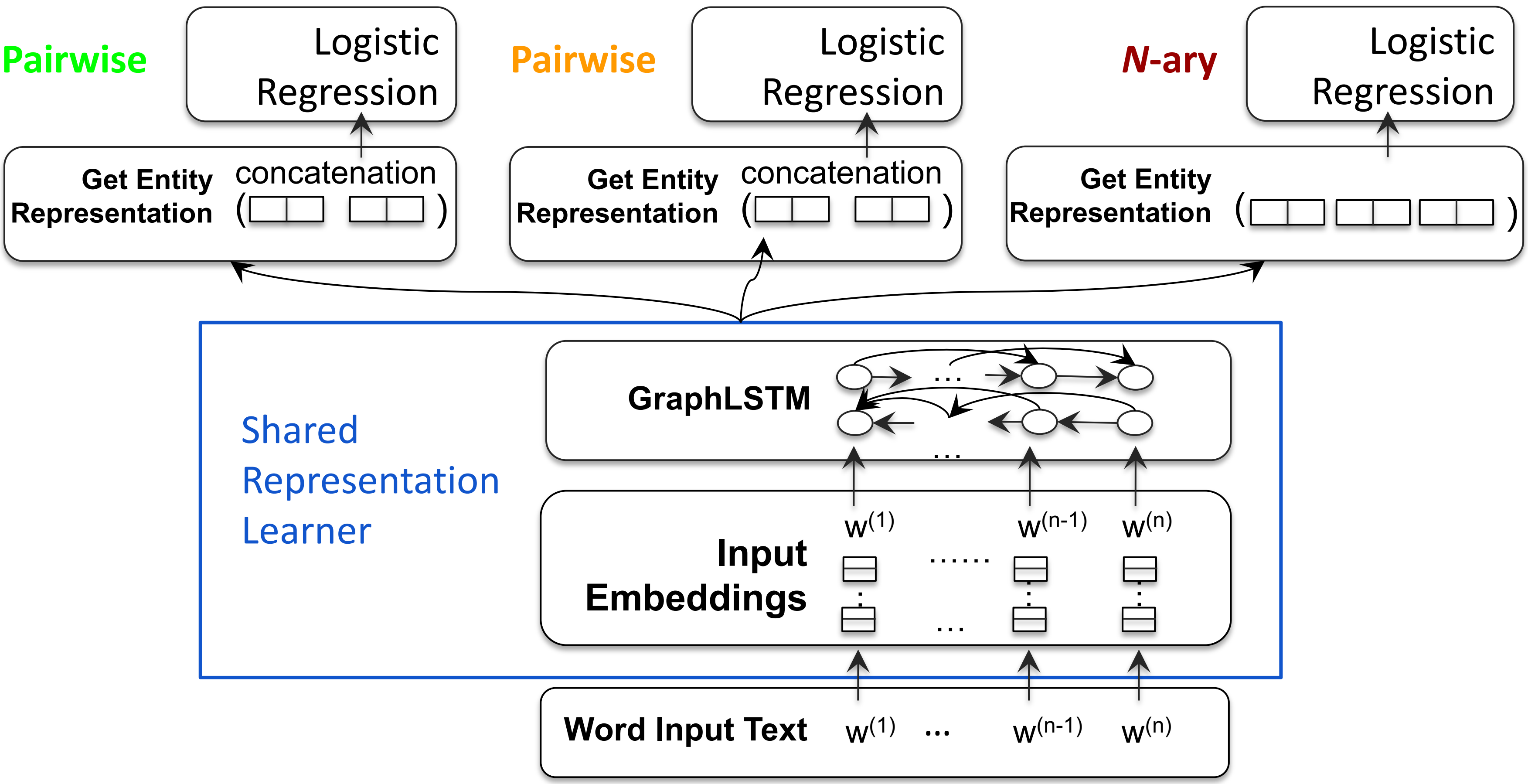
$$
\begin{aligned}
i_t &= \sigma(W_i x_t + \boxed{\sum_{j \in P(t)} U_i^{m(t,j)} h_j + b_i}) \\
o_t &= \sigma(W_o x_t + \boxed{\sum_{j \in P(t)} U_o^{m(t,j)} h_j + b_o}) \\
\tilde{c}_t &= \tanh(W_c x_t + \boxed{\sum_{j \in P(t)} U_c^{m(t,j)} h_j} + b_c) \\
\boxed{f_{tj}} &= \sigma(W_f x_t + U_f^{m(t,j)} h_j + b_f) \\
c_t &= i_t \odot \tilde{c}_t + \boxed{\sum_{j \in P(t)} f_{tj} \odot c_j} \\
h_t &= o_t \odot \tanh(c_t)
\end{aligned}
$$

Multi-task Learning



Peng et. al., 2017

Multi-task Learning



Peng et. al., 2017

Domain: Molecular Tumor Board

- Ternary interaction: (drug, gene, mutation)
- Distant supervision
  - Knowledge bases: GDKD + CIVIC
  - Text: PubMed Central articles (~ 1 million full-text articles)
- We got 3,462 paragraphs about drug-gene-mutation relations from distant supervision.

Absolute Recall

	Drug	Gene	Mutation	Interaction
DGKD + CiViC	16	12	41	59
Single-Sent	68	228	221	530
Cross-Sent	103	512	445	1461

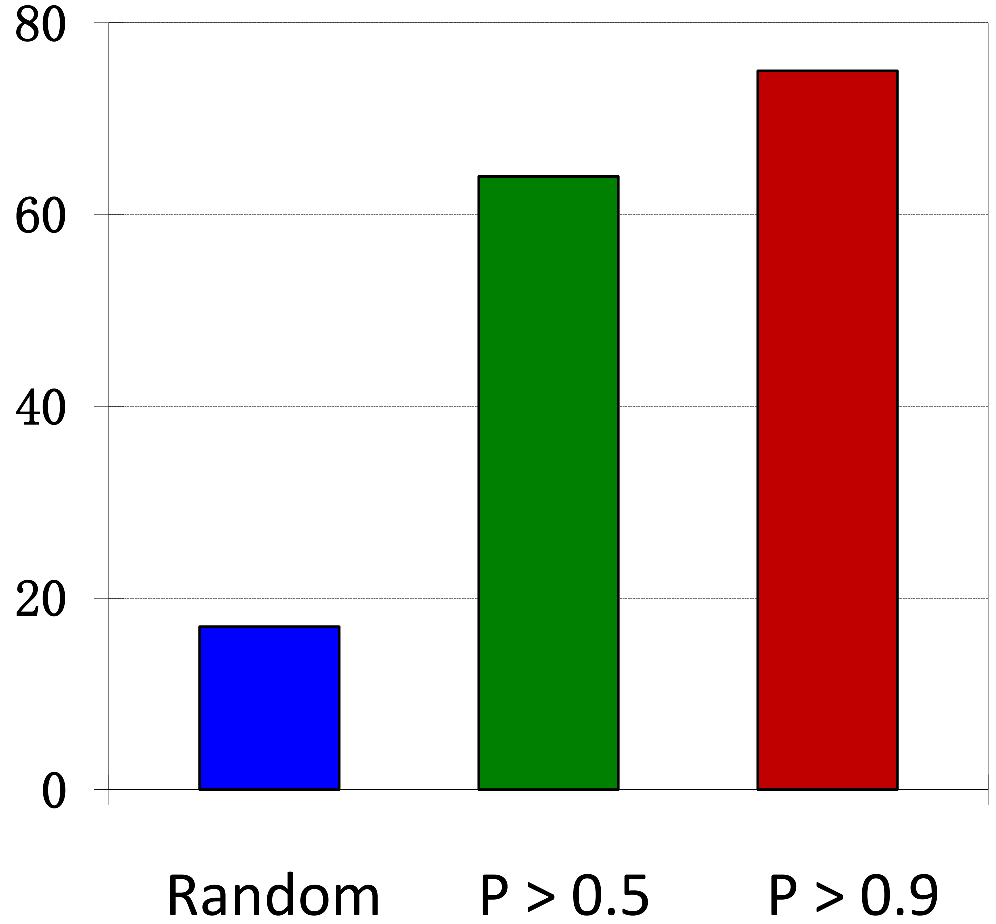
Numbers of *distinct* drugs, genes and mutations and their interactions in the knowledge bases vs. PubMed scale automatic extraction.

Machine reading extracted orders of magnitudes more knowledge

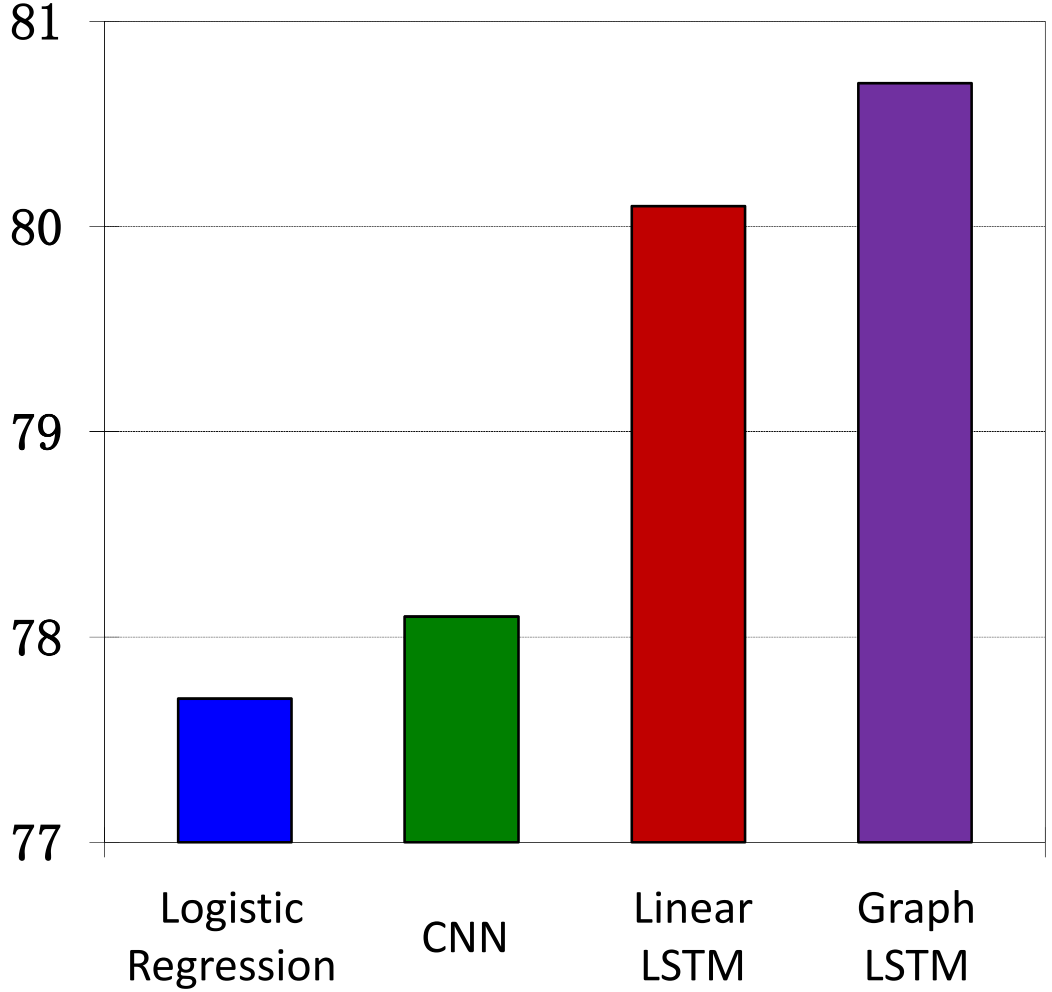
Cross-sentence extraction triples the yield

Sample Precision

Precision



Automatic Evaluation



Multi-Task Learning

Code and data available at: http://hanover.azurewebsites.net/

	Drug-Gene-Mutation	Drug-Mutation
Graph LSTM	80.7	76.7
+ Multi-task	82.0	78.5

Conclusion

- Jointly learning comprehensive representations from *heterogeneous sources:*
  - *Data and code available at: https://github.com/hltcoe/golden-horse/*
- Encoding linguistic structures to learn robust representations:
  - Data and code available at: *http://hanover.azurewebsites.net/*

Knowledge Graph Reasoning: Past, Present, and Future

William Wang

Department of Computer Science

UC SANTA BARBARA

NAACL 2018 Tutorial

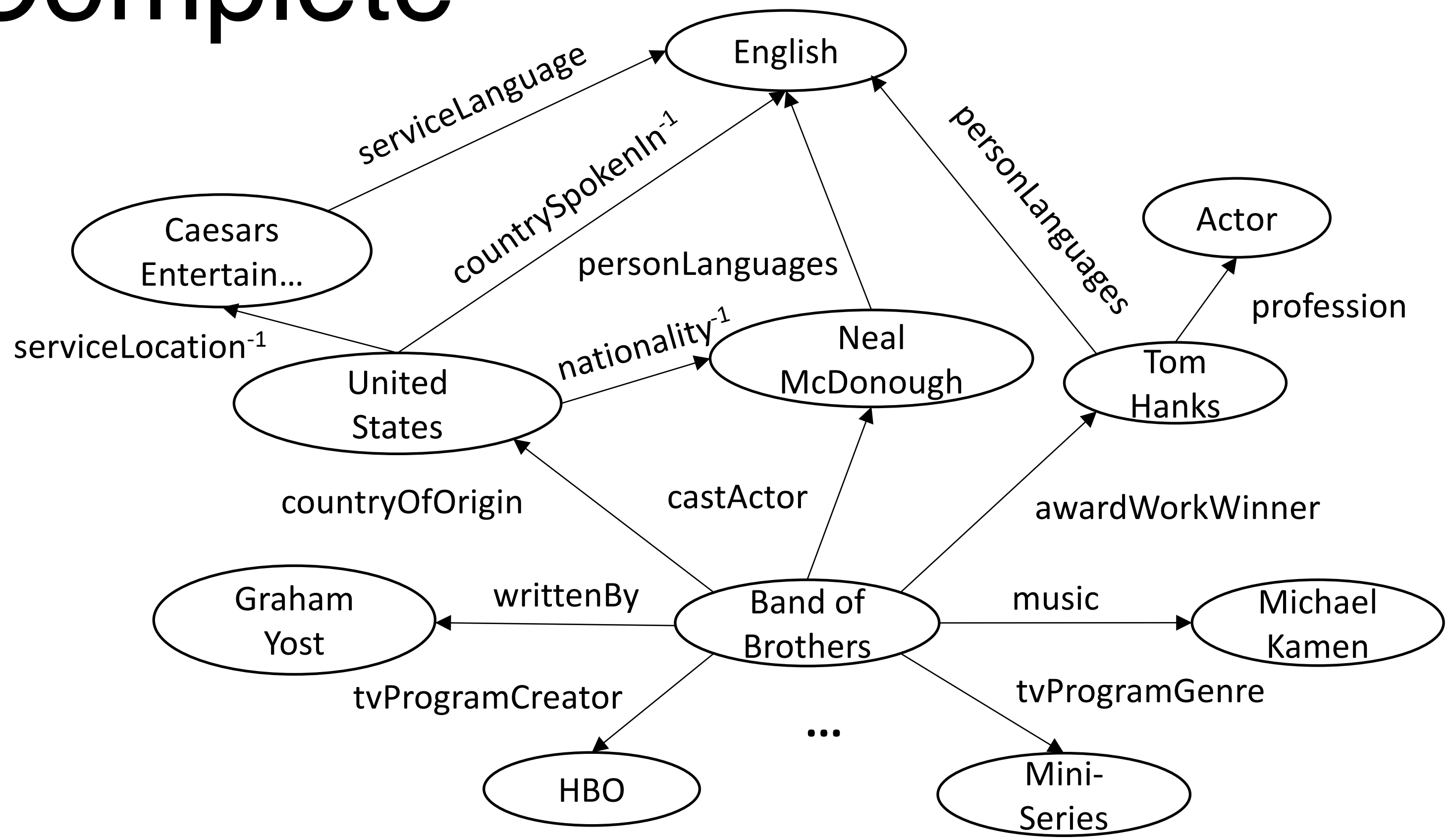
w. Xiang Ren and Nanyun Peng (USC)

UCSB

Agenda

- Motivation
- Path-Based Reasoning
- Embedding-Based Reasoning
- Bridging Path-Based and Embedding-Based Reasoning: DeepPath, MINERVA, and DIVA
- Conclusions

Knowledge Graphs are Not Complete



UCSB

Benefits of Knowledge Graph

- Support various applications
  - Structured Search
  - Question Answering
  - Dialogue Systems
  - Relation Extraction
  - Summarization
- Knowledge Graphs can be constructed via information extraction from text, but…
  - There will be a lot of missing links.
  - Goal: complete the knowledge graph.

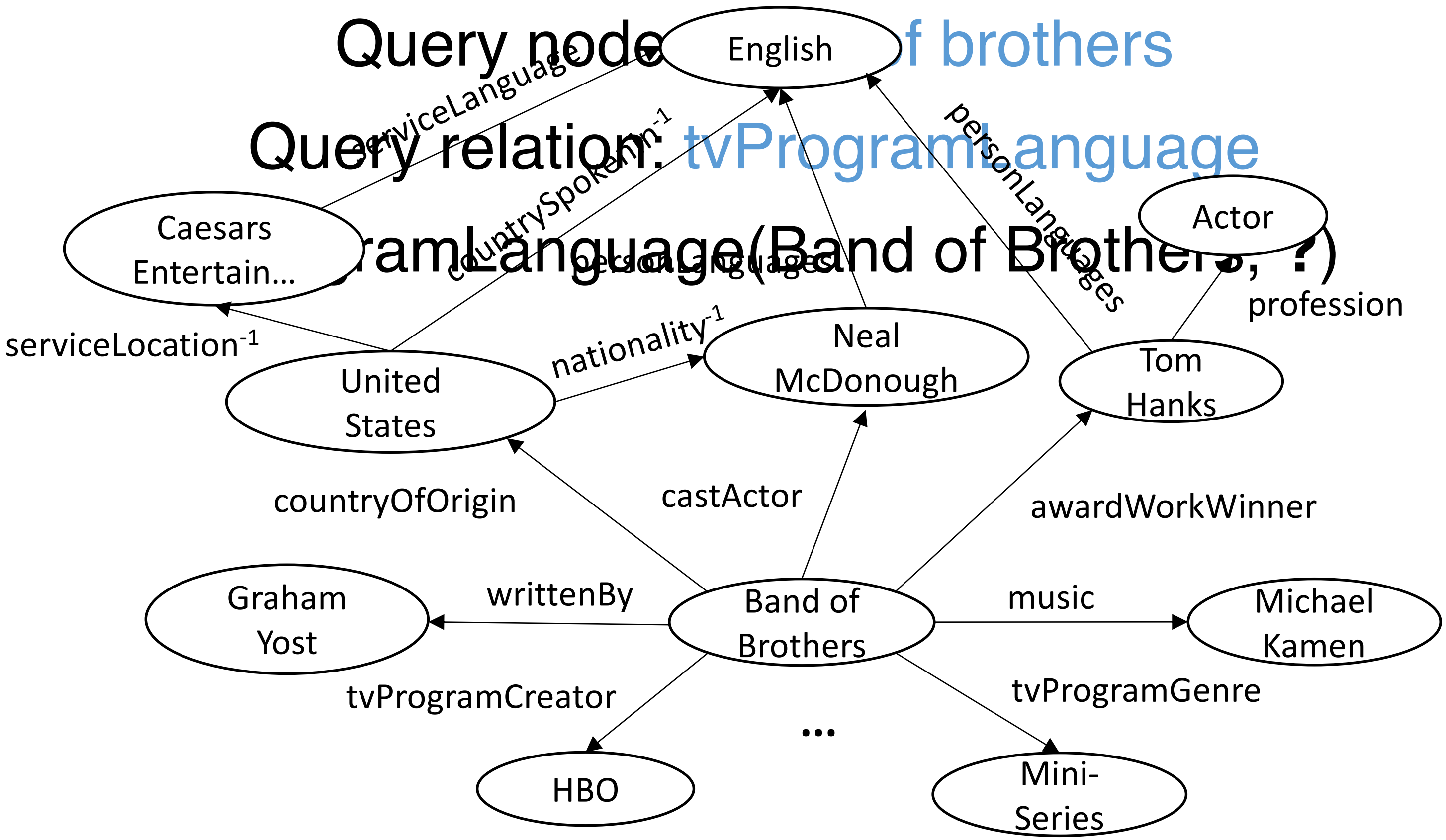
UCSB

Reasoning on Knowledge Graph

Query node f brothers

Query relation: tvProgramLanguage

ramLanguage(Band of Brothers, ?)



KB Reasoning Tasks

- Predicting the missing link.
  - Given e1 and e2, predict the relation r.
- Predicting the missing entity.
  - Given e1 and relation r, predict the missing entity e2.
- Fact Prediction.
  - Given a triple, predict whether it is true or false.

Related Work

- **Path-based methods**
  - Path-Ranking Algorithm, Lao et al. 2011
  - ProPPR, Wang et al, 2013 (My PhD thesis)
  - Subgraph Feature Extraction, Gardner et al, 2015
  - RNN + PRA, Neelakantan et al, 2015
  - Chains of Reasoning, Das et al, 2017

Why do we need path-based methods?

It's accurate and explainable!

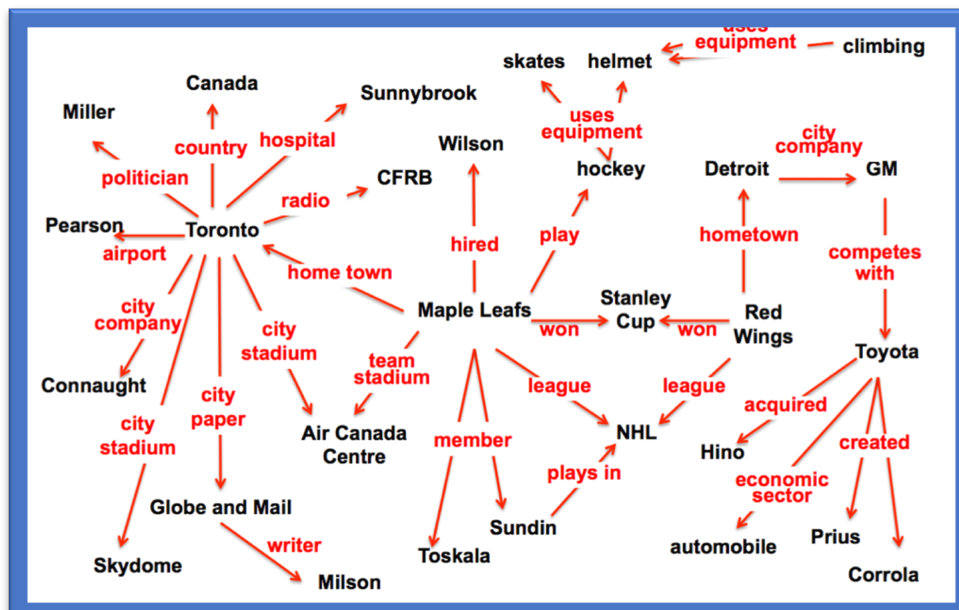
UCSB

Path-Ranking Algorithm (Lao et al., 2011)

- 1. Run random walk with restarts to derive many paths.
- 2. ent paths.

teamHomeStadium

$$c \xrightarrow{\text{teamPlaysInCity}} c \xrightarrow{\text{cityStadiums}} c$$



UCSB

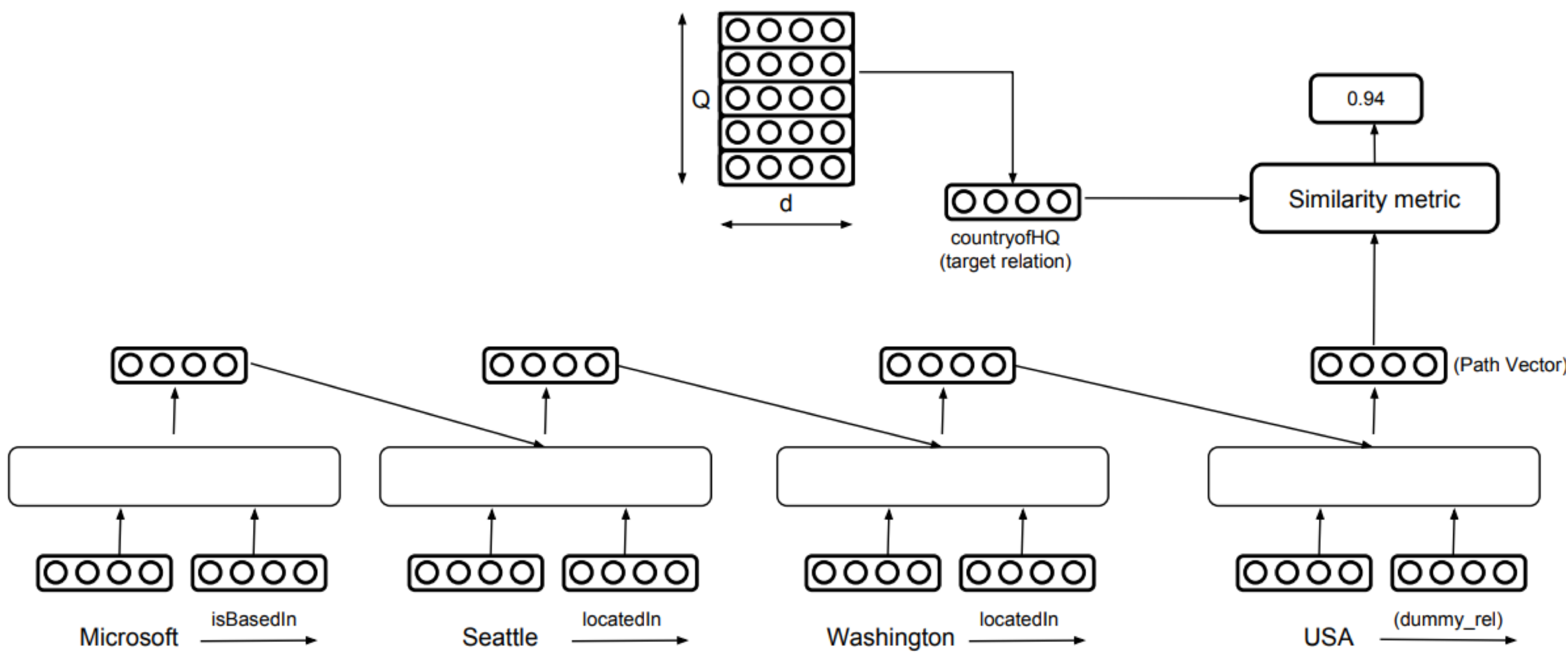
ProPPR (Wang et al., 2013;2015)

- ProPPR generalizes PRA with recursive probabilistic logic programs.
- You may use other relations to jointly infer this target relation.

```
about(X,Z):- handLabeled(X,Z)                         # base
about(X,Z):- sim(X,Y),about(Y,Z)                      # prop
sim(X,Y):- link(X,Y)                                  # sim,link
sim(X,Y):-
    hasWord(X,W),hasWord(Y,W),
    linkedBy(X,Y,W)                                   # sim,word
linkedBy(X,Y,W):- true                                # by(W)
```

Chain of Reasoning (Das et al, 2017)

- 1. Use PRA to derive the path.
- 2. Use RNNs to perform reasoning of the target relation.



UCSB

Related Work

- **Embedding-based method**
  - RESCAL, Nickel et al, 2011
  - TransE, Bordes et al, 2013
  - Neural Tensor Network, Socher et al, 2013
  - TransR/CTransR, Lin et al, 2015
  - Complex Embeddings, Trouillon et al, 2016

Embedding methods allow us to compare, and find similar entities in the vector space.

UCSB

RESCAL (Nickel et al., 2011)

- Tensor factorization on the
  - (head)entity-(tail)entity-relation tensor.

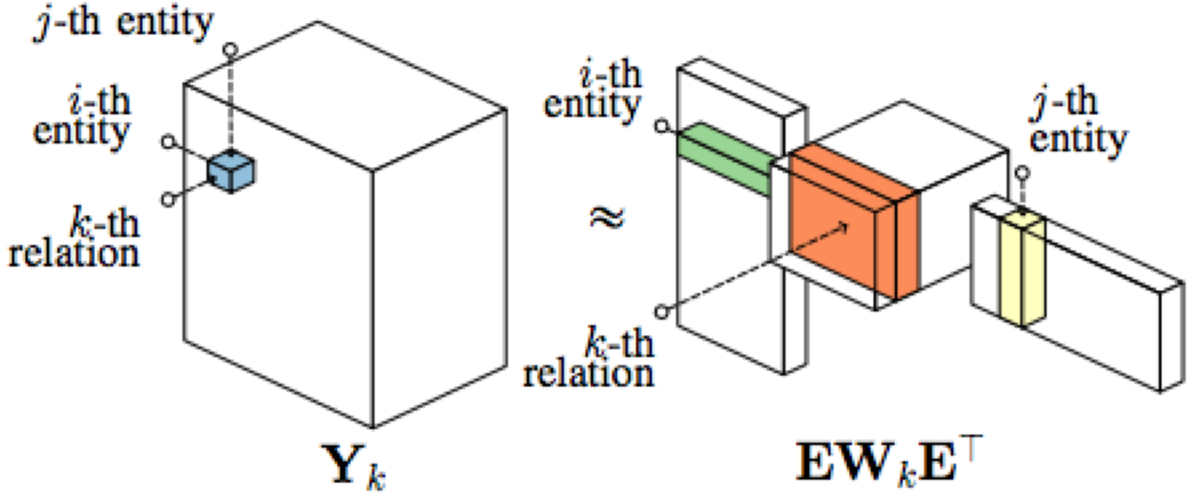


Fig. 4. RESCAL as a tensor factorization of the adjacency tensor $\mathbf{Y}$.

UCSB

TransE (Bordes et al., 2013)

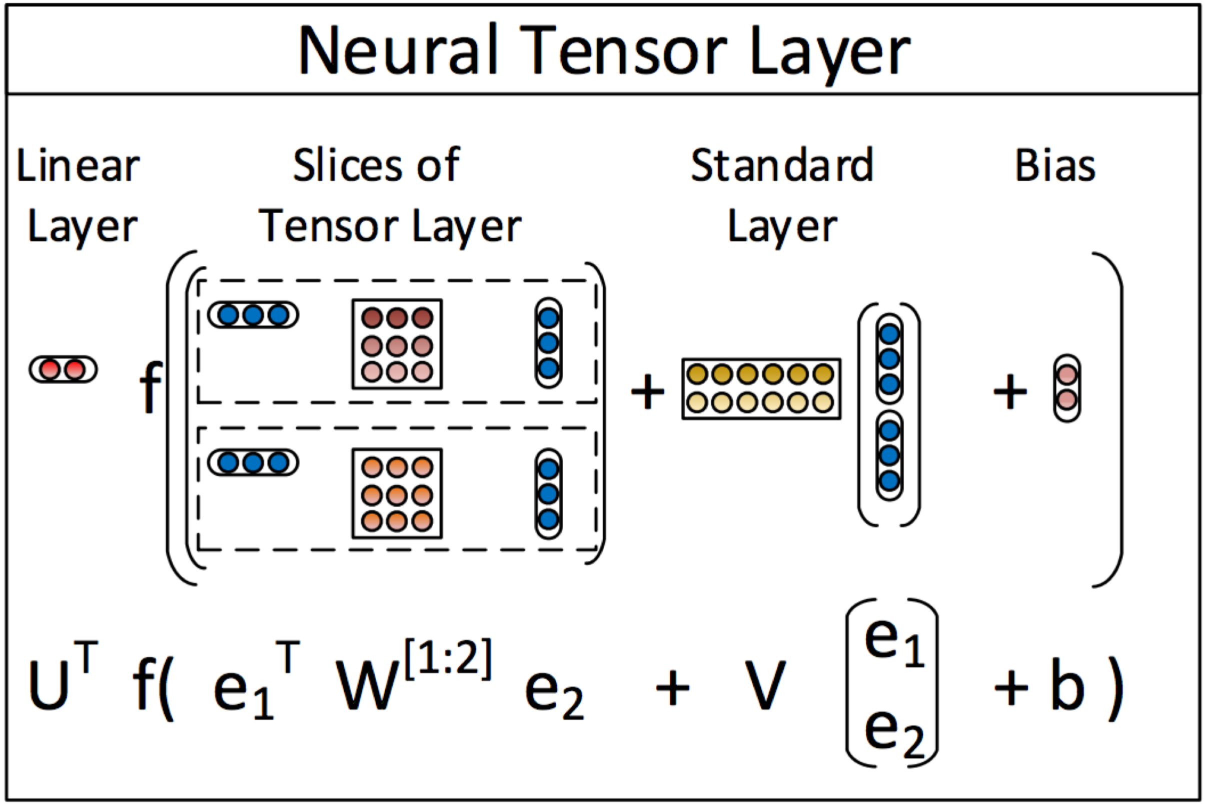
- Assumption: in the vector space, when adding the relation to the head entity, we should get close to the target tail entity.
- Margin based loss function:
  - Minimize the distance between (h+l) and t.
  - Maximize the distance between (h+l) to a randomly sampled tail t' (negative example).

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

UCSB

Neural Tensor Networks (Socher et al., 2013)

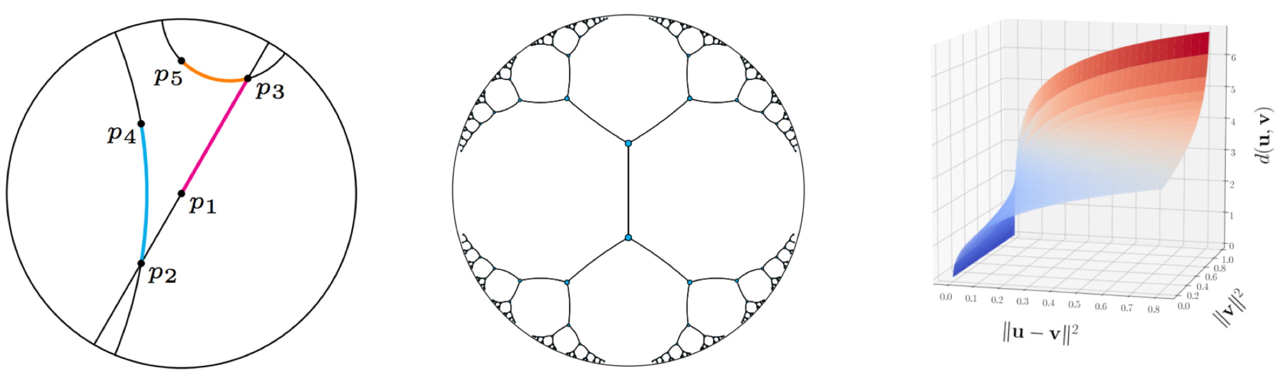
- Model the bilinear interaction between entity pairs with tensors.



Poincaré Embeddings (Nickel and Kiela, 2017)

- Idea: learn hierarchical KB representations by looking at hyperbolic space.

$$d(\boldsymbol{u}, \boldsymbol{v}) = \operatorname{arcosh}\left(1 + 2\frac{\|\boldsymbol{u} - \boldsymbol{v}\|^2}{(1 - \|\boldsymbol{u}\|^2)(1 - \|\boldsymbol{v}\|^2)}\right).$$



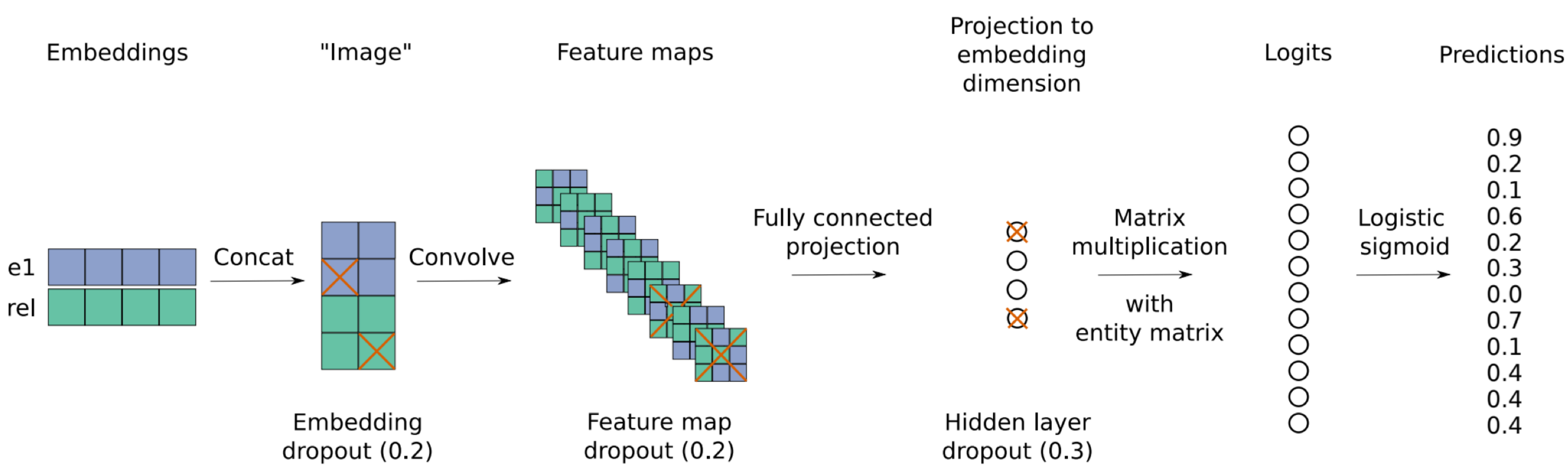
(a) Geodesics of the Poincaré disk (b) Embedding of a tree in $\mathcal{B}^2$ (c) Growth of Poincaré distance

Figure 1: (a) Due to the negative curvature of $\mathcal{B}$, the distance of points increases exponentially (relative to their Euclidean distance) the closer they are to the boundary. (c) Growth of the Poincaré distance $d(\boldsymbol{u}, \boldsymbol{v})$ relative to the Euclidean distance and the norm of $\boldsymbol{v}$ (for fixed $\|\boldsymbol{u}\| = 0.9$). (b) Embedding of a regular tree in $\mathcal{B}^2$ such that all connected nodes are spaced equally far apart (i.e., all black line segments have identical hyperbolic length).

UCSB

ConvE (Dettmers et al, 2018)

- 1. Reshape the head and relation embeddings into "images".
- 2. Use CNNs to learn convolutional feature maps.



UCSB

Bridging Path-Based and Embedding-Based Reasoning with Deep Reinforcement Learning: DeepPath (Xiong et al., 2017)

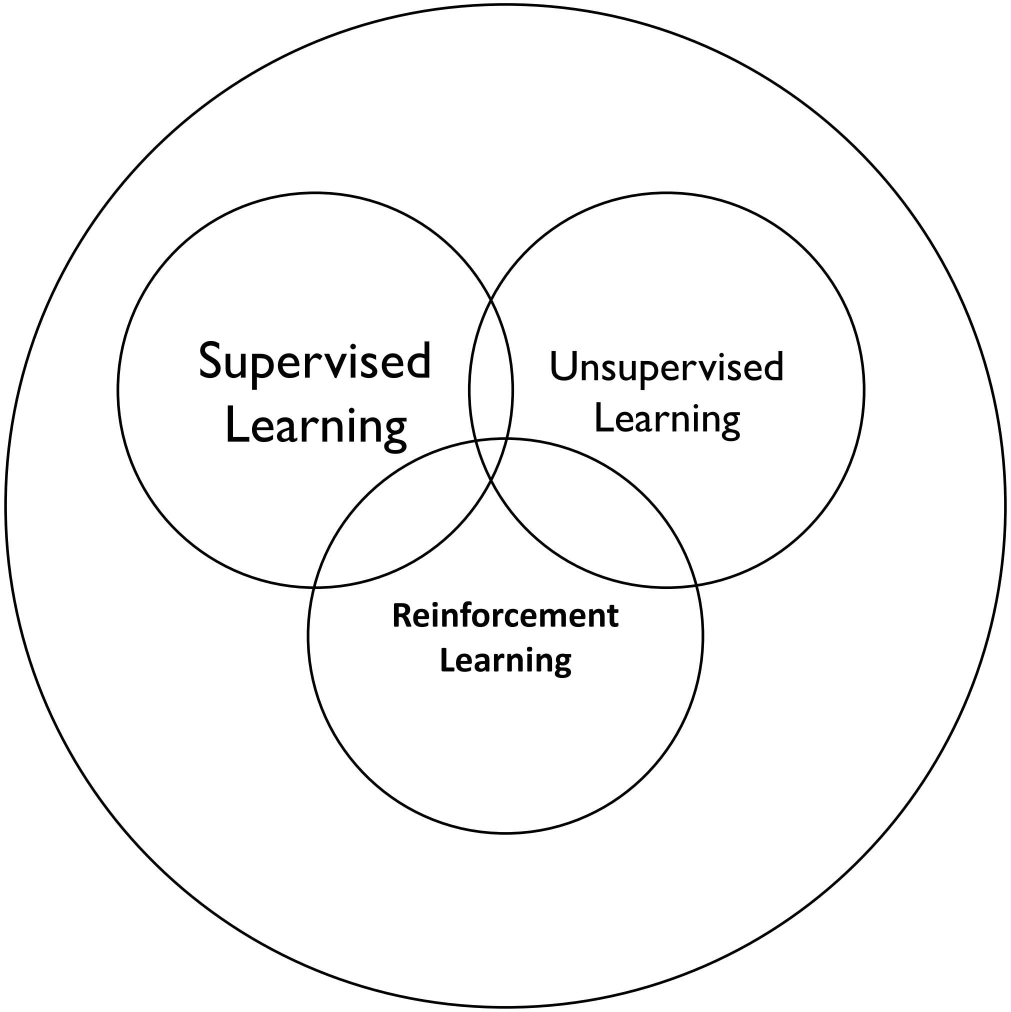
UCSB

RL for KB Reasoning: DeepPath (Xiong et al., 2017)

- Learning the paths with RL, instead of using random walks with restart
- Model the path finding as a MDP
- Train a RL agent to find paths
- Represent the KG with pretrained KG embeddings
- Use the learned paths as logical formulas

UCSB

Machine Learning



Supervised v.s. Reinforcement

Supervised Learning
- Training based on supervisor/label/annotation
- Feedback is instantaneous
- Not much temporal aspects

Reinforcement Learning
- Training only based on reward signal
- Feedback is delayed
- Time matters
- Agent actions affect subsequent exploration

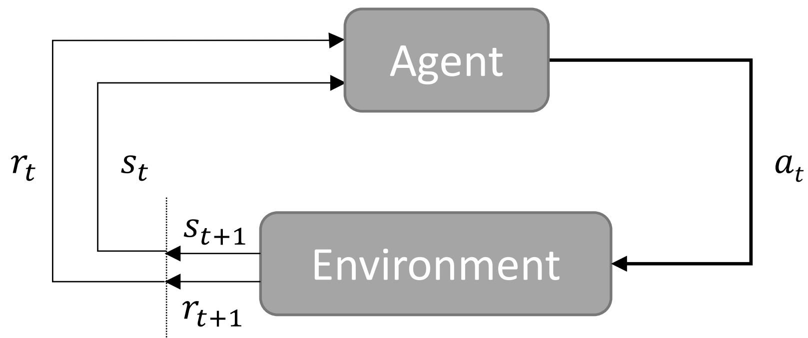
UCSB

Reinforcement Learning

- RL is a general purpose framework for **decision making**
  - RL is for an *agent* with the capacity to *act*
  - Each *action* influences the agent's future *state*
  - Success is measured by a scalar *reward* signal
  - Goal: *select actions to maximize future reward*

UCSB

Reinforcement Learning

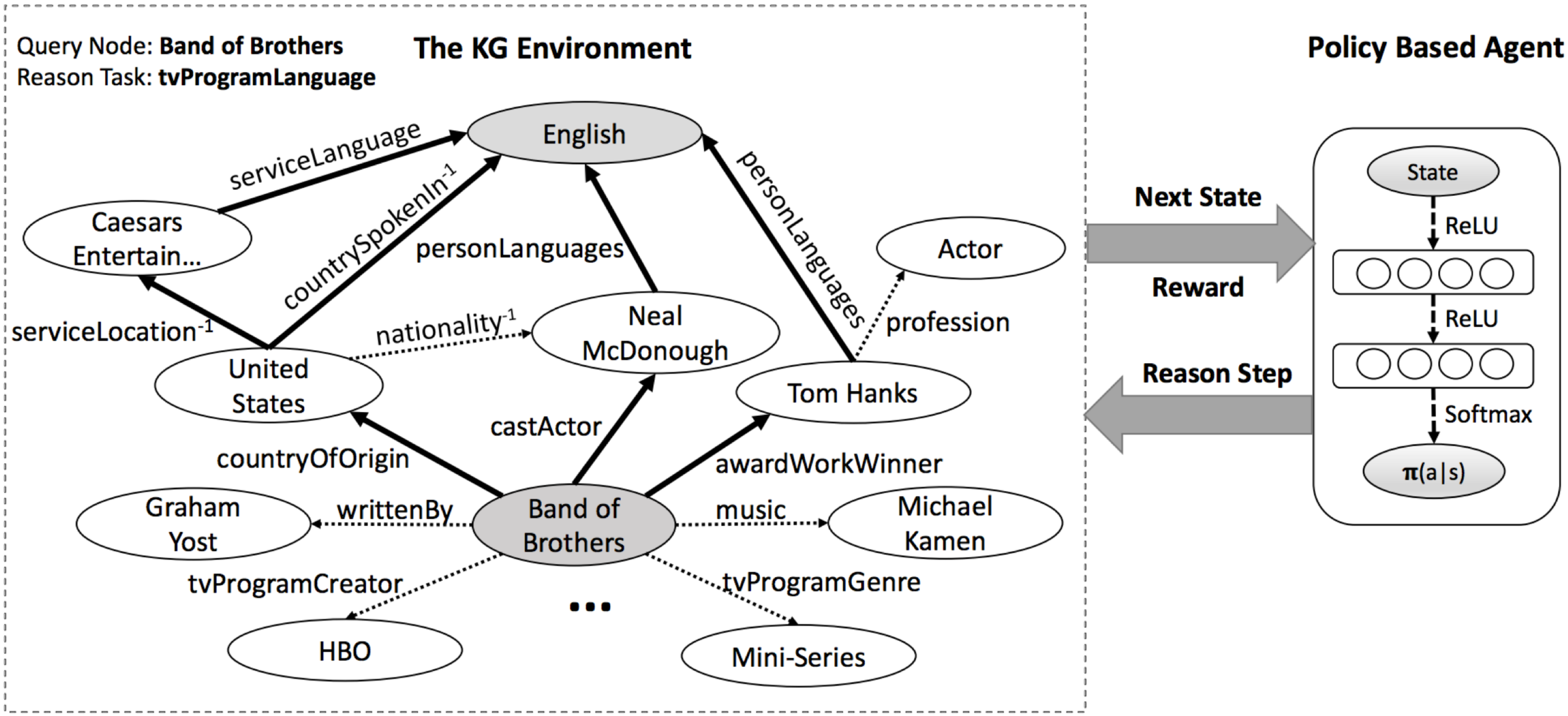


Multi-layer neural nets $\psi(s_t)$

KG modeled as a MDP

UCSB

DeepPath: RL for KG Reasoning



UCSB

Components of MDP

- Markov decision process $< S, A, P, R >$
  - S: continuous states represented with embeddings
  - A: action space (relations)
  - $P(S_{t+1} = s' | S_t = s, A_t = a)$: transition probability
  - $R(s, a)$: reward received for each taken step

- With pretrained KG embeddings
  - $s_t = e_t \oplus (e_{target} - e_t)$
  - $A = \{r_1, r_2, \dots, r_n\}$, all relations in the KG

UCSB

Reward Functions

- Global Accuracy

$$r_{\text{GLOBAL}} = \begin{cases} +1, & \text{if the path reaches } e_{target} \\ -1, & \text{otherwise} \end{cases}$$

- Path Efficiency

$$r_{\text{EFFICIENCY}} = \frac{1}{length(p)}$$

- Path Diversity

$$r_{\text{DIVERSITY}} = -\frac{1}{|F|}\sum_{i=1}^{|F|} cos(\mathbf{p}, \mathbf{p}_i)$$

UCSB

Training with Policy Gradient

- Monte-Carlo Policy Gradient (REINFORCE, William, 1992)

$$\nabla_\theta J(\theta) = \sum_t \sum_{a \in \mathcal{A}} \pi(a|s_t;\theta) \nabla_\theta \log \pi(a|s_t;\theta) R(s_t, a_t)$$
$$\approx \nabla_\theta \sum_t \log \pi(a = r_t|s_t;\theta) R(s_t, a_t)$$

$$R(s_t, a_t) = \lambda_1 r_{global} + \lambda_2 r_{efficiency} + \lambda_3 r_{diversity}$$

Challenge

- Typical RL problems
  - Atari games (Mnih et al., 2015): 4~18 valid actions
  - AlphaGo (Silver et al. 2016): ~250 valid actions
  - Knowledge Graph reasoning: >= 400 actions

Issue:

- large action (search) space -> poor convergence properties

UCSB

Supervised (Imitation) Policy Learning

- Use randomized BFS to retrieve a few paths
- Do imitation learning using the retrieved paths
- All the paths are assigned with +1 reward

$$\nabla_\theta J(\theta) = \sum_t \sum_{a \in \mathcal{A}} \pi(a|s_t;\theta) \nabla_\theta \log \pi(a|s_t;\theta)$$
$$\approx \nabla_\theta \sum_t \log \pi(a = r_t|s_t;\theta)$$

UCSB

Datasets and Preprocessing

Dataset	# of Entities	# of Relations	# of Triples	# of Tasks
FB15k-237	14,505	237	310,116	20
NELL-995	75,492	200	154,213	12

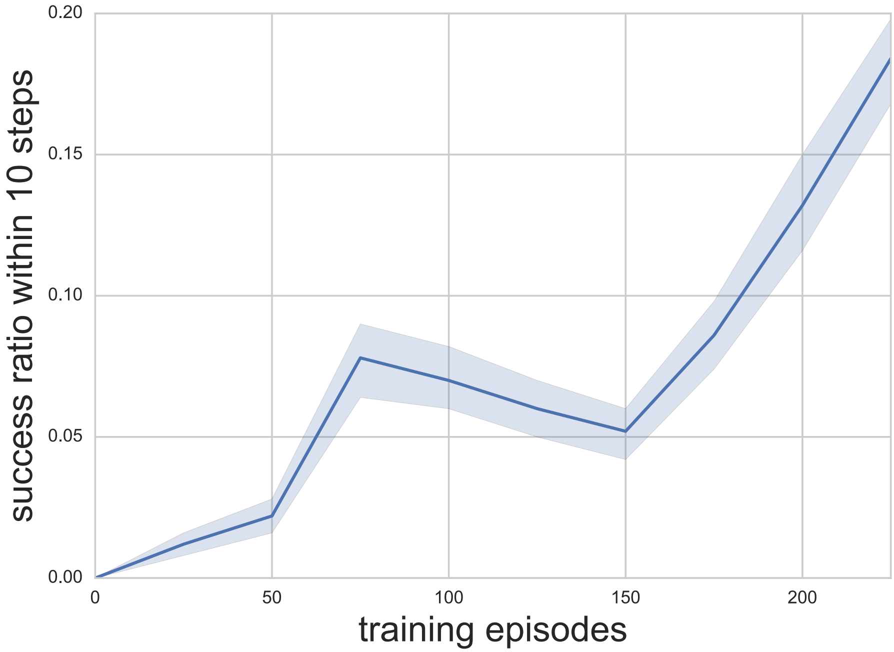
FB15k-237: Sampled from FB15k (Bordes et al., 2013), redundant relations removes

NELL-995: Sampled from the 995th iteration of NELL system (Carlson et al., 2010b)

- **Dataset processing**
  - Remove useless relations: *haswikipediaurl*, *generalizations, etc*
  - Add inverse relation links to the knowledge graph
  - Remove the triples with task relations

UCSB

Effect of Supervised Policy Learning



- **x-axis:** number of training epochs
- **y-axis:** success ratio (probability of reaching the target) on test set

-> Re-train the agent using reward functions

UCSB

Inference Using Learned Paths

- Path as logical formula
  - FilmCountry: actionFilm^{-1} -> personNationality
  - PersonNationality: placeOfBirth -> locationContains^{-1}
  - etc …
- Bi-directional path-constrained search
  - Check whether the formulas hold for entity pairs

Uni-directional search **bi-directional search**

Link Prediction Result

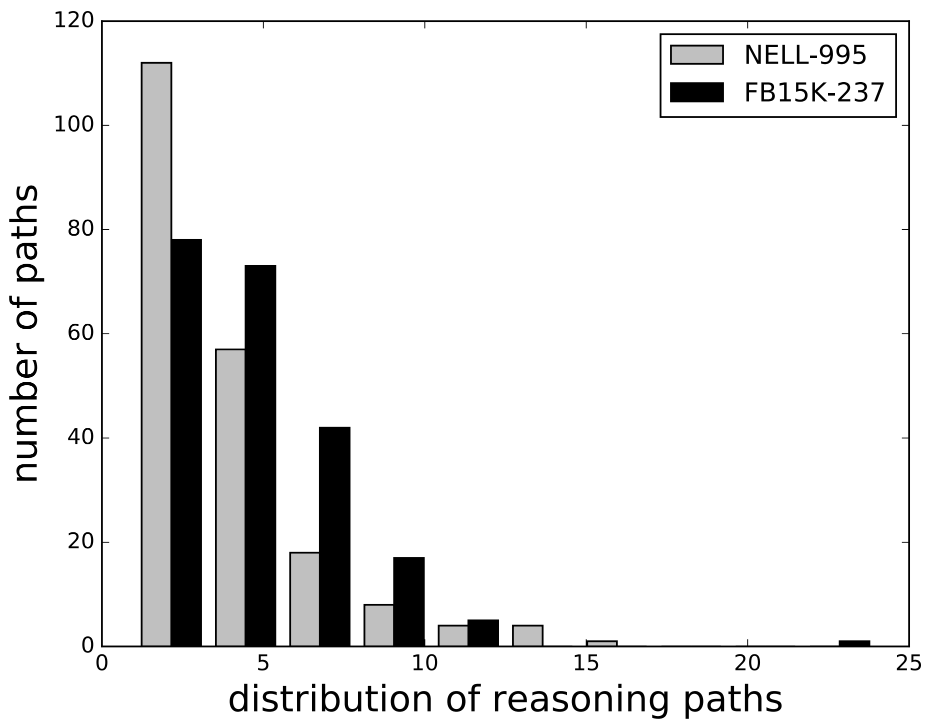
Tasks	PRA	Ours	TransE	TransR
worksFor	0.681	**0.711**	0.677	0.692
atheletPlaysForTeam	**0.987**	0.955	0.896	0.784
athletePlaysInLeague	0.841	**0.960**	0.773	0.912
athleteHomeStadium	0.859	**0.890**	0.718	0.722
teamPlaysSports	0.791	0.738	0.761	**0.814**
orgHirePerson	0.599	**0.742**	0.719	0.737
personLeadsOrg	0.700	**0.795**	0.751	0.772
…				
Overall	0.675	**0.796**	0.737	0.789

Mean average precision on NELL-995

UCSB

Qualitative Analysis

Path length distributions



UCSB

Qualitative Analysis

Example Paths

personNationality:
- placeOfBirth -> $\text{locationContains}^{-1}$
- peoplePlaceLived -> $\text{locationContains}^{-1}$
- peopleMariage -> locationOfCeremony -> $\text{locationContains}^{-1}$

tvProgramLanguage:
- tvCountryOfOrigin -> countryOfficialLanguage
- tvCountryOfOrigin -> filmReleaseRegion-1 -> filmLanguage
- tvCastActor -> personLanguage

athletePlaysForTeam:
- athleteHomeStadium -> $\text{teamHomeStadium}^{-1}$
- athletePlaysSports -> $\text{teamPlaysSports}^{-1}$
- atheleteLedSportsTeam

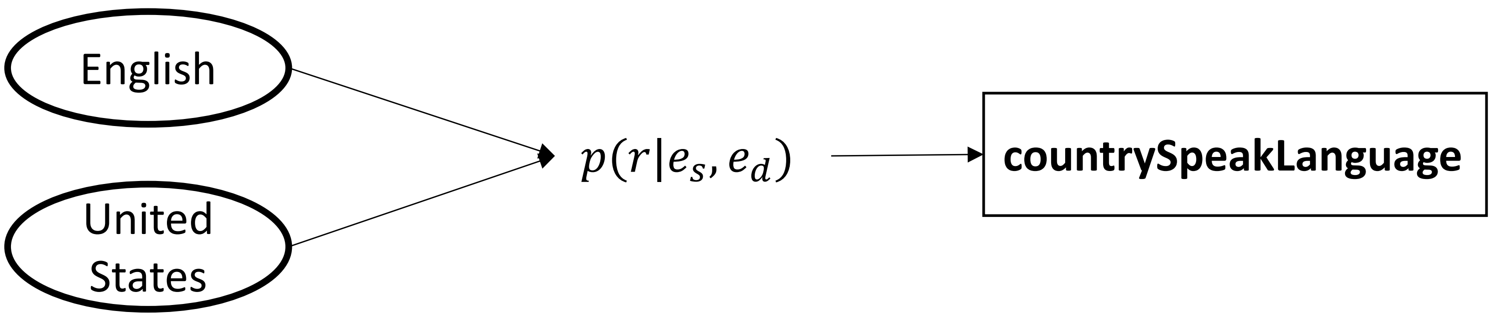
UCSB

Bridging Path-Finding and Reasoning w. Variational Inference (teaser): DIVA (Chen et al., NAACL 2018)

UCSB

DIVA: Variational KB Reasoning (NAACL 2018, Monday Morning)

- Inferring latent paths connecting entity nodes.



Condition (e_s, e_d)

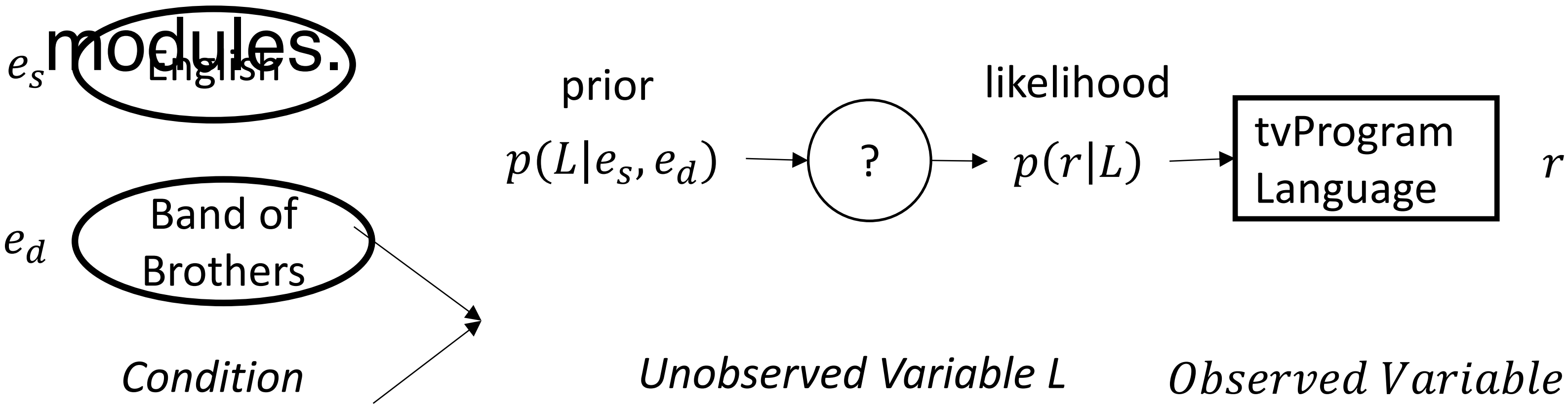
Observed Variable r

$$\bar{p} = argmax_p \log p(r|e_s, e_d)$$

UCSB

DIVA: Variational KB Reasoning (NAACL 2018, Monday Morning)

- Inferring latent paths connecting entity nodes by parameterizing likelihood (path reasoning) and prior (path finding) with neural network modules.

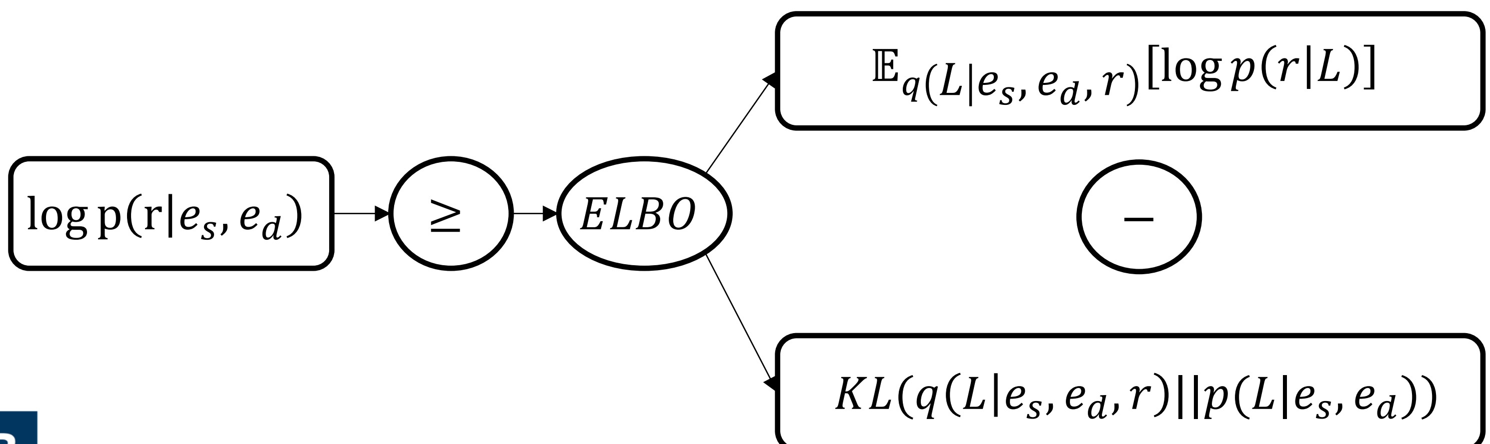


$$p = argmax_p\, p(r|e_s, e_d) = argmax_p \log \int_L^{\infty} p(r|L)p(L|e_s, e_d)$$

UCSB

DIVA: Variational KB Reasoning (NAACL 2018, Monday Morning)

- Marginal likelihood $\log \int_L p(r|L)p(L|e_s, e_d)$is intractable
- We resort to Variational Bayes by introduce a posterior distribution $q(L|e_s, e_d, r)$



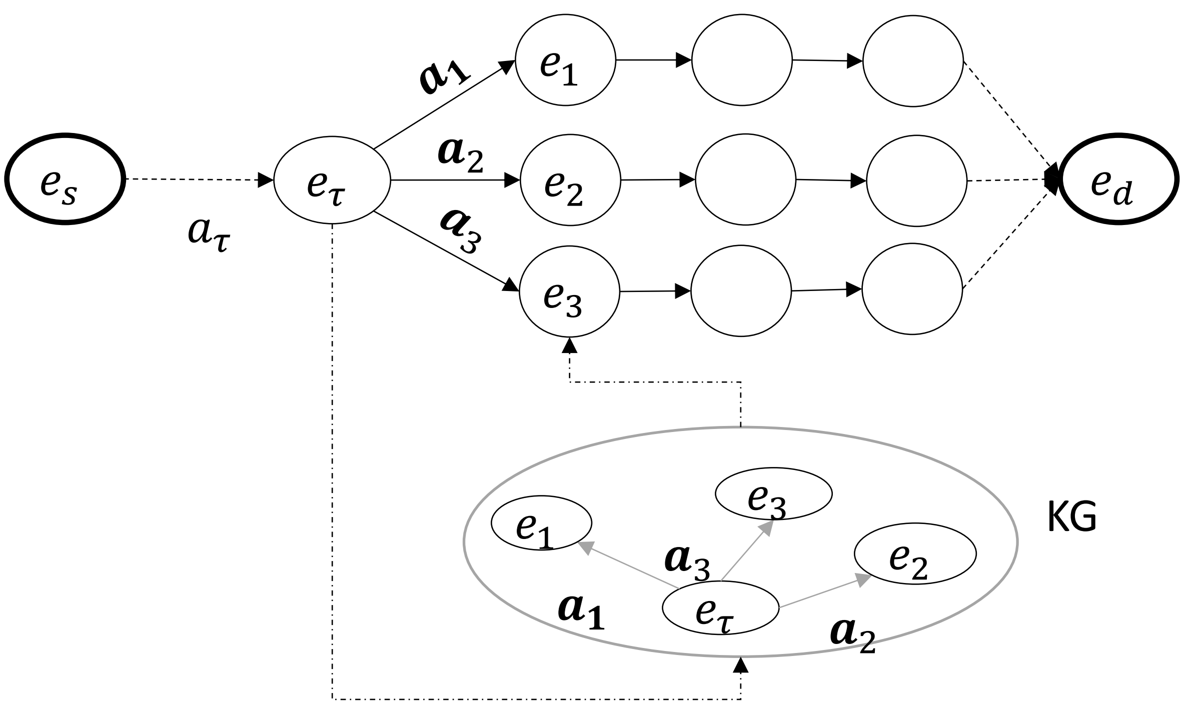
UCSB

DP Kingma et al. 2013

Parameterization – Path-finder

- Approximate posterior $q_\varphi(L|e_s, e_d, r)$ and prior $p_\beta\ (L|e_s, e_d)$: parameterize with RNN

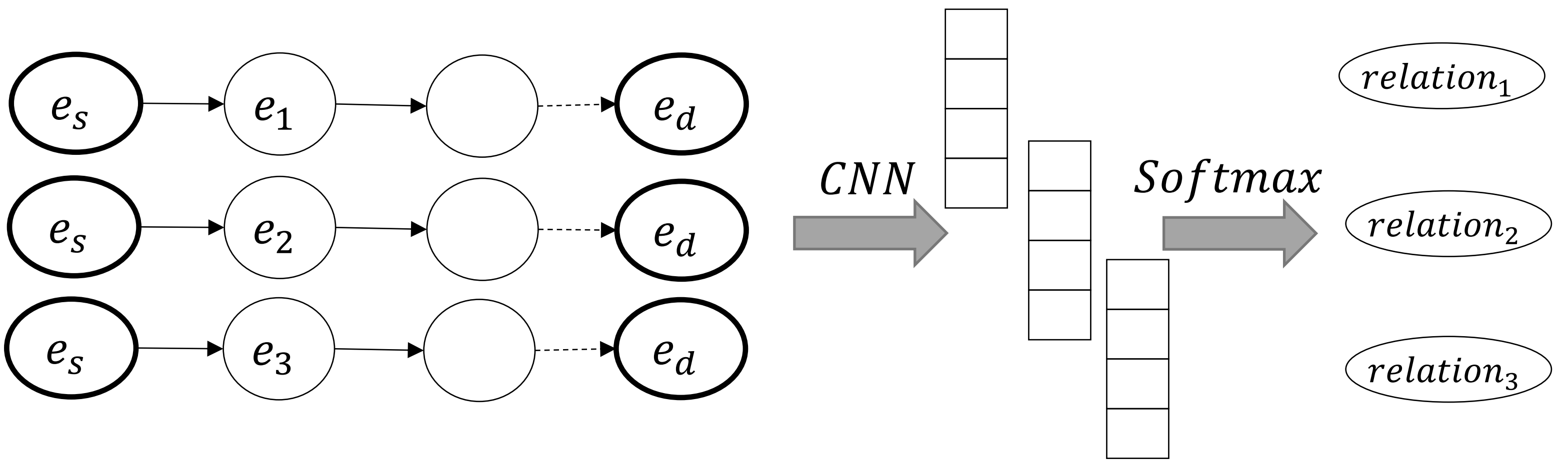
Transition Probability: $p(a_{\tau+1}, e_{\tau+1}|a_{1:\tau}, e_{1:\tau})$



UCSB

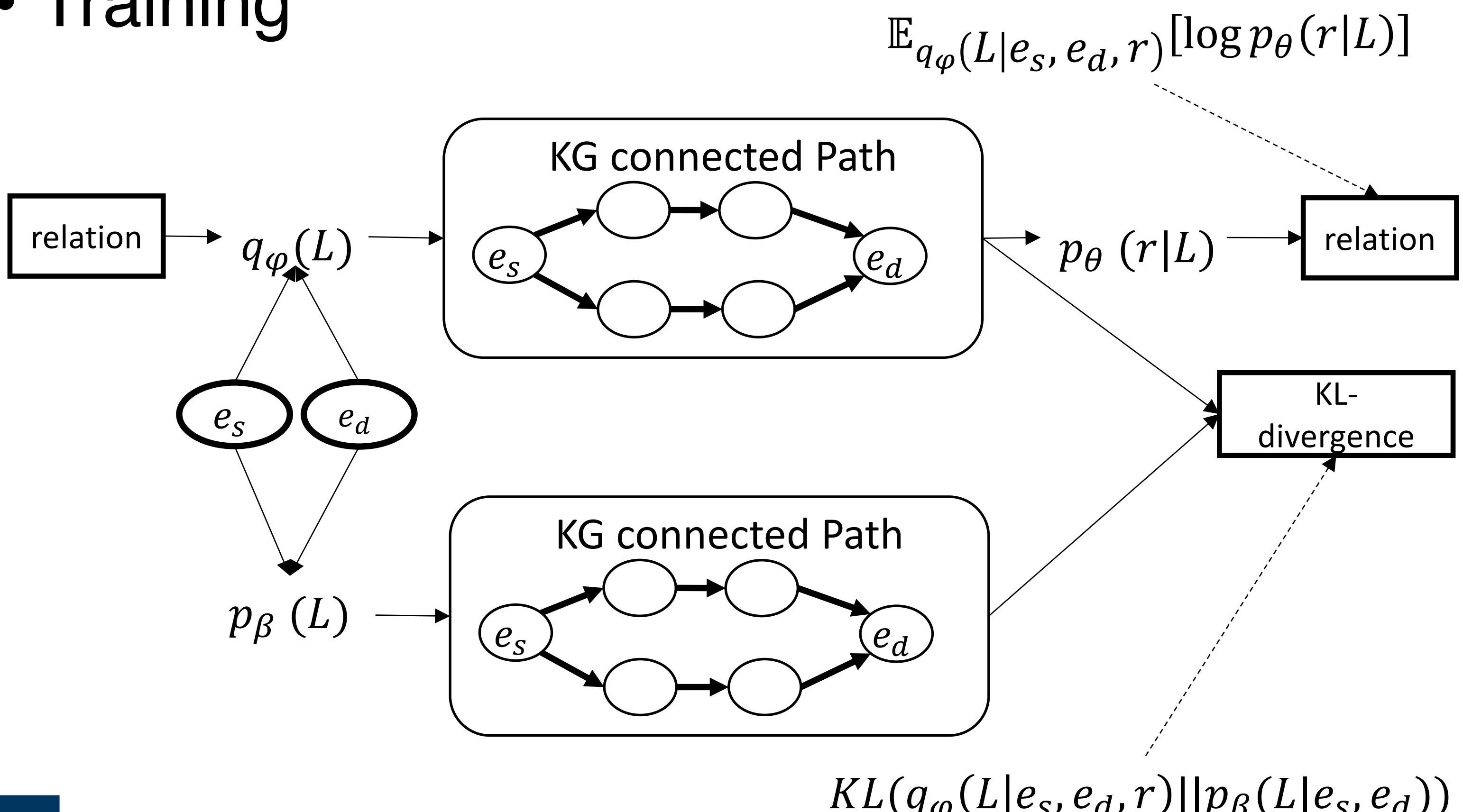
Parameterization – Path Reasoner

- Likelihood $p_\theta\ (r|L)$: parameterize with CNN



DIVA: Variational KB Reasoning (NAACL 2018, Monday Morning)

- Training

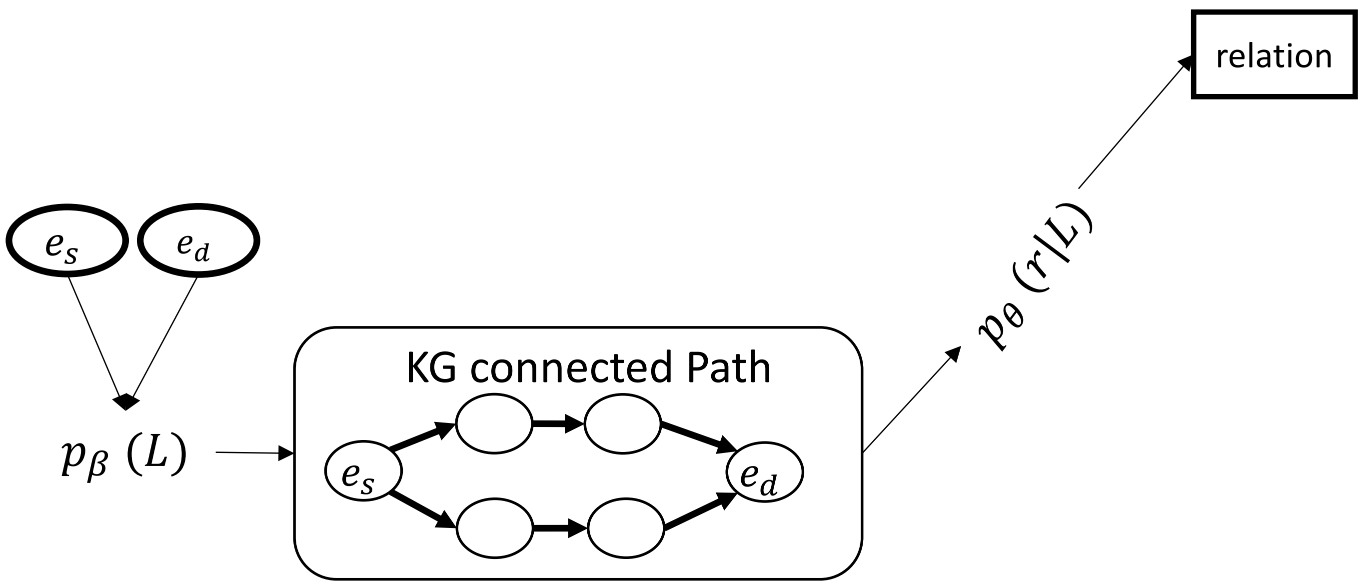


UCSB

$posterior$: q_φ, likelihood: $p_\theta\ (r|L)$, prior: $p_\beta(L|e_s,e_d)$

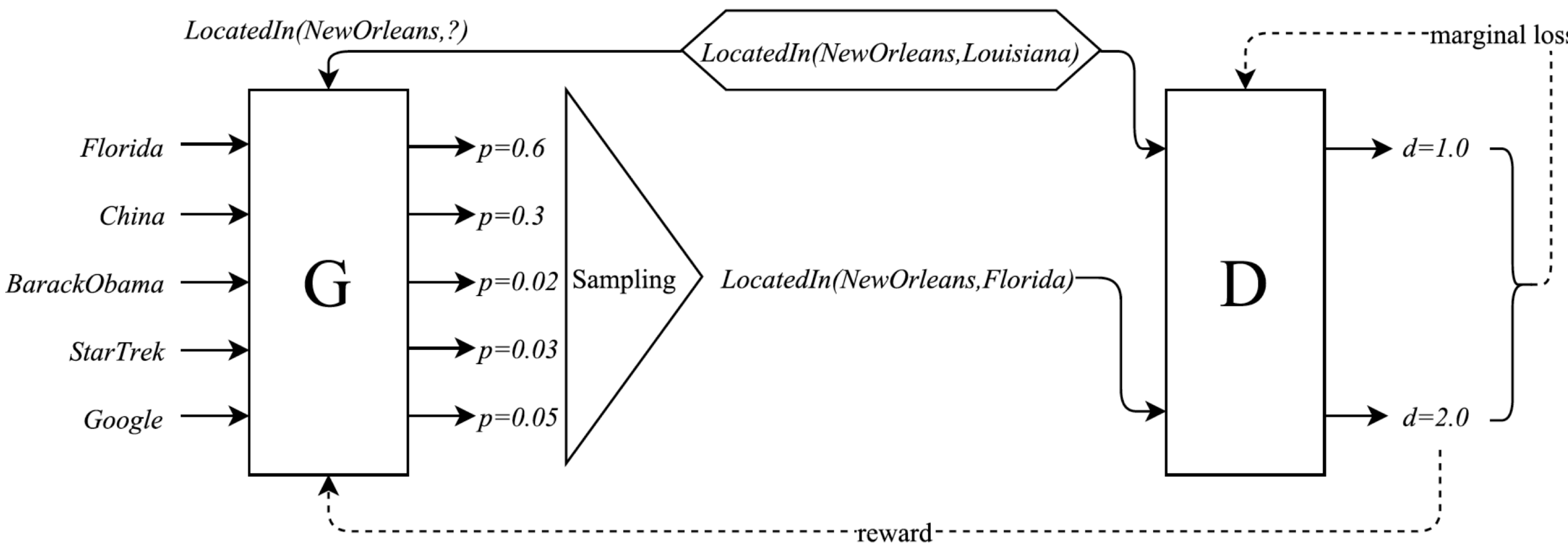
DIVA: Variational KB Reasoning (NAACL 2018, Monday Morning)

- Testing



$posterior$: q_φ, likelihood: $p_\theta\ (r|L)$, prior: $p_\beta(L|e_s, e_d)$

KBGAN: Adversarial Learning for Knowledge Graph Completion (NAACL 2018, Monday Morning)



Idea: use adversarial learning to replace random sampling (from a uniform distribution).

UCSB

Conclusions

- Embedding-based methods are very scalable and robust.
- Path-based methods are more interpretable.
- There are some recent efforts in unifying embedding and path-based approaches.
- DIVA integrates path-finding and reasoning in a principled variational inference framework.

UCSB

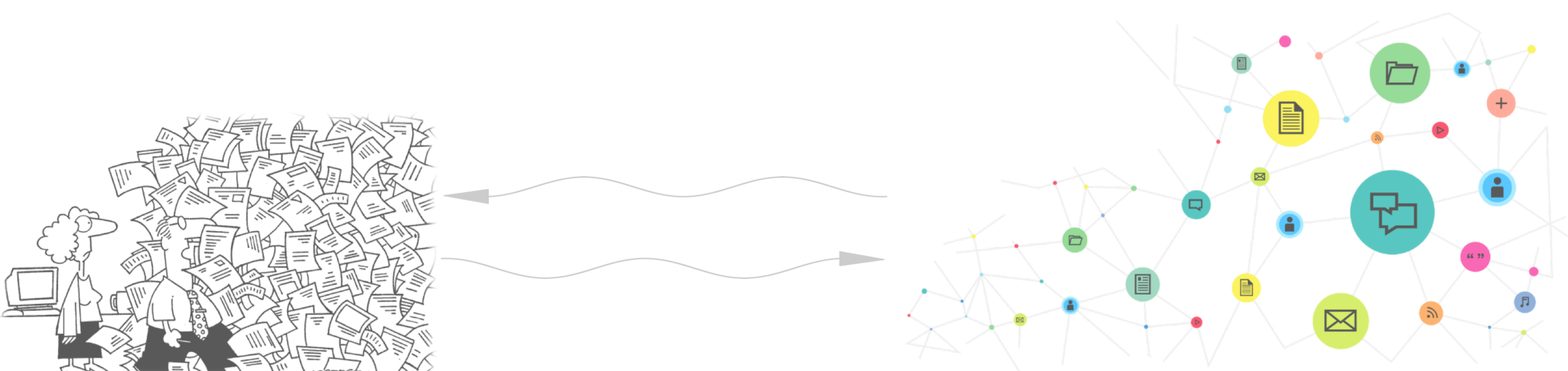
Thanks!

DeepPath Source code:
https://github.com/xwhan/DeepPath
KBGAN Source code:
https://github.com/cai-lw/KBGAN
ProPPR Source code:
https://github.com/TeamCohen/ProPPR

UCSB

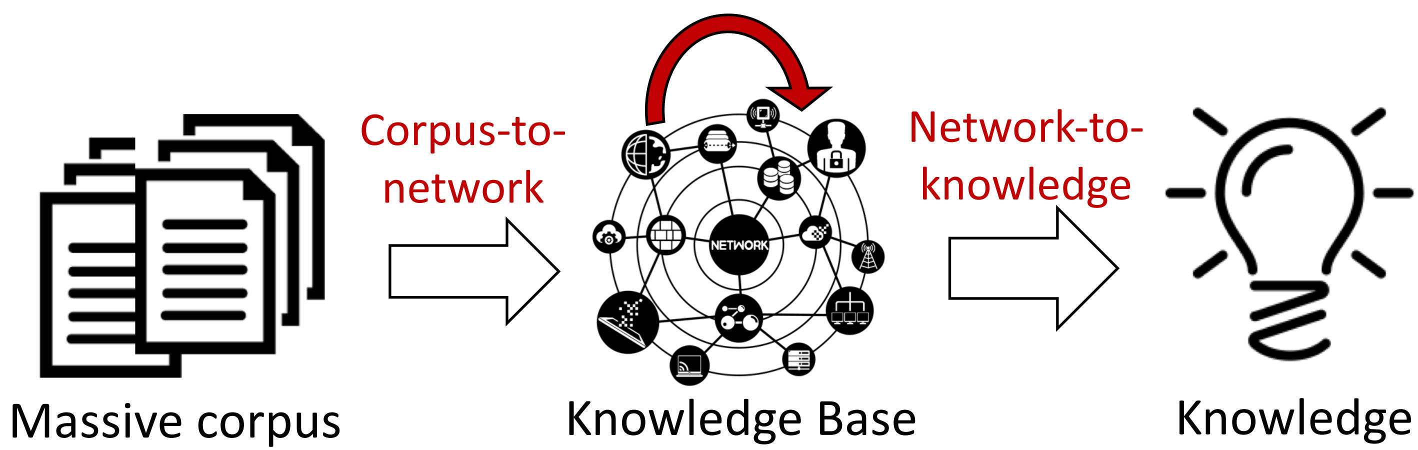
Scalable Construction and Reasoning of Massive Knowledge Bases

Summary



Overall Contributions

- **Effort-Light Structure Extraction**
  → Corpus-specific labeling free, domain/language-independent
- **Joint Models for Low-resource IE:** jointly learning representations from unlabeled data, linguistic structures, annotations from other tasks, domains, and languages. → Reusable knowledge
- **Reasoning**: learning to infer missing links from background knowledge.
- A principled approach to manage, explore, and analyze "Big Text Data"



Looking Forward: **What's Next?**

Text corpus

Data-driven text segmentation
(SIGMOD'15, WWW'16, ...)

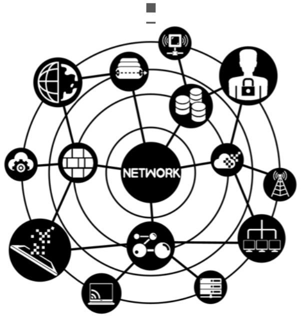
entity names & context units

Knowledge bases

Partially-labeled corpus

Learning Corpus-specific Model
(KDD'15, KDD'16, EMNLP'16, WWW'17, ...)

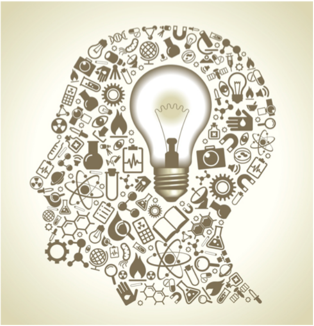
Structures from the remaining *unlabeled* data



Knowledge Base

Network-to-knowledge

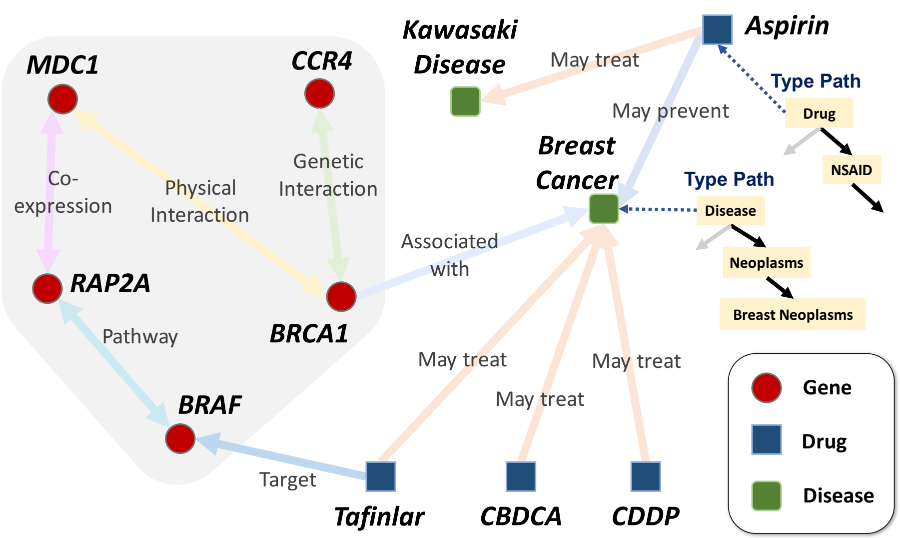
?



Knowledge & Insights

Looking Forward: Analyzing Literature to **Facilitate Scientific Research**

- Literature → Knowledge Base→ Scientific Discovery
- More disciplines & More structure analysis functions



Scientific Hypothesis Generation by predicting missing relationships

Gaining insights for various research tasks in different disciplines

Collaborate with life scientists, chemists, physicists, computer scientists, …

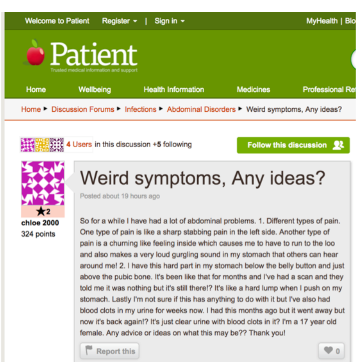
Looking Forward: Engaging with **Human Behaviors**

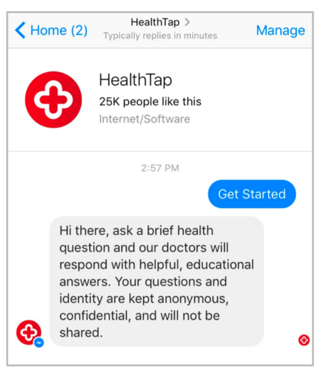
User-generated Content **(Structured Network)** + Structured Behavior Data ? → Personalized Intelligent Systems

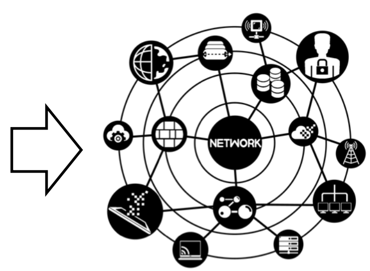
Social media post,
Customer review,
Chats & messages

Social network,
Electronic health record,
Transaction record

Smart Health,
Business intelligence,
Conversational agent

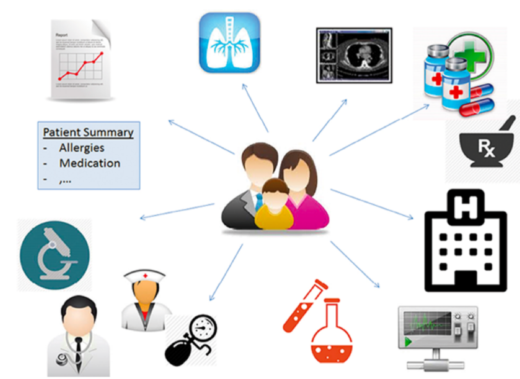






User content to structured network

+

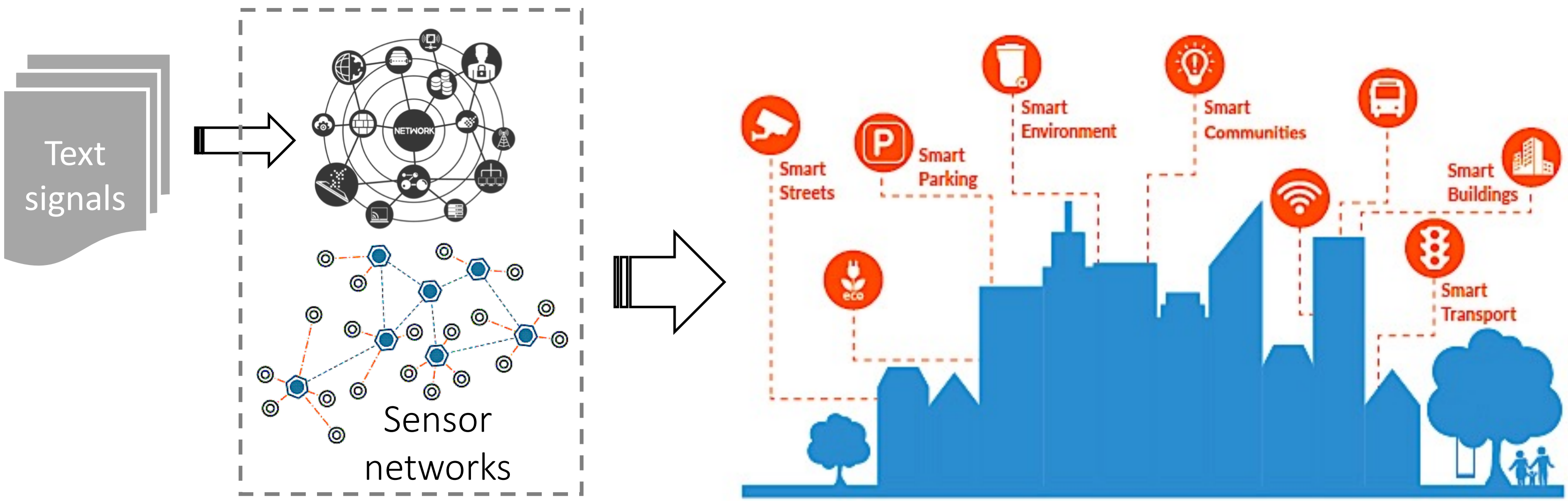


Structured information from eHealth records

Collaborate with doctors, social scientists, economists, …

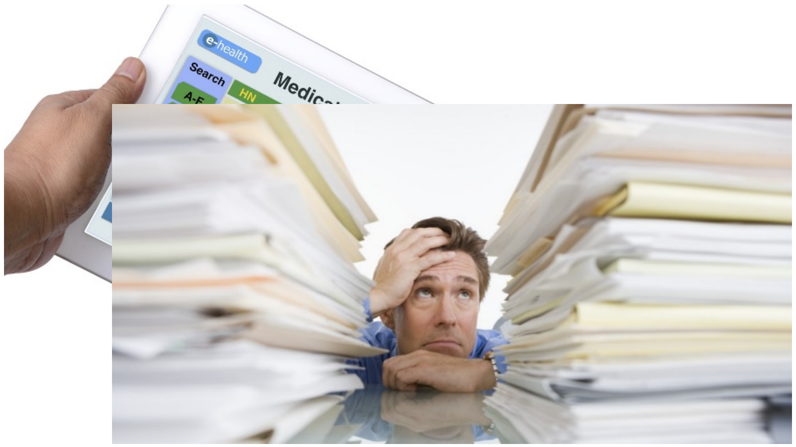
Looking Forward: Integrating with **Our Physical World**

Textual Signals **(Structured Network)**	+	Physical Sensor (Network) Signals	**?** → **Data Analytics**	Smart-City Operating Systems
News streams, Social media post, Organization report		Geo-sensors, Audio/video sensors, Bio-sensors		**Traffic management, Sustainable urban system, Cyber-physical system**



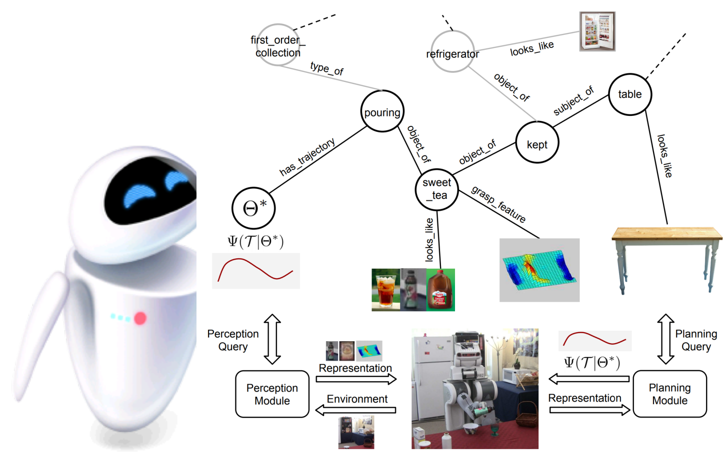
Collaborate with network & system researchers, environmental scientists, …

Application to Vertical Domains



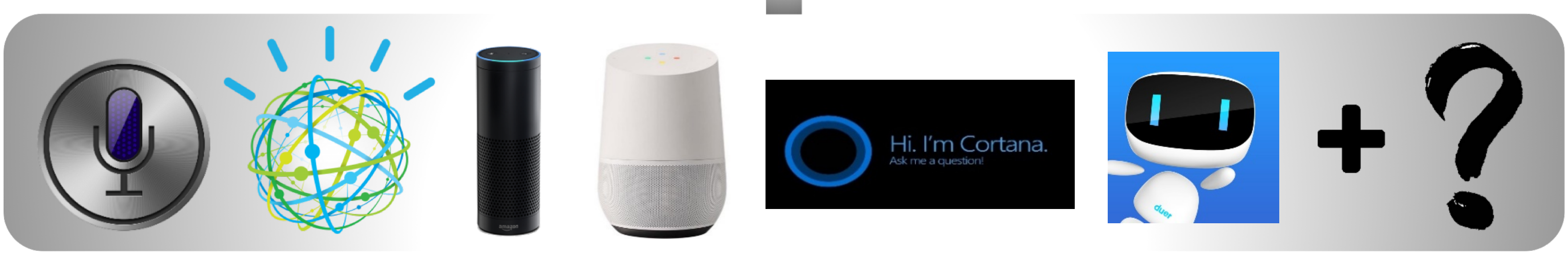
"Which cement stocks go up the most when a Category 3 hurricane hits Florida?"

KENSHO

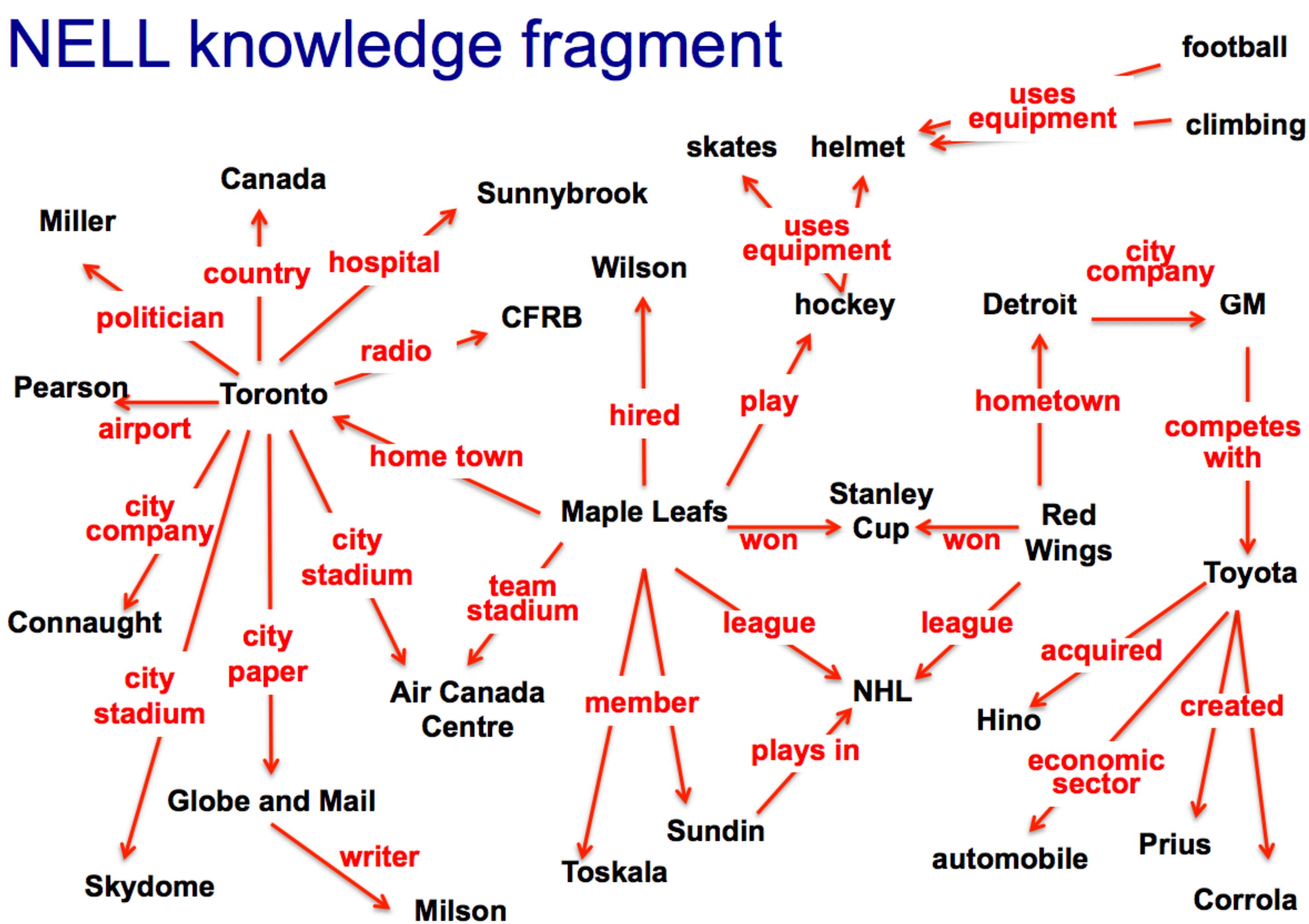


One Interface for All

- All domains in a unified knowledge base
- Incrementally learn new domains without forgetting (or instead boosting) existing ones



Learning to Reason for KB Completion



Slide from Tom Mitchell.

A Tale of Three Stories

- Embedding-Based Approaches:
  - Light-weight, scalable, and robust.
- Path-Based Approaches:
  - Explainable and interpretable.
- Deep Reinforcement Learning Based:
  - Integrate embedding and path based methods seamlessly.

SOTAs for Reasoning on KBs

- ConvE (Dettmers et al., AAAI 2018)
- Poincare (Nickel and Kiela, NIPS 2017)

- DeepPath (Xiong et al., EMNLP 2017).
- MINERVA (Das et al., ICLR 2018).

- DIVA (Chen et al., NAACL 2018).

Open-sourced Software

- Entity recognition and typing:
  - ClusType: http://shanzhenren.github.io/ClusType/
  - LM-LSTM-CRF: https://github.com/LiyuanLucasLiu/LM-LSTM-CRF
  - CrossType Name Tagger: https://github.com/yuzhimanhua/LM-LSTM-CRF
  - Multi-tasking LSTM-CRF: https://github.com/hltcoe/golden-horse/
- Relation extraction:
  - CoType: https://github.com/shanzhenren/CoType
  - ReQuest: https://github.com/shanzhenren/ReQuest
  - GraphLSTM: http://hanover.azurewebsites.net/
- KB reasoning:
  - DeepPath: https://github.com/xwhan/DeepPath
  - KBGAN: https://github.com/cai-lw/KBGAN
  - ProPPR: https://github.com/TeamCohen/ProPPR

Thank you! Q&A

- **Effort-Light Structure Extraction**
  → Corpus-specific labeling free, domain/language-independent
- **Joint Models for Low-resource IE:** jointly learning representations from unlabeled data, linguistic structures, annotations from other tasks, domains, and languages. → Reusable knowledge
- **Reasoning**: leverage embedding and path based models for discovering new knowledge.
- A principled approach to manage, explore, and analyze "Big Text Data"