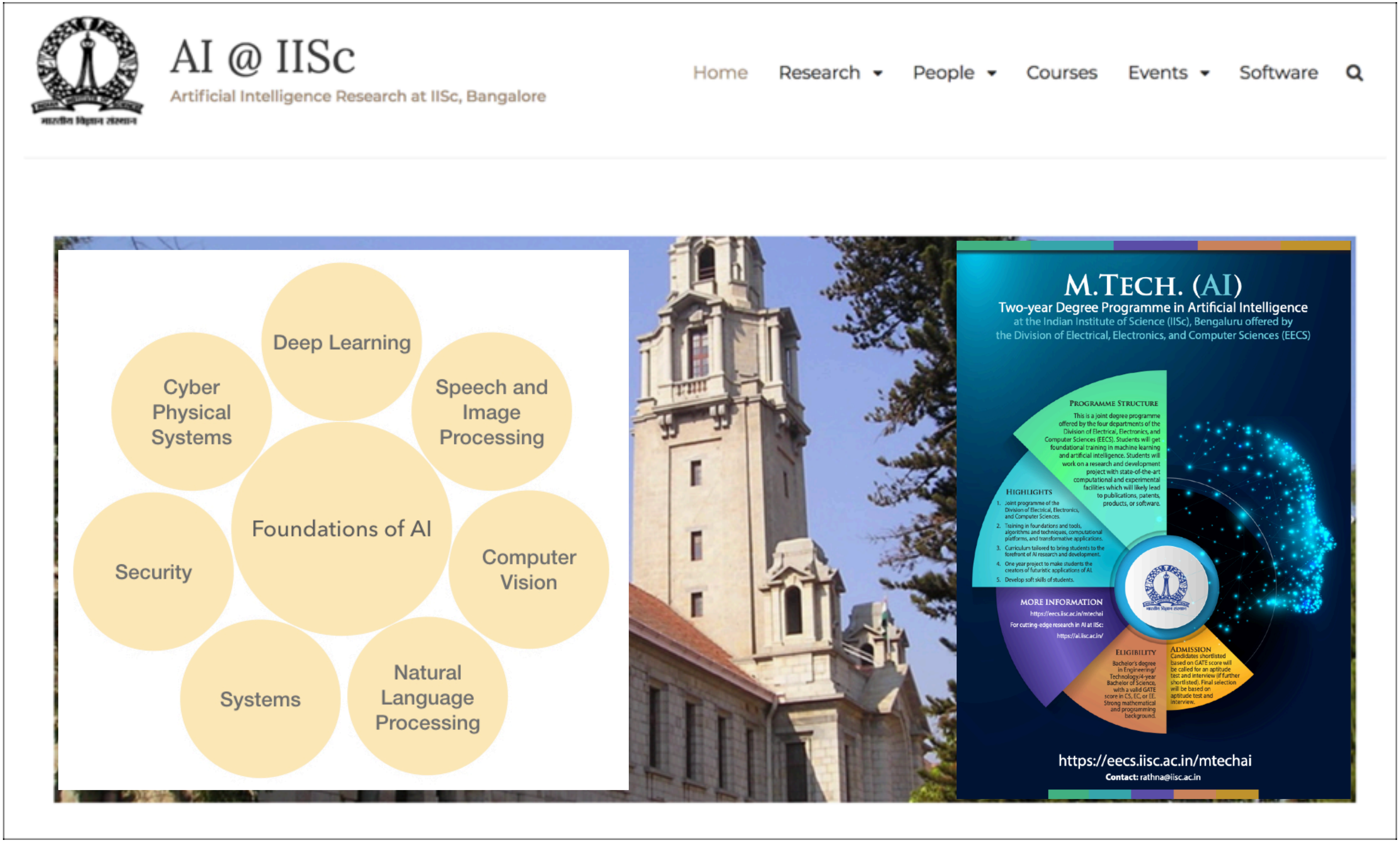
Knowledge Infused Deep Learning

Partha Talukdar
IISc Bangalore and KENOME
ppt@iisc.ac.in



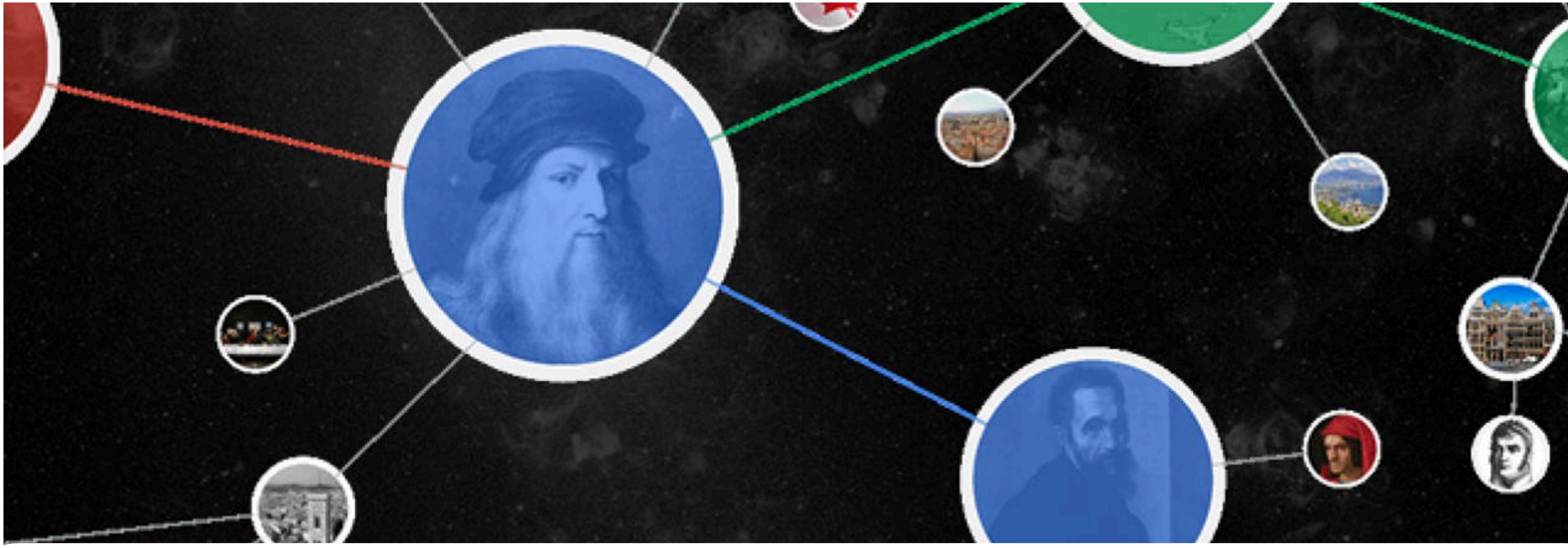
http://ai.iisc.ac.in

Thesis

Background knowledge is key to Intelligent Decision Making

State Farm stocks tumbled along with **Berkshire Hathaway**

insured? ***subsidiary***



Knowledge Graph (KG): Things, not Strings

For this talk:

KG = Multi-relational Graphs

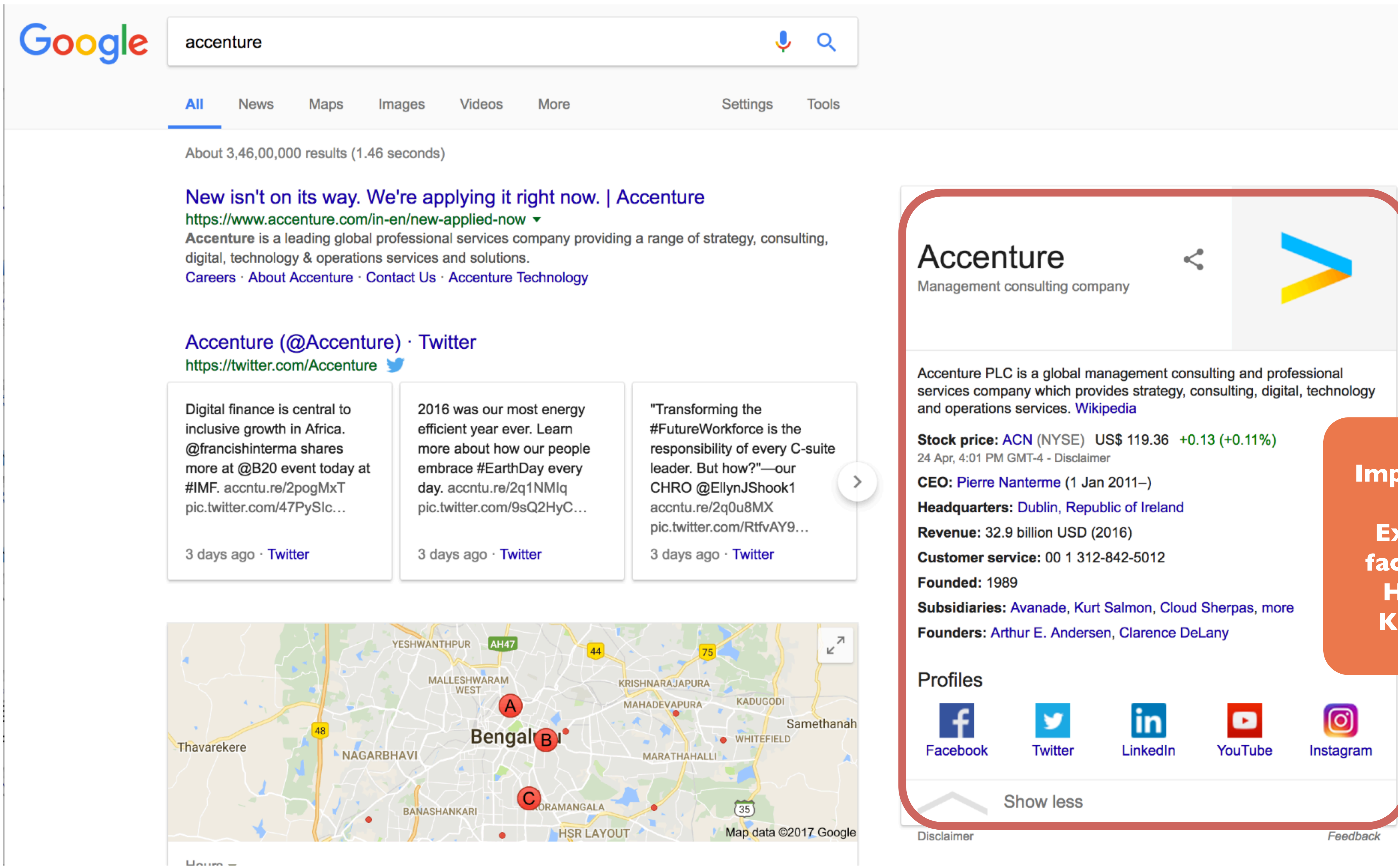
(e.g., factual, syntactic, temporal, etc.)

Thesis

Deep Learning (DL) augmented with KG improves performance

DL + KG > DL

Use case: Web Search



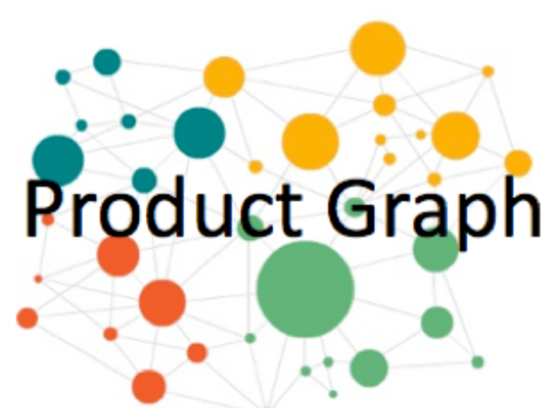
Improved Web Search Experience, facilitated by Harvested Knowledge

Use case: Conversational AI

Knowledge Graphs can provide a shared context

Google Knowledge Graph

Amazon



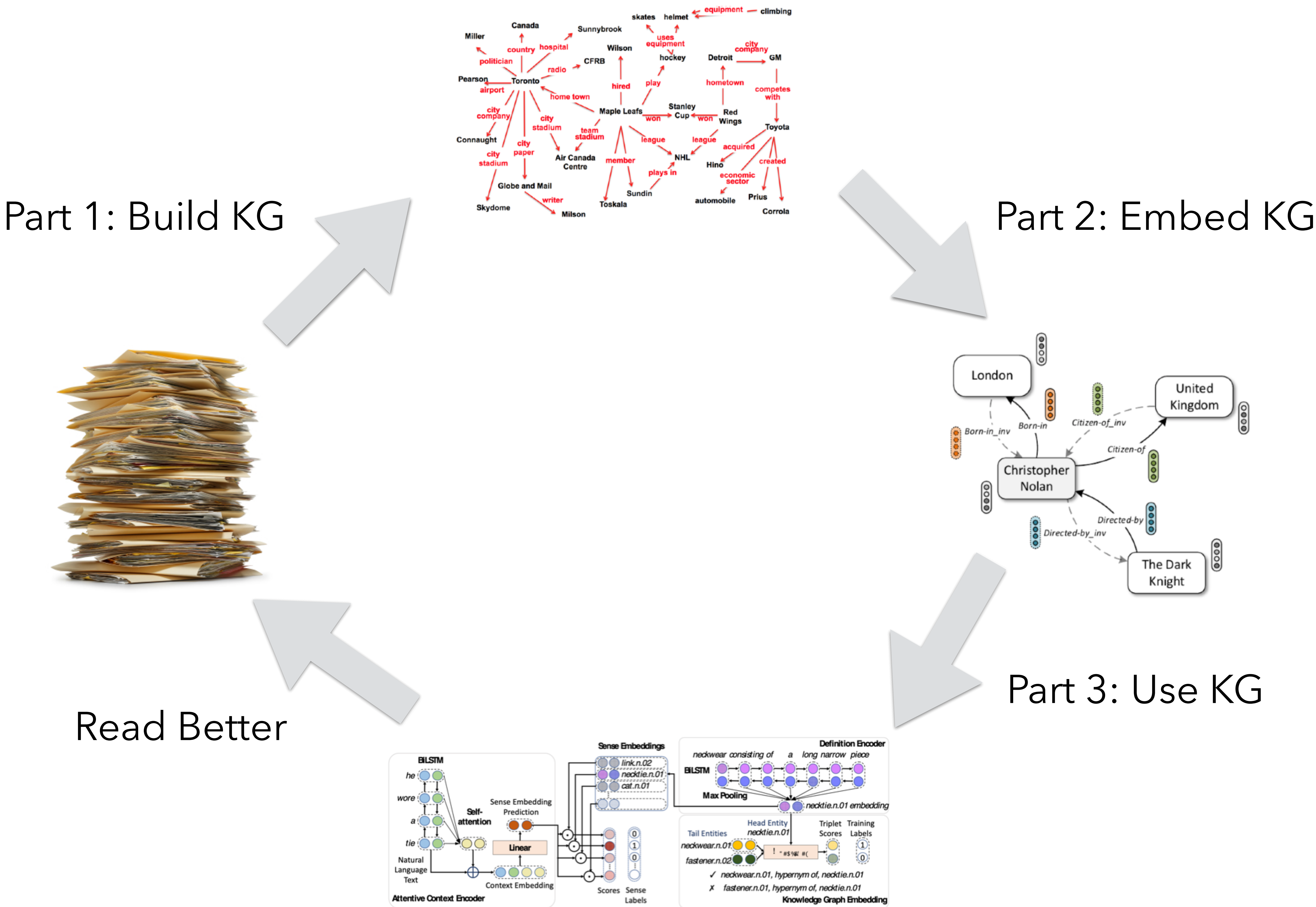
Microsoft Satori

Facebook Entity Graph

LinkedIn Graph

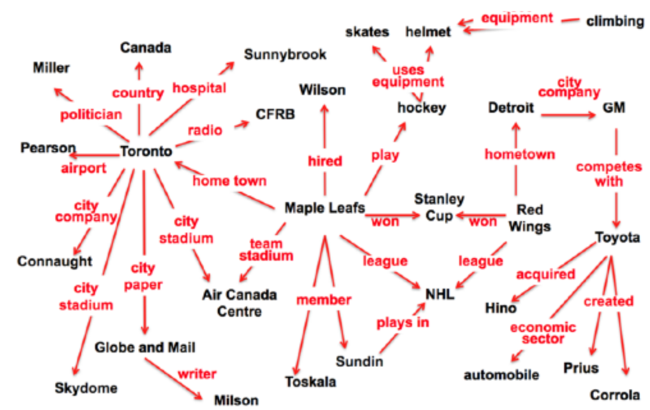
This Talk

- **Part 1 (Build KG):** Where do we get KGs from?
- **Part 2 (Embed KG):** How do we embed KGs?
- **Part 3 (Use KG):** How to use embedded KG in DL?



Part 1: Build KG
Part 2: Embed KG
Part 3: Use KG
Read Better
London
United Kingdom
Christopher Nolan
The Dark Knight
Born-in
Born-in_inv
Citizen-of
Citizen-of_inv
Directed-by
Directed-by_inv
Toronto
Maple Leafs
Stanley Cup
Red Wings
NHL
Toyota
Sense Embeddings
Definition Encoder
Attentive Context Encoder
Knowledge Graph Embedding

Part 1: Build KG



Part 2: Embed KG

Part 3: Use KG

Read Better

KG Construction Efforts

Cycorp®
Home of smarter solutions

Freebase

High Supervision

Amazon
Product Graph

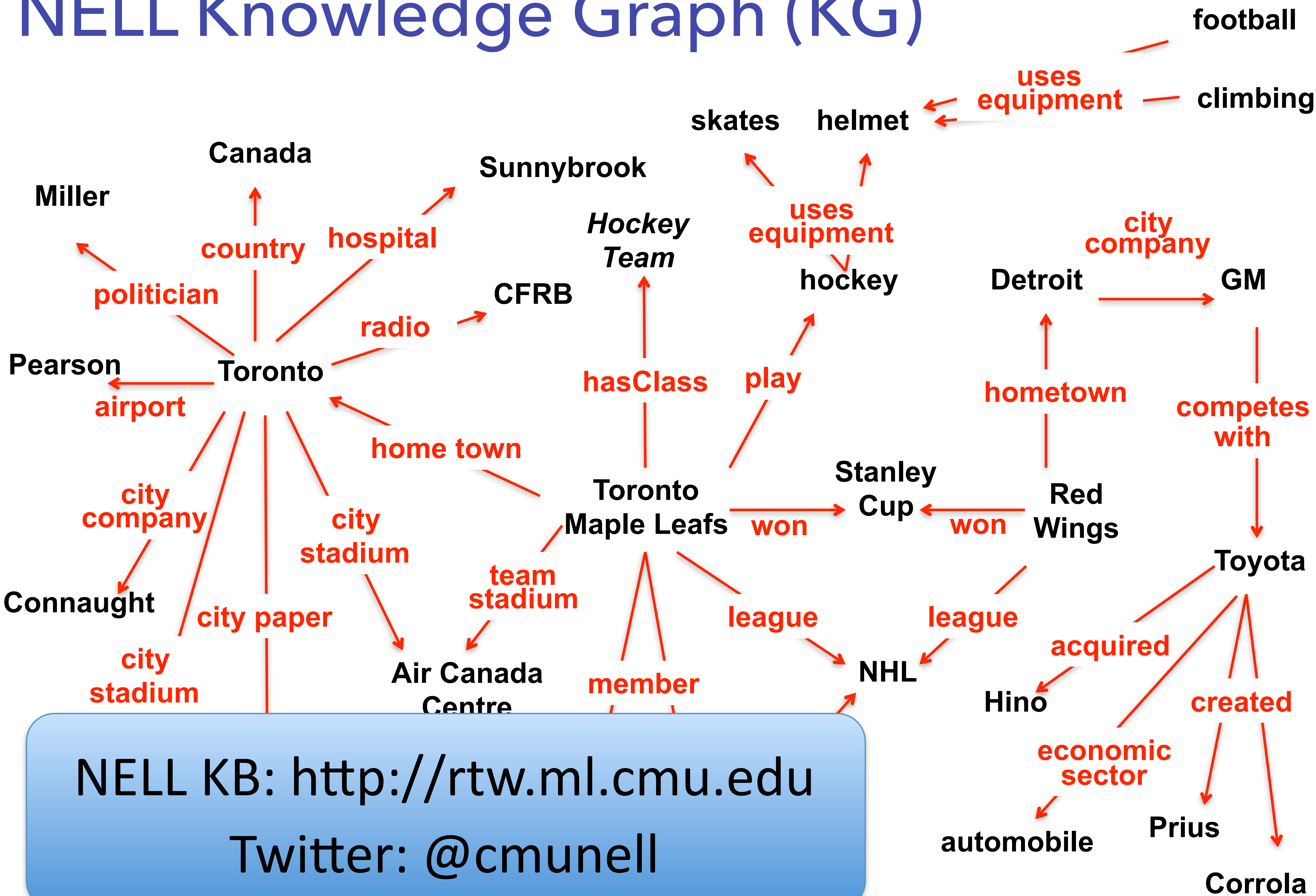
yago
select knowledge

DeepDive

NELL

Low Supervision

NELL Knowledge Graph (KG)



NELL KB: http://rtw.ml.cmu.edu

Twitter: @cmunell

DOI:0 0.0000/000

Never-Ending Learning

By T. Mitchell, W. Cohen, E. Hruschka, P. Talukdar, B. Yang, J. Betteridge, A. Carlson, B. Dalvi, M. Gardner, B. Kisiel, J. Krishnamurthy, N. Lao, K. Mazaitis, T. Mohamed, N. Nakashole, E. Platanios, A. Ritter, M. Samadi, B. Settles, R. Wang, D. Wijaya, A. Gupta, X. Chen, A. Saparov, M. Greaves, and J. Welling

Abstract

Whereas people learn many different types of knowledge from diverse experiences over many years, and become better learners over time, most current machine learning systems are much more narrow, learning just a single function or data model based on statistical analysis of a single data set. We suggest that people learn better than computers precisely because of this difference, and we suggest a key direction for machine learning research is to develop software architectures that enable intelligent agents to also learn many types of knowledge, continuously over many years, and to become better learners over time. In this paper we define more precisely this *never-ending learning* paradigm for machine learning, and we present one case study: the Never-Ending Language Learner (NELL), which achieves a number of the desired properties of a never-ending learner. NELL has been learning to read the Web 24hrs/day since January 2010, and so far has acquired a knowledge base with 120mn diverse, confidence-weighted beliefs (e.g., *servedWith(tea,biscuits)*), while learning thousands of interrelated functions that continually improve its reading competence over time. NELL has also learned to reason over its knowledge base to infer new beliefs it has not yet read from those it has, and NELL is inventing new relational predicates to extend the ontology it uses to represent beliefs. We describe the design of NELL, experimental results illustrating its behavior, and discuss both its successes and shortcomings as a case study in never-ending learning. NELL can be tracked online at http://rtw.ml.cmu.edu, and followed on Twitter at @CMUNELL.

1. INTRODUCTION

Machine learning is a highly successful branch of Artificial Intelligence (AI), and is now widely used for tasks from spam filtering, to speech recognition, to credit card fraud detection, to face recognition. Despite these successes, the ways in which computers learn today remain surprisingly narrow when compared to human learning. This paper explores an alternative paradigm for machine learning that more closely models the diversity, competence and cumulative nature of human learning. We call this alternative paradigm *never-ending learning.*

To illustrate, note that in each of the above machine learning applications, the computer learns only a single function to perform a
human labeled traini
of that function. In s
examples consist of s
labels for each. This s
vised function approxi
problem is to approxi

CACM 2018

(e.g., the spam filter) given a training set of input/output pairs $\{\langle x_i, y_i\rangle\}$ of that function. Other machine learning paradigms exist as well (e.g., unsupervised clustering, topic modeling, reinforcement learning) but these paradigms also typically acquire only a single function or data model from a single dataset.

In contrast to these paradigms for learning single functions from well organized data sets over short time-frames, humans learn many different functions (i.e., different types of knowledge) over years of accumulated diverse experience, using extensive background knowledge learned from earlier experiences to guide subsequent learning. For example, humans first learn to crawl, then to walk, run, and perhaps ride a bike. They also learn to recognize objects, to predict their motions in different circumstances, and to control those motions. Importantly, they learn *cumulatively*: as they learn one thing this new knowledge helps them to more effectively learn the next, and if they revise their beliefs about the first then this change refines the second.

The thesis of our research is that *we will never truly understand machine or human learning until we can build computer programs that, like people,*

- learn many different types of knowledge or functions,
- from years of diverse, mostly self-supervised experience,
- in a staged curricular fashion, where previously learned knowledge enables learning further types of knowledge,
- where self-reflection and the ability to formulate new representations and new learning tasks enable the learner to avoid stagnation and performance plateaus.

We refer to this learning paradigm as "never-ending learning." The contributions of this paper are to (1) define more precisely the never-ending learning paradigm, (2) present as a case study a computer program called the NELL which implements several of these capabilities, and which has been learning to read the Web 24hrs/day since January 2010, and (3) identify from NELL's strengths and weaknesses a number of key design features important to any never-ending learning system. This paper is an elaboration and extension to an earlier overview of the NELL system.[27]

2. RELATED WORK

...nsidered the problem of design-
...nts that persist over long periods

...f this paper appeared in the
... *AAAI Conference on Artificial*
...an. 25–30, 2015), 2302–2310.

Never-Ending Learning
Tom Mitchell · Partha Talukdar

Tutorial

Mon Jun 10th 09:15 -- 11:30 AM @ Hall B

in Tutorials Hall B »

There exists a stark difference between today's machine learning methods and the lifelong learning capabilities of humans. Humans learn many different functions and skills, from diverse experiences gained over many years, from a staged curriculum in which they first learn easier and later more difficult tasks, retain the learned knowledge and skills, which are used in subsequent learning to make it easier or more effective. Furthermore, humans self-reflect on their evolving skills, choose new learning tasks over time, teach one another, learn new representations, read books, discuss competing hypotheses, and more. This tutorial will focus on the question of how to design machine learning agents with similar capabilities. The tutorial will include research on topics such as reinforcement learning and other agent learning architectures, transfer and multi-task learning, representation learning, amortized learning, learning by natural language instruction and demonstration, learning from experimentation.

Slides » Video »

Author Information

Tom Mitchell (Carnegie Mellon University)

Tom M. Mitchell is the Founders University Professor and Interim Dean of the School of Computer Science at Carnegie Mellon University. Mitchell has worked in Machine Learning for many years, and co-founded the ICML conference (with Jaime Carbonell and Ryszard Michalski). Recently, he directed the Never-Ending Language Learning (NELL) project, which operated continuously for over eight years, providing a case study for how to architect never-ending learning systems. Mitchell is a member of the U.S. National Academy of Engineering, a member of the American Academy of Arts and Sciences, and a Fellow and Past President of the Association for the Advancement of Artificial Intelligence (AAAI).

Partha Talukdar (IISc Bangalore / KENOME)

Partha Talukdar is a faculty member in the Department of Computational and Data Sciences (CDS) at the Indian Institute of Science (IISc), Bangalore. He is also the founder of KENOME, an enterprise Knowledge graph company with the mission to help enterprises make sense of big dark data. Previously, Partha was a Postdoctoral Fellow in the Machine Learning Department at Carnegie Mellon University, working with Tom Mitchell on the NELL project. Partha received his PhD (2010) in CIS from the University of Pennsylvania, working under the supervision of Fernando Pereira, Zack Ives, and Mark Liberman. Partha is broadly interested in Machine Learning, Natural Language Processing, and Cognitive Neuroscience, with particular interest in large-scale learning and inference. Partha is a recipient of IBM Faculty Award, Google's Focused Research Award, and Accenture Open Innovation Award. He is a co-author of a book on Graph-based Semi-Supervised Learning published by Morgan Claypool Publishers. Homepage: http://talukdar.net

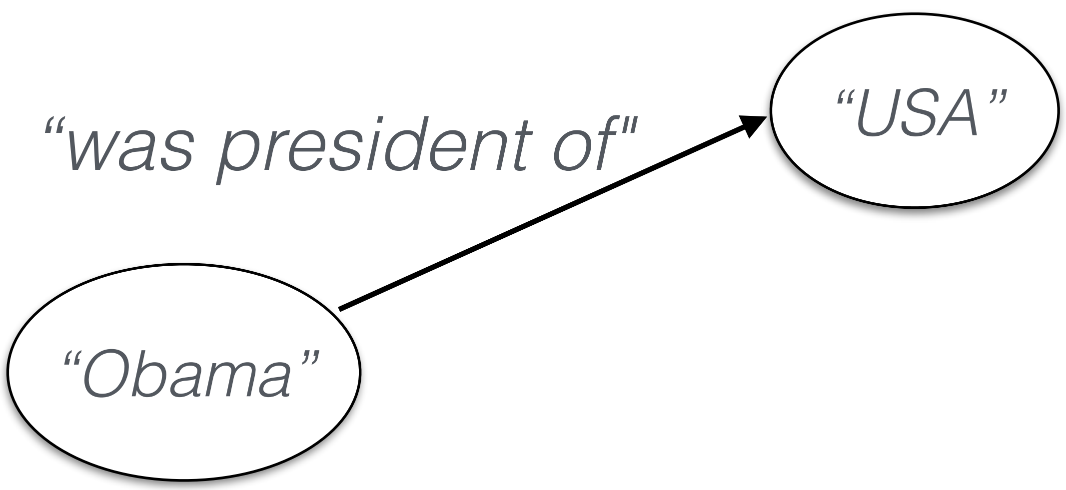
Tutorial @ ICML 2019
bit.ly/nel-icml19-tutorial

Tutorial @ KDD 2018
bit.ly/kg-kdd18-tutorial

Two Types of Knowledge Graphs

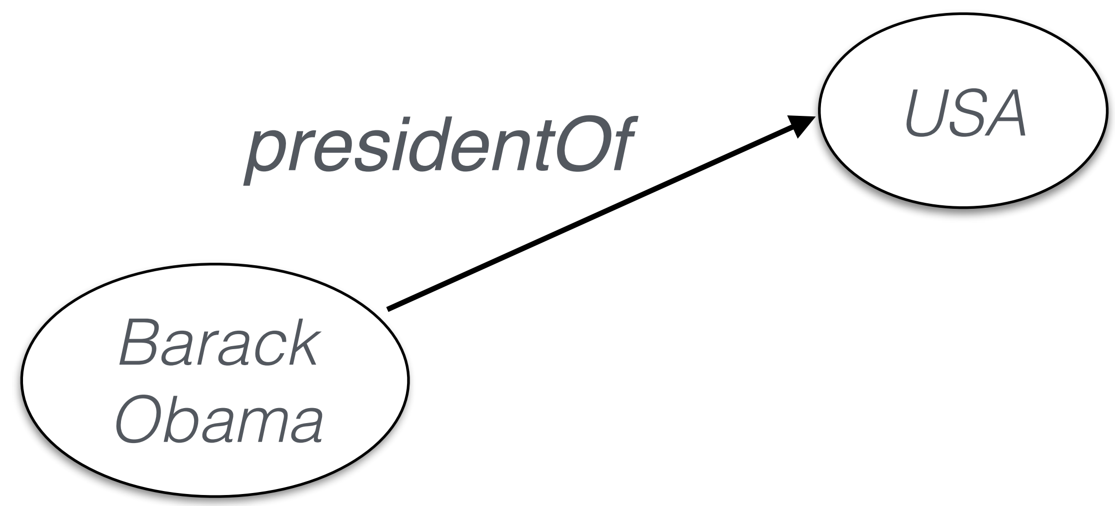
“Obama was the President of USA.”

Open KG
(Ontology Free)



- + easy to build, available tools
- + high recall
- - fragmented (more later)

Ontological KG



- + high precision
- + canonicalized/normalized
- - requires supervision

How to transition from Open to Ontological KG?

Domain-specific Relation Schema Induction

Joint work with

Madhav N.

Uday Saini

Manish Gupta
(Microsoft)

EMNLP 2016, ACL 2018

https://github.com/malllabiisc/sictf

Domain-specific Knowledge Graphs (KG)

- Need KGs in specific domains (e.g., insurance, automotives, etc.)

- General purpose KGs (e.g., Freebase, YAGO, NELL, etc.) are good starting points, but often not enough

- Problem: how to build KG out of documents from a given domain, with minimal supervision?

Relation Schema Induction

- Relation Schemas [e.g., *undergo(Patient, Surgery)*]
  - starting point in ontological KG construction
  - prepared by experts: expensive and incomplete

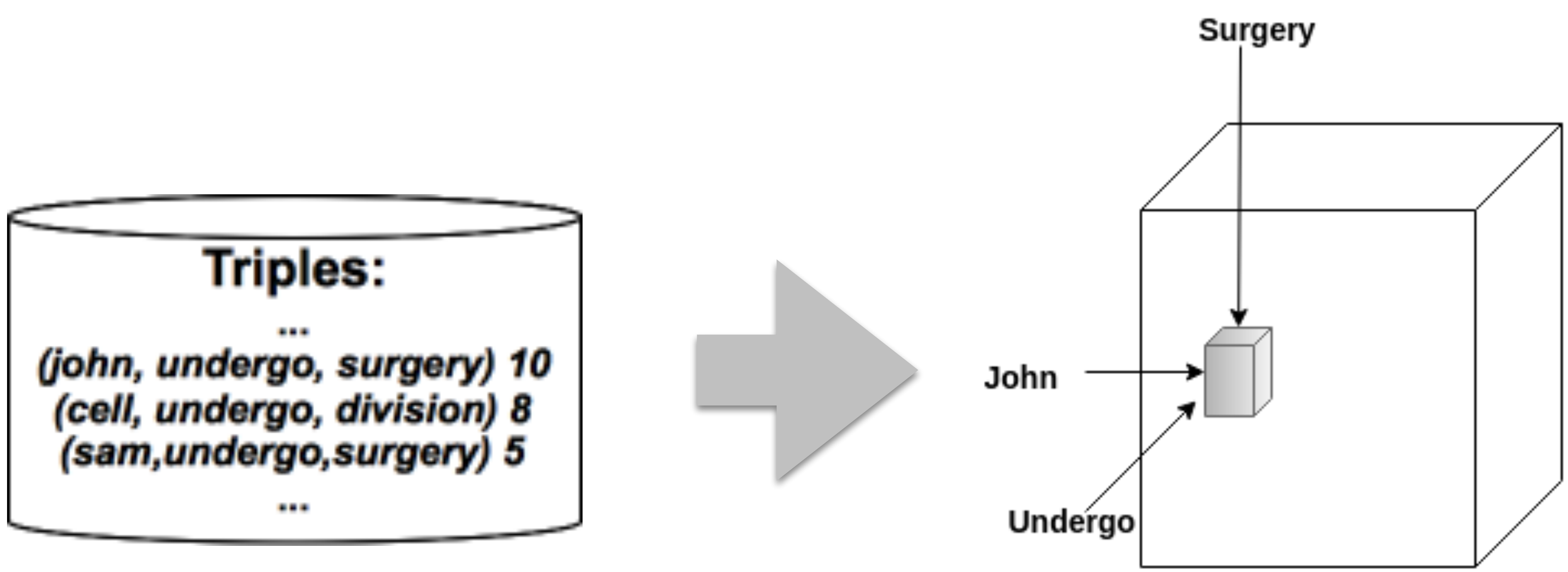
"… John underwent angioplasty last Tuesday …"
"… Sam will undergo Tonsillectomy …"
…
"… cells that undergo meiosis …"

…
undergo(Patient, Surgery)
undergo(Cell, Division)
…

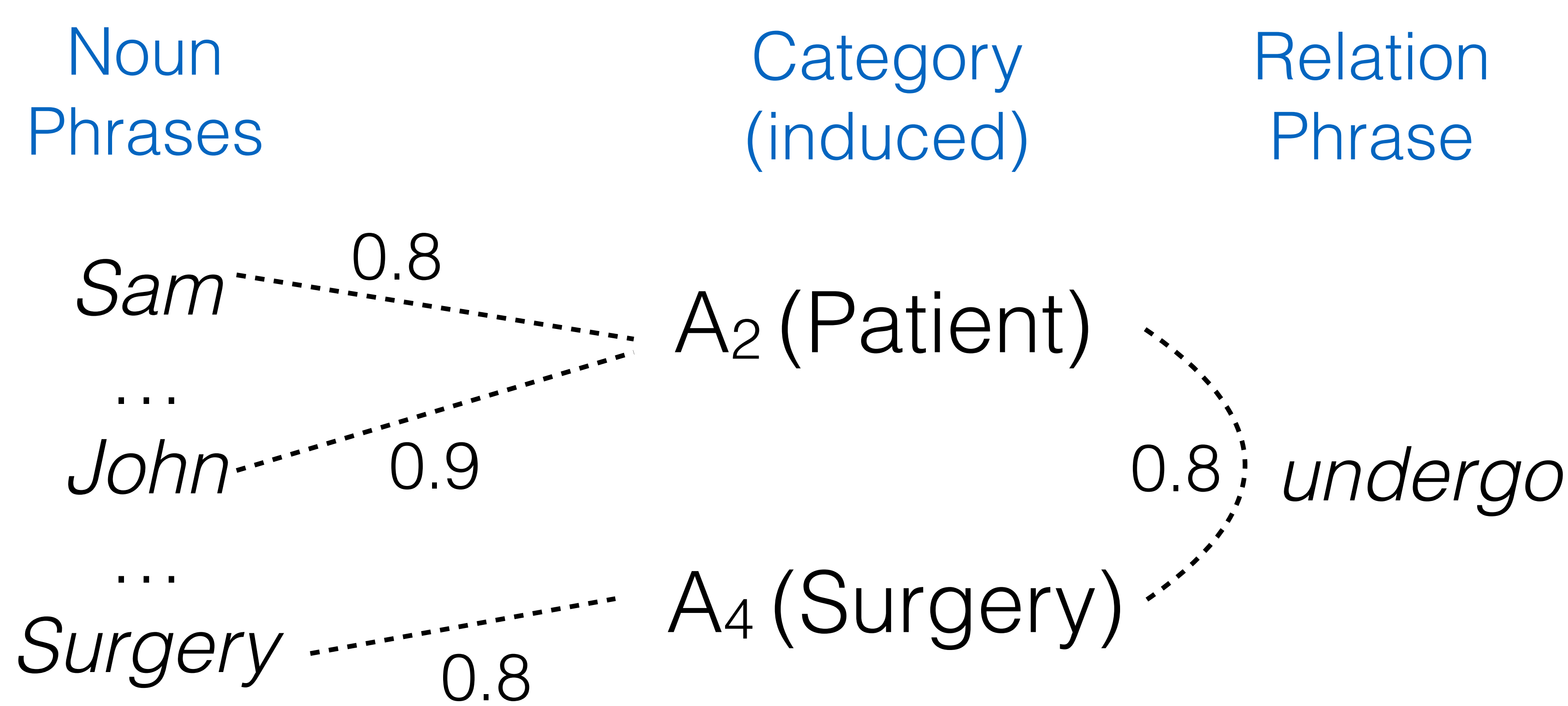
How to automatically identify relations and their schemas from domain documents?

OpenIE Tensor

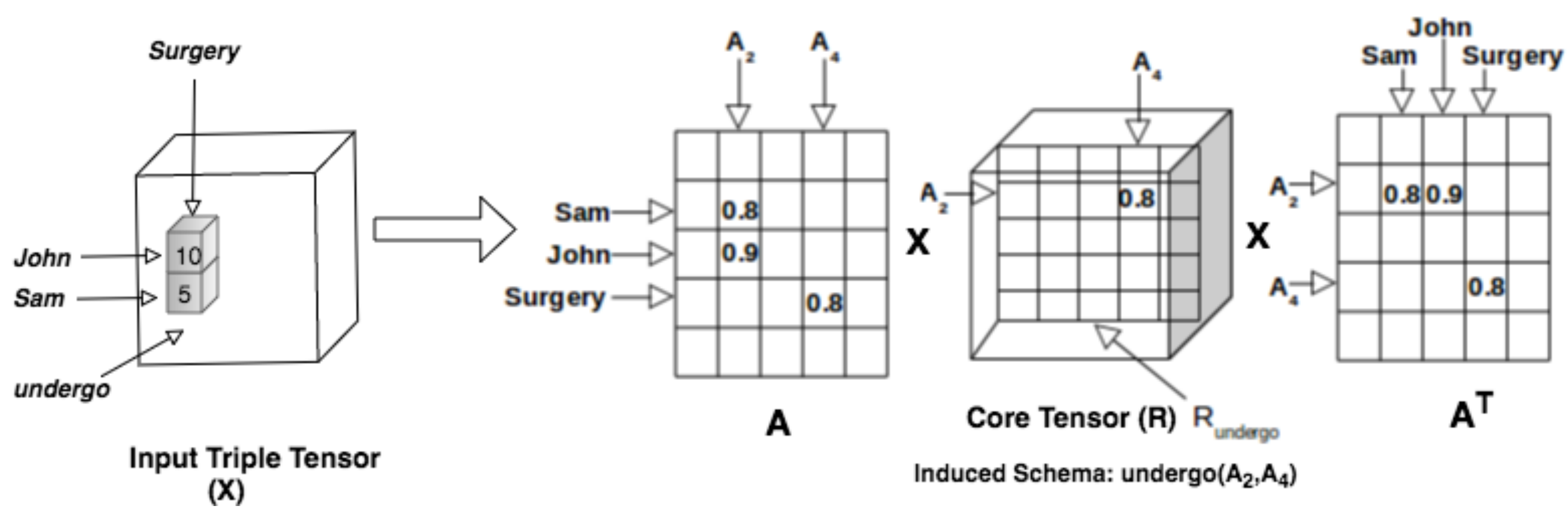
- *“Sam will undergo Surgery next Tuesday”*
  => (Sam, undergo, Surgery)
- OpenIE triples is a good starting point
- Tensors provide a natural way to represent such triples



RSI as Tensor Factorization



RESCAL



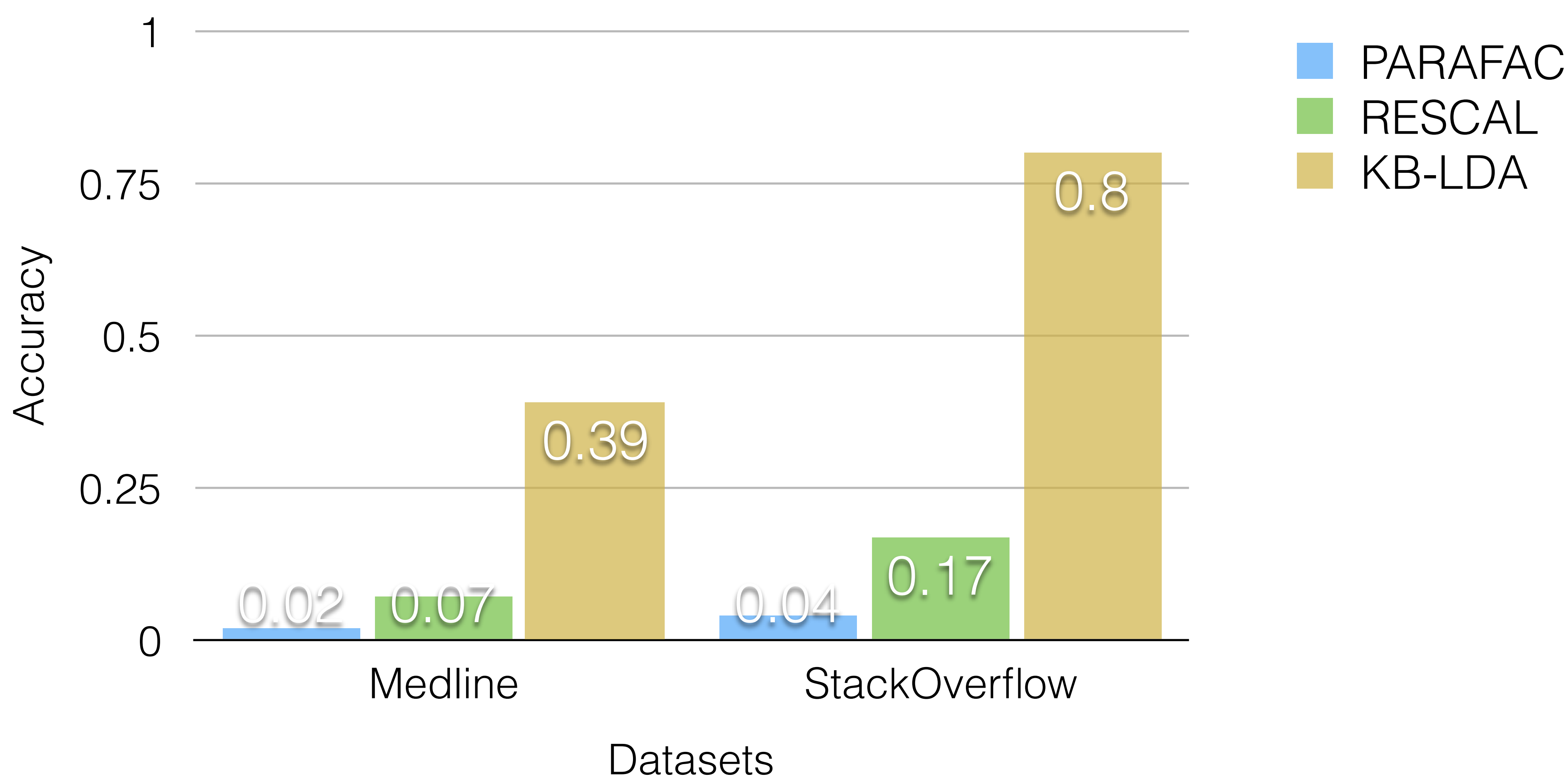
Evaluation Protocol

Dataset	**# Docs**	**# Triples**
MEDLINE	50,216	13,308
StackOverflow	5.5m	37,439

Table 4: Datasets used in the experiments.

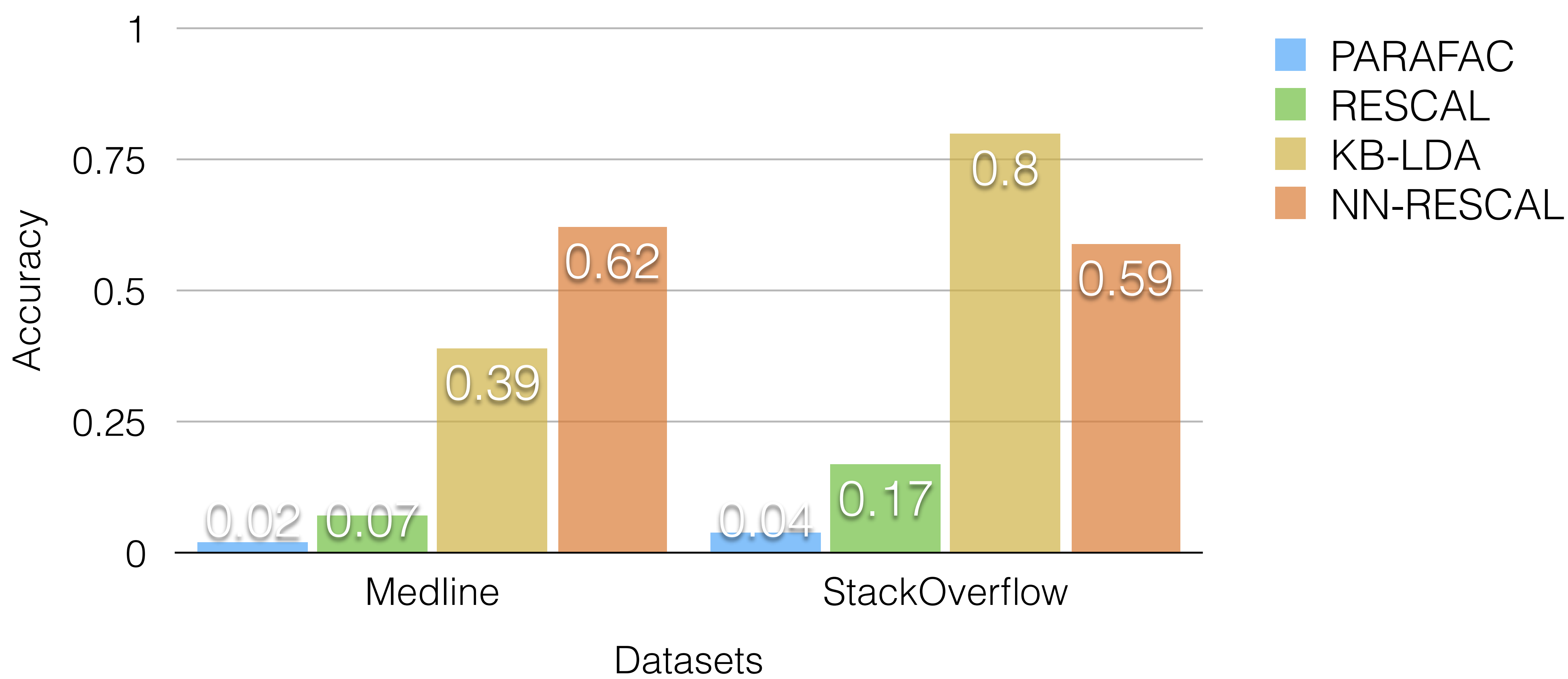
Relation Schema	Top 3 NPs in Induced Categories which were presented to annotators	Annotator Judgment
	StackOveflow	
clicks(A_0, A_1)	A_0: *users, client, person* A_1: *link, image, item*	valid
refreshes(A_{19}, A_{13})	A_{19}: *browser, window, tab* A_{13}: *page, activity, app*	valid
can_parse(A_{41}, A_{17})	A_{41}: *access, permission, ability* A_{17}: *image file, header file, zip file*	invalid
	MEDLINE	
receive(A_1, A_{18})	A_1: *patient, NUM patients, one patient* A_{18}: *flecainide, aerosolized pentamidine, prophylaxis*	valid
undergo(A_1, A_3)	A_1: *patient, NUM patients, one patient* A_3: *surgery, abdominal surgery, open heart surgery*	valid
fail_to(A_{32}, A_{36})	A_{32}: *chest pain, bacteriologic failure, unresectable disease* A_{36}: *nodular disease, valvular disease, Crohn disease*	invalid

Vanilla Tensor Factorization Fails



RESCAL outperforms PARAFAC, but both are significantly worse than state-of-the-art KB-LDA [Movshovitz-Attias & Cohen 2015]

Non-Negativity Helps



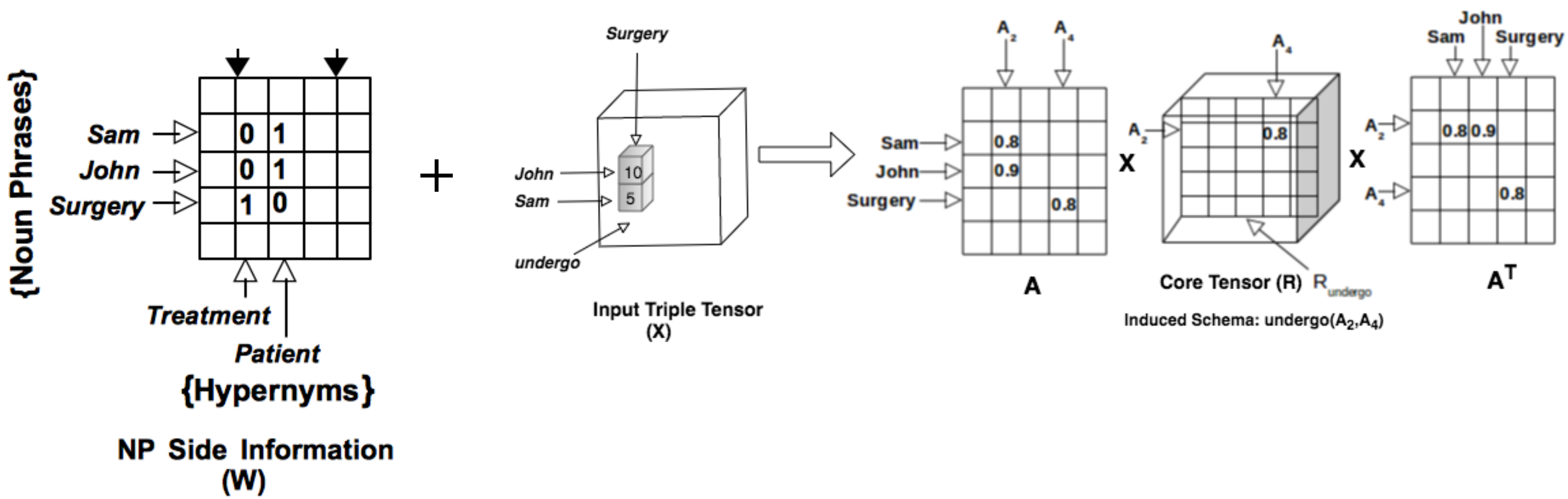
NN-RESCAL outperforms KB-LDA in MEDLINE dataset.

Noun Phrase (NP) Side Information

Capture NP type information using Hearst Patterns, e.g., "… *disease such as hypertension …*"

MEDLINE
(hypertension, disease), (hypertension, state), (hypertension, disorder) , (neutrophil, blood element), (neutrophil, effector cell), (neutrophil, cell type)
StackOverflow
(image, resource), (image, content), (image, file), (perl, language), (perl, script), (perl, programs)

Table 2: Noun Phrase (NP) side information in the form of (Noun Phrase, Hypernym) pairs extracted using Hearst patterns from two different datasets.



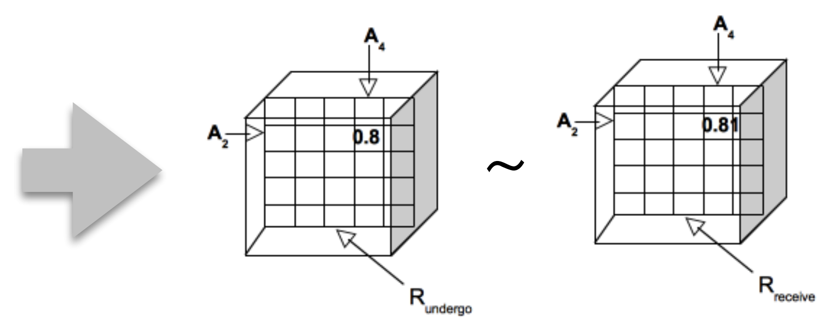
Factorization with NP Side Info

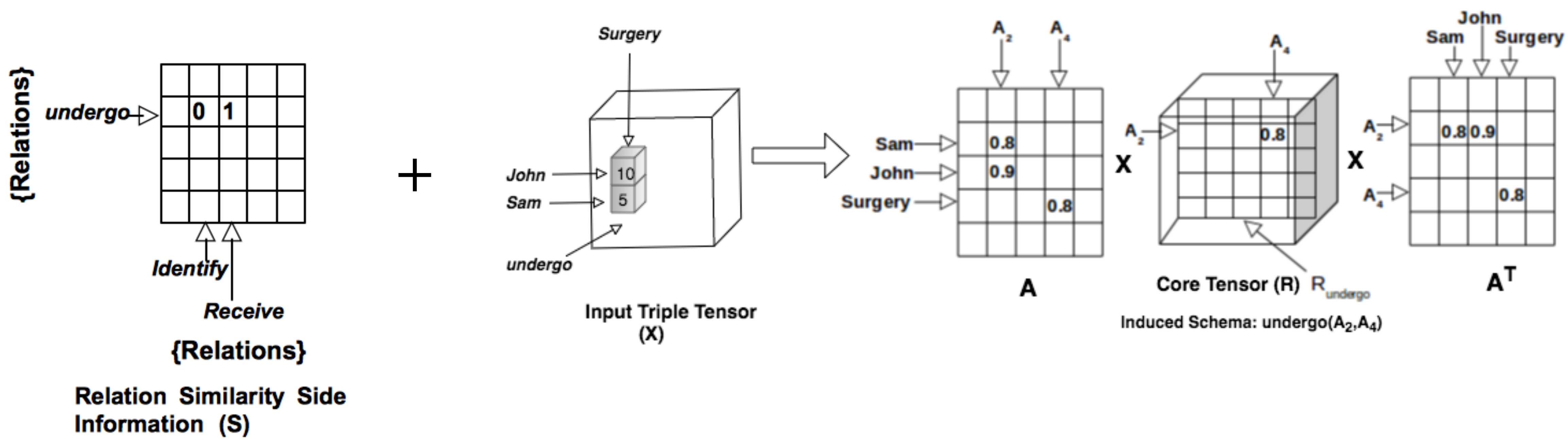
NP Side information is helpful, KB-LDA already makes use of such side information

Relational Side Information

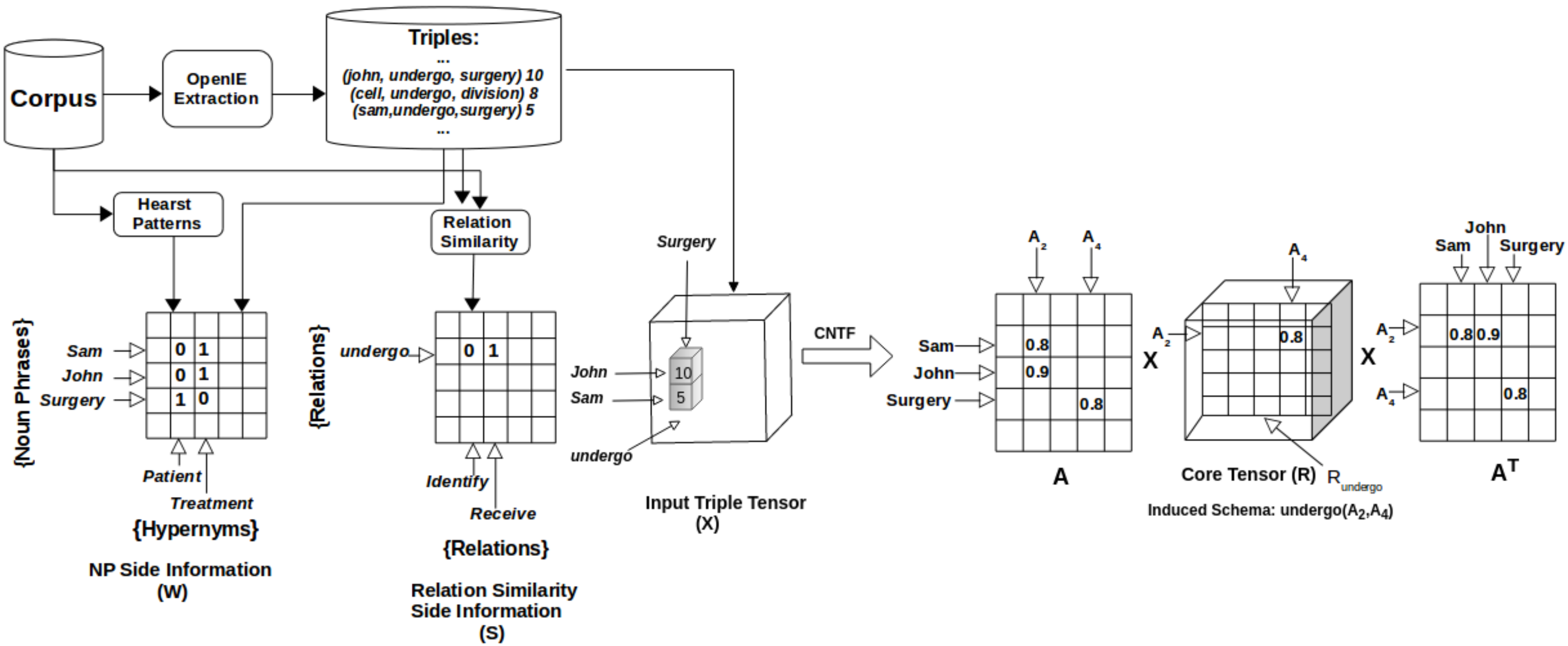
Similar relation phrases should have similar schemas
Embeddings can help calculate similarity

undergo ~ receive

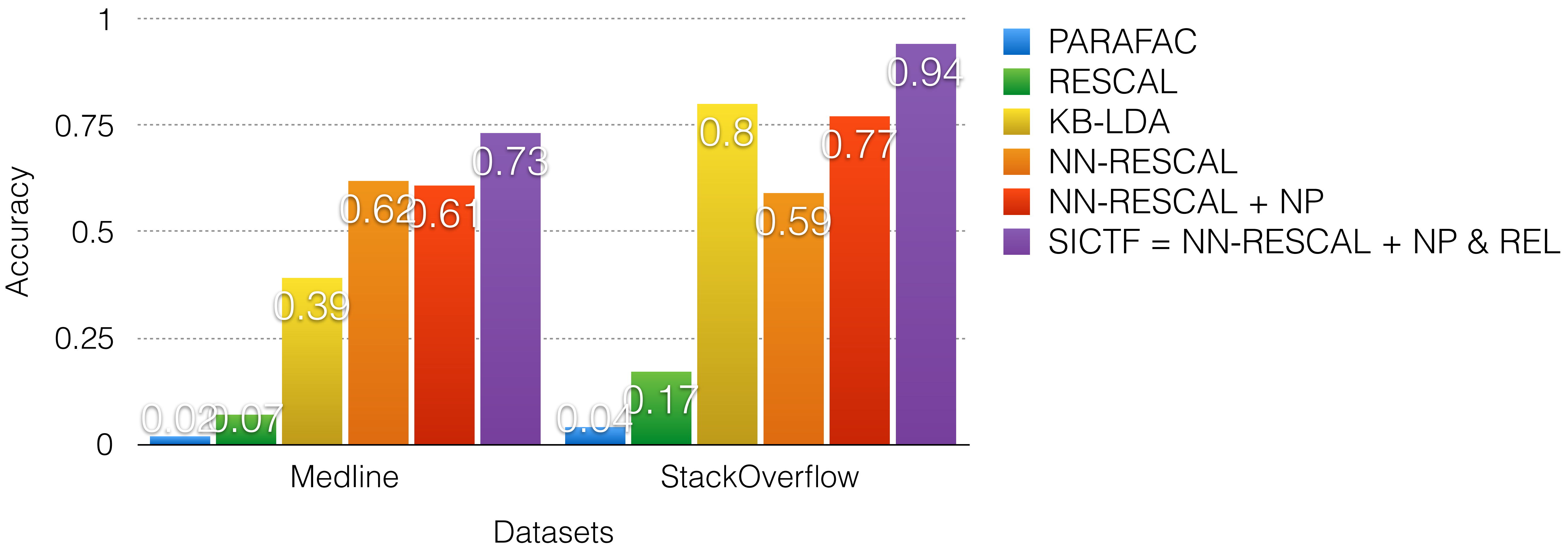




SICTF Architecture



SICTF: Factorization with NP & Relation Side Info



SICTF outperforms all baselines on both datasets

SICTF Objective

$$\min_{A,\{R_k\}} \sum_k [f(X_k, A, R_k)] + \overbrace{f_{np}(W, A, V)}^{\text{NP sideinfo}} + \underbrace{f_{rel}(S, R)}_{\text{relation similarity}}$$

where,

$$f(X_k, A, R_k) = \| X_k - AR_kA^T \|_F^2 + \lambda_R \| R_k \|_F^2$$

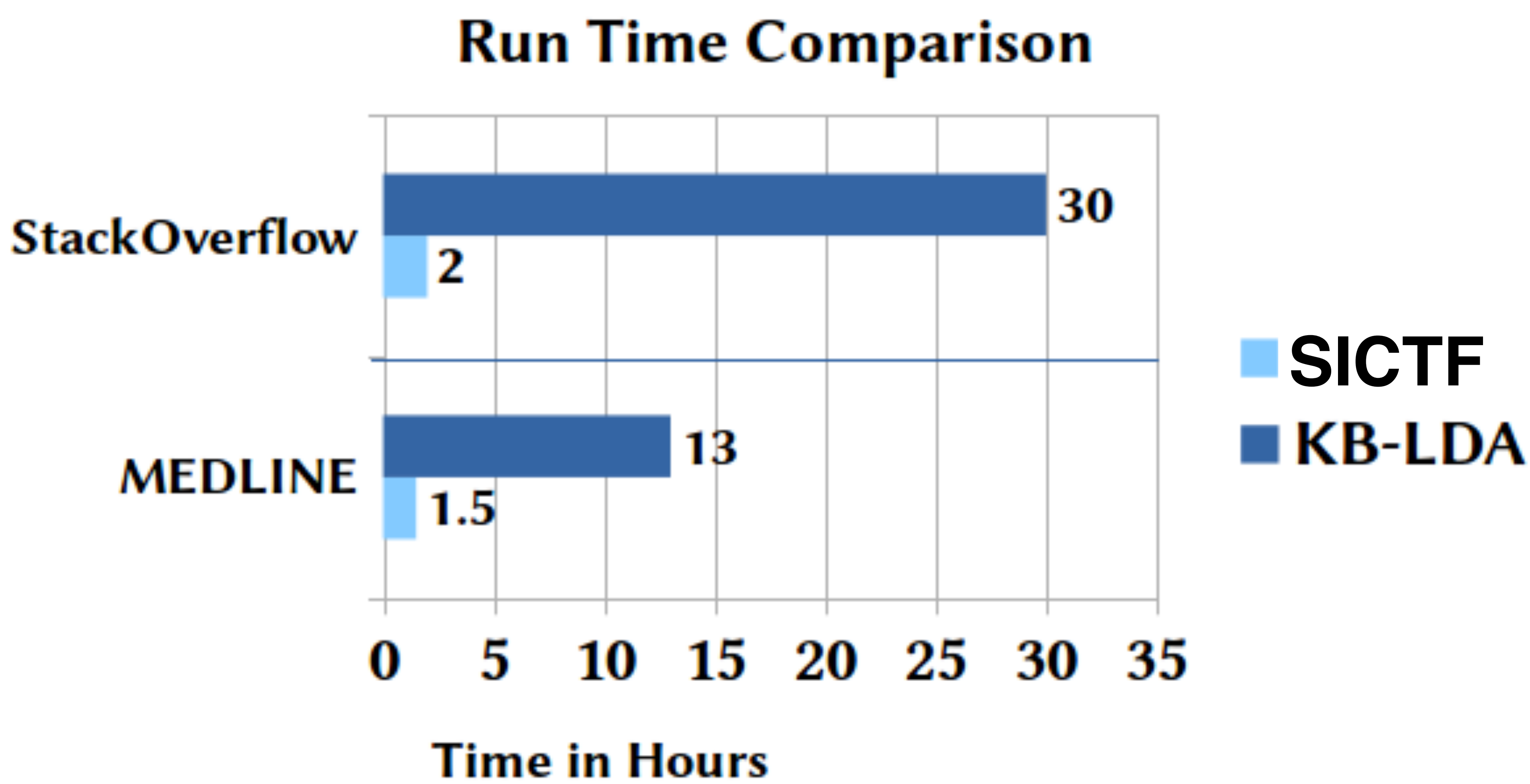
$$f_{np}(W, A, V) = \lambda_{np} \| W - AV \|_F^2 + \lambda_A \| A \|_F^2 + \lambda_V \| V \|_F^2$$

$$f_{rel}(S, R) = \lambda_{rel} \sum_{i=1}^{m} \sum_{j=1}^{m} S_{ij} \| R_i - R_j \|_F^2$$

Sample schemas induced by SICTF

Relation Schema	Top 3 NPs in Induced Categories which were presented to annotators	Annotator Judgment
	StackOveflow	
clicks(A_0, A_1)	A_0: *users, client, person* A_1: *link, image, item*	valid
refreshes(A_{19}, A_{13})	A_{19}: *browser, window, tab* A_{13}: *page, activity, app*	valid
can_parse(A_{41}, A_{17})	A_{41}: *access, permission, ability* A_{17}: *image file, header file, zip file*	invalid
	MEDLINE	
receive(A_1, A_{18})	A_1: *patient, NUM patients, one patient* A_{18}: *flecainide, aerosolized pentamidine, prophylaxis*	valid
undergo(A_1, A_3)	A_1: *patient, NUM patients, one patient* A_3: *surgery, abdominal surgery, open heart surgery*	valid
fail_to(A_{32}, A_{36})	A_{32}: *chest pain, bacteriologic failure, unresectable disease* A_{36}: *nodular disease, valvular disease, Crohn disease*	invalid

Runtime Comparison



SICTF achieves 11x+ speedup over KB-LDA, so better and faster!

TFB: Going beyond binary induction

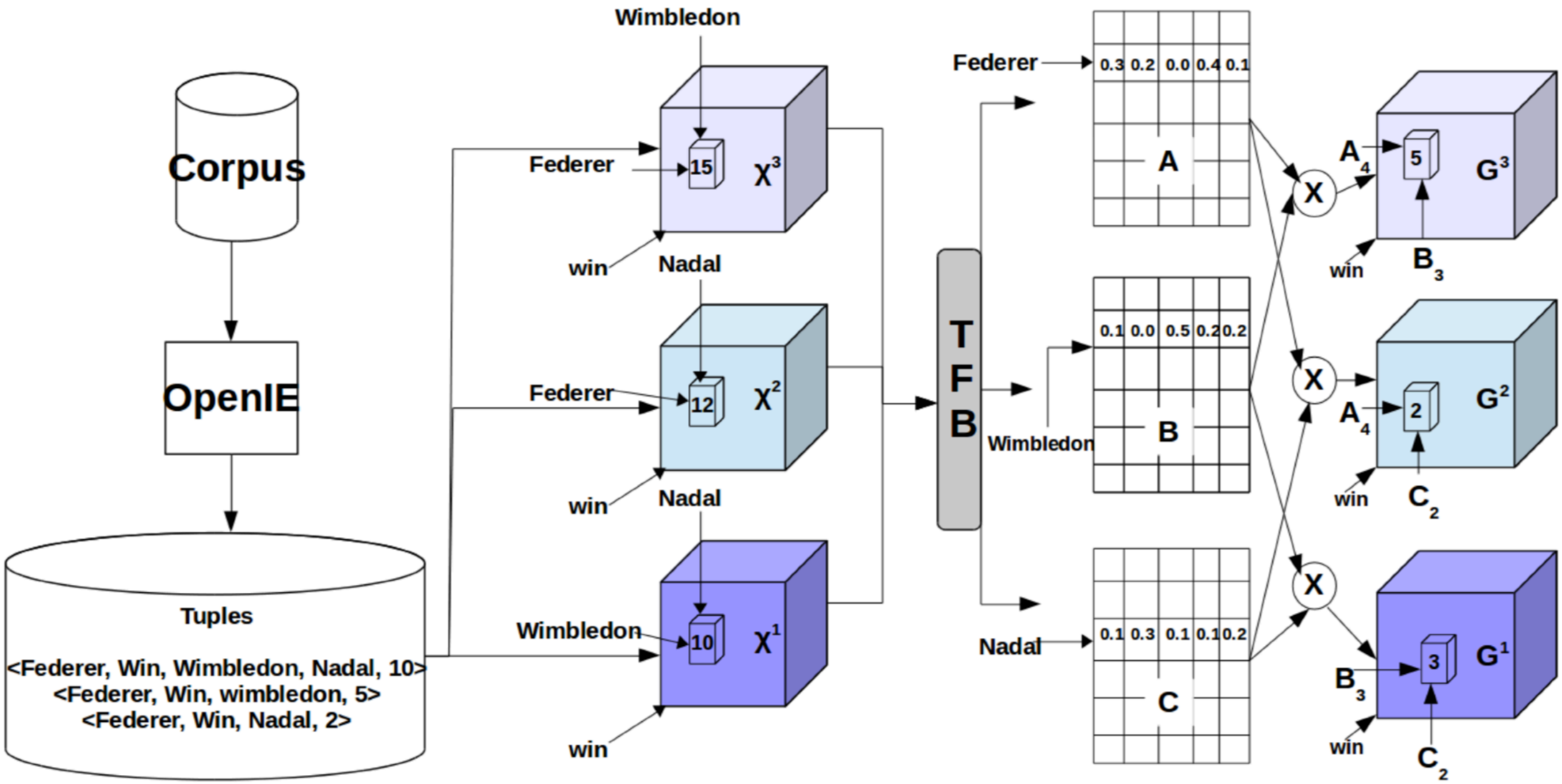


Figure : Tensor Factorization with Back-off

Schema Induction by Constrained Clique Mining

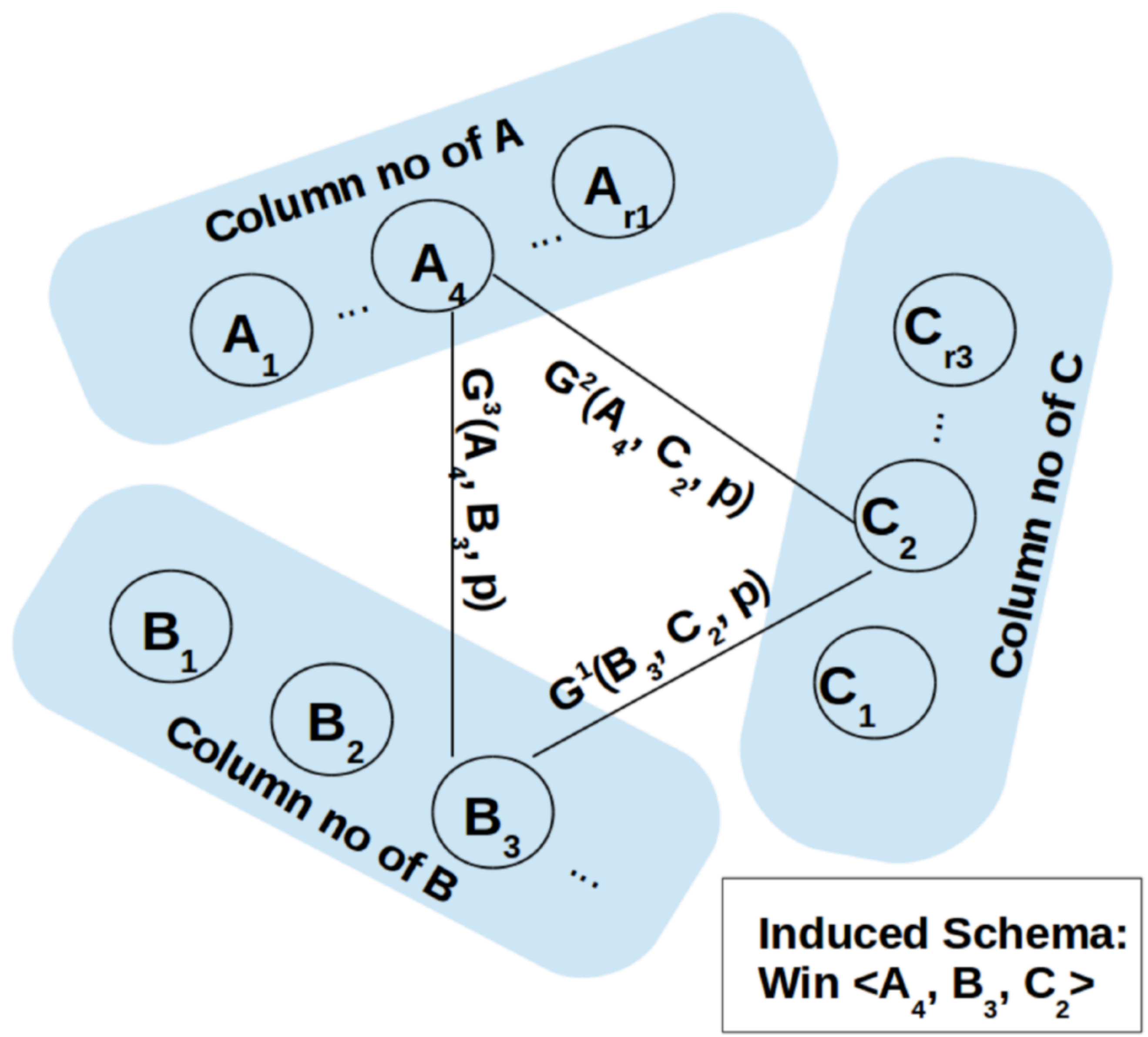


Figure : Schema Induction

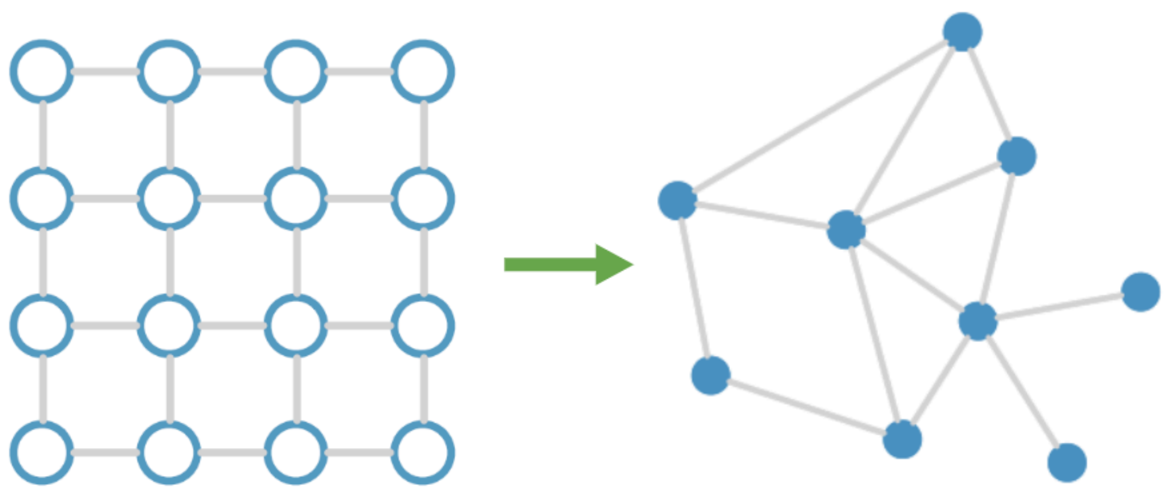
Schemas induced by TFBA

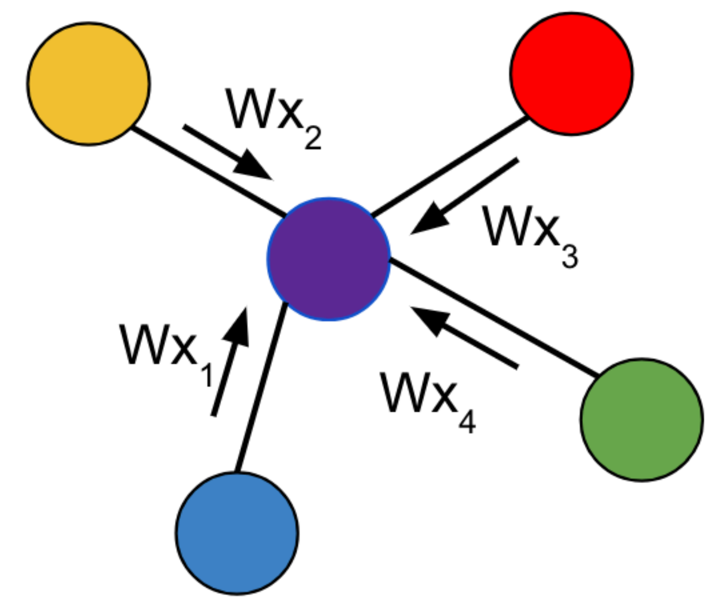
Predicate Schema	NPs from the induced categories	Annotator Judgment	Suggested Label
	Shootings		
left(A_6, B_0, C_7)	A_6: *shooting, shooting incident, double shooting* B_0: *one person, two people, three people* C_7: *dead, injured, on edge*	valid	< shooting > < people > <injured >
identify(A_1, B_1, C_5, C_6)	A_1: *police, officers, huntsville police* B_1: *man, victims, four victims* C_5: *sunday, shooting staurday, wednesday afternoon* C_6: *apartment, bedroom, building in the neighborhood*	valid	< police > < victim(s)> <day/time > <place >
say(A_1, B_1, C_5)	A_1: *police, officers, huntsville police* B_1: *man, victims, four victims* C_5: *sunday, shooting staurday, wednesday afternoon*	invalid	–
	NYT sports		
spend(A_0, B_{16}, C_3)	A_0: *yankees, mets, jets* B_{14}: *\$ <num> million, \$ <num>, \$ <num> billion* C_3: *<num>, year, last season*	valid	< team > < money > < year >
win(A_2, B_{10}, C_3)	A_2: *red sox, team, yankees* B_{10}: *world series, title, world cup* C_3: *<num>, year, last season*	valid	< team > < championship > < year >
get(A_4, B_4, C_1)	A_4: *umpire, mike cameron, andre agassi* B_4: *ball, lives, grounder* C_1: *back, forward, <num>-yard line*	invalid	–

Table : Examples of schemas induced by TFB. Please note that some of them have 3 dimensions and some of them have four.

Graph Convolutional Networks (GCN)

Generalization of CNNs over Graphs.





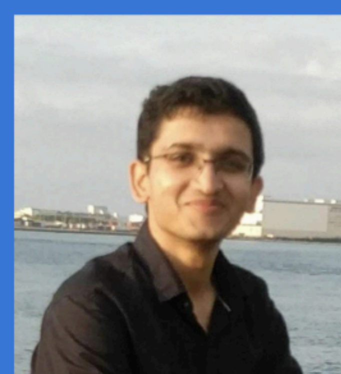
$$h_v = f\left(\sum_{u\in\mathcal{N}(v)} (Wx_u + b)\right), \quad \forall v \in \mathcal{V}$$

EMNLP 2019, Hong Kong

Tutorial Homepage: **github/svjan5/GNNs-for-NLP**

A Tutorial on

Graph Neural Networks for Natural Language Processing



Shikhar Vashishth[1] — shikhar@iisc.ac.in

Y. Naganand[1] — naganand@iisc.ac.in

Partha Talukdar[1,2] — ppt@iisc.ac.in

[1]**Indian Institute of Science, Bangalore** [2]**KENOME**

bit.ly/gnn-nlp-tutorial

RESIDE: Distantly-supervised Relation Extraction with Side Information

Joint work with


Shikhar Vashishth


Rishabh Joshi


Sai Suman


Chiru

EMNLP 2018

https://github.com/malllabiisc/RESIDE

Relation Extraction (RE)

"Obama was elected as the President of USA yesterday."

presidentOf(Obama, USA)

- Two approaches
  - Supervised: sentence annotations, doesn't scale
  - Distantly-supervised: seed-based supervision

Distantly-supervised RE

Multi-instance, Multi-label Learning Problem

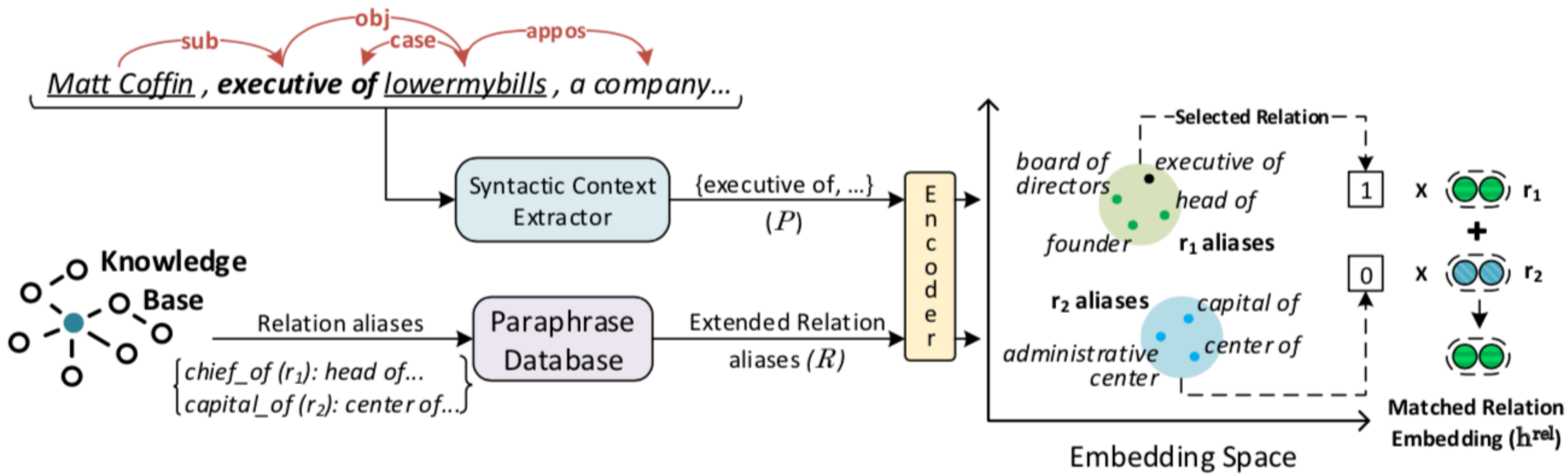
Entity 1 => Narendra Modi
Entity 2 => India

Relation => Prime_Minister_of

Instance Set =>
* Narendra Modi, Prime Minister of India visited USA
* Narendra Modi payed respects at Amar Jawan Jyoti on India's Independence day
*Narendra Modi represented India at recent BRICS Summit
*Narendra Modi, Indian Prime Minister welcomed delegates from UK visiting Delhi

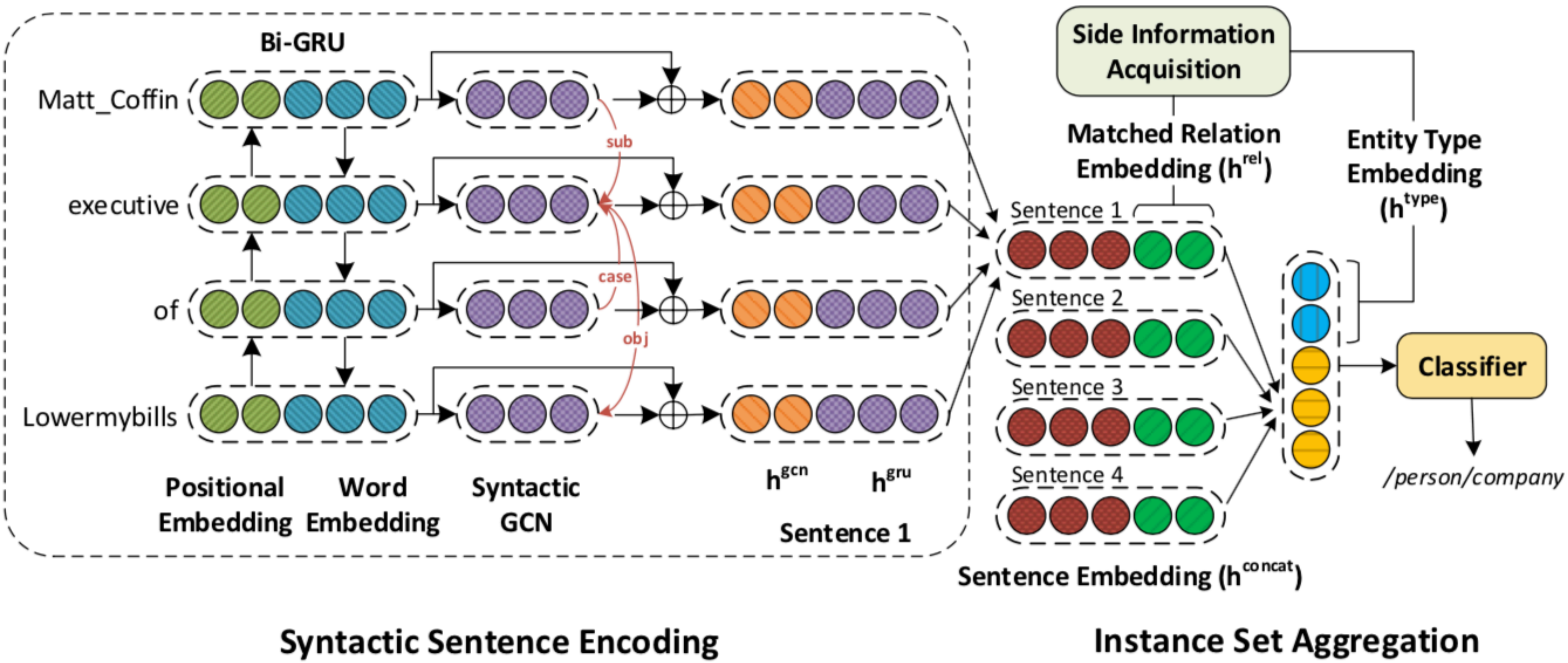
Figure : Example of an Instance Set in Distant Supervision training set

RESIDE: Side Information

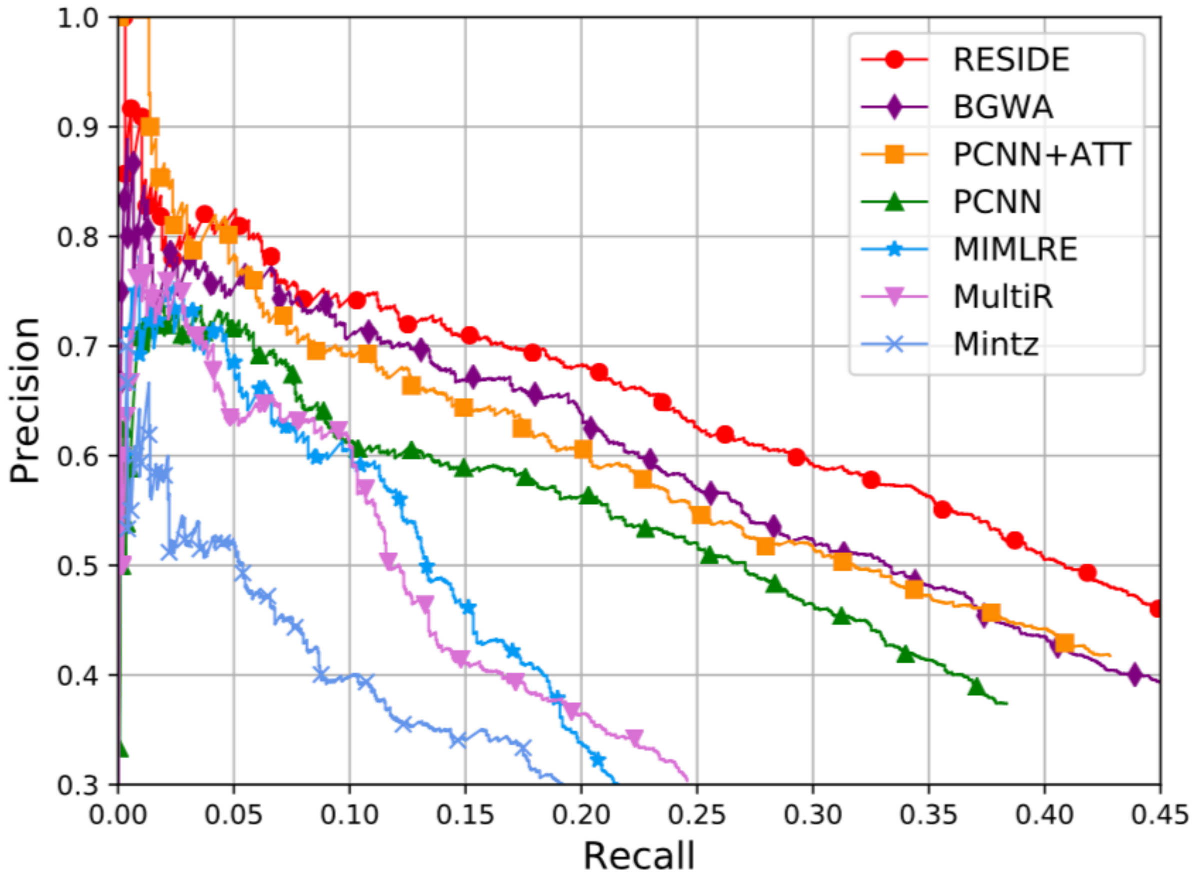


- **Relation alias side information**: Extract relation phrases between target entities and link them to KG based on closeness in embedding space

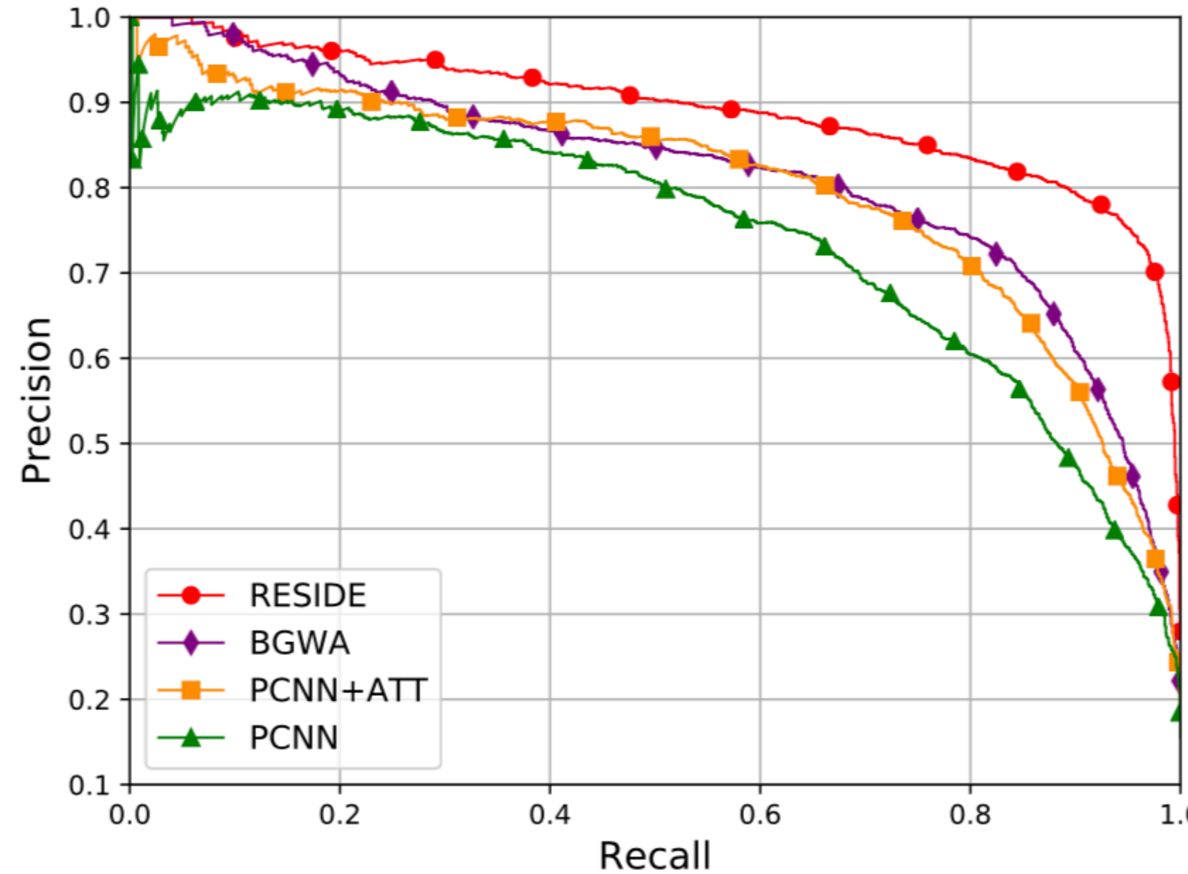
RESIDE: Overview



RESIDE: Results

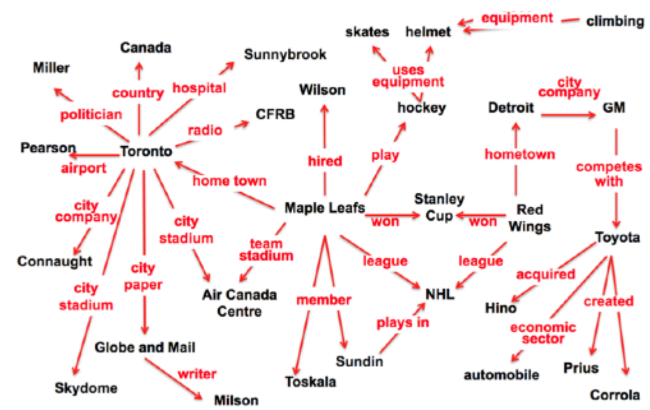


(a) Riedel dataset



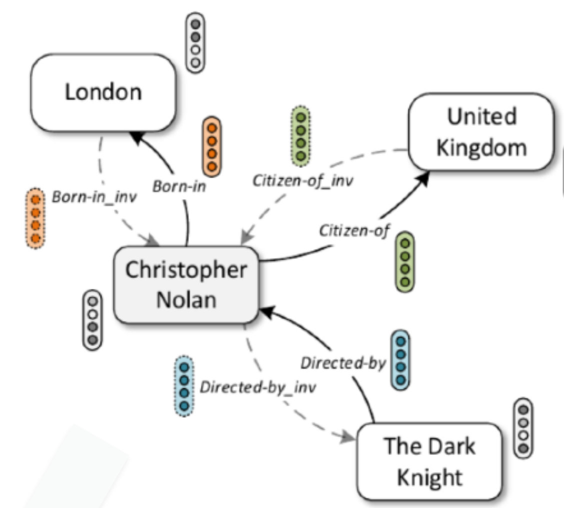
(b) GIDS dataset

Figure : Comparison of Precision-recall curve.



Part 1: Build KG

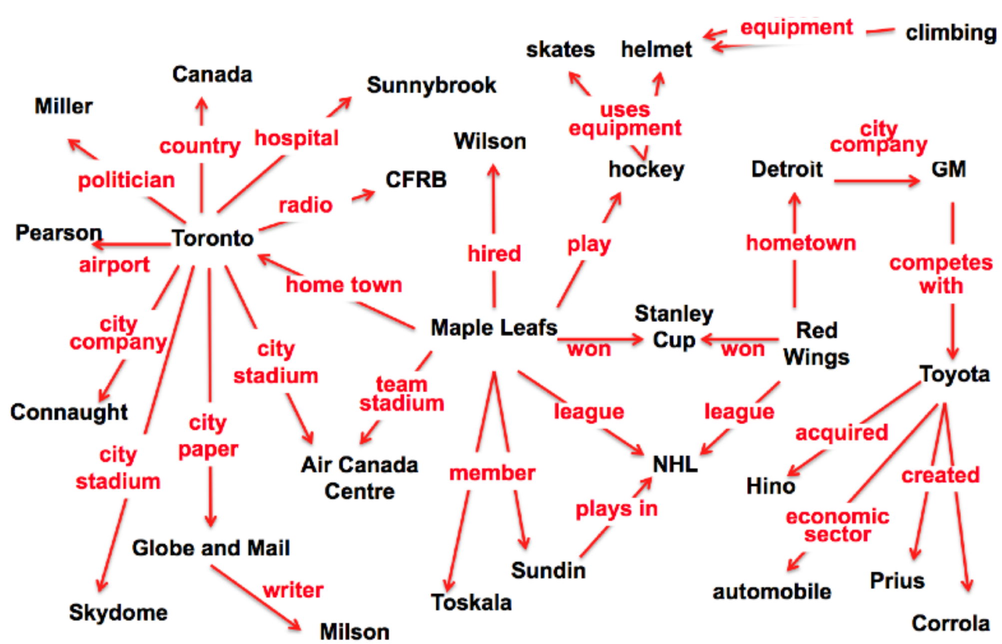
Part 2: Embed KG



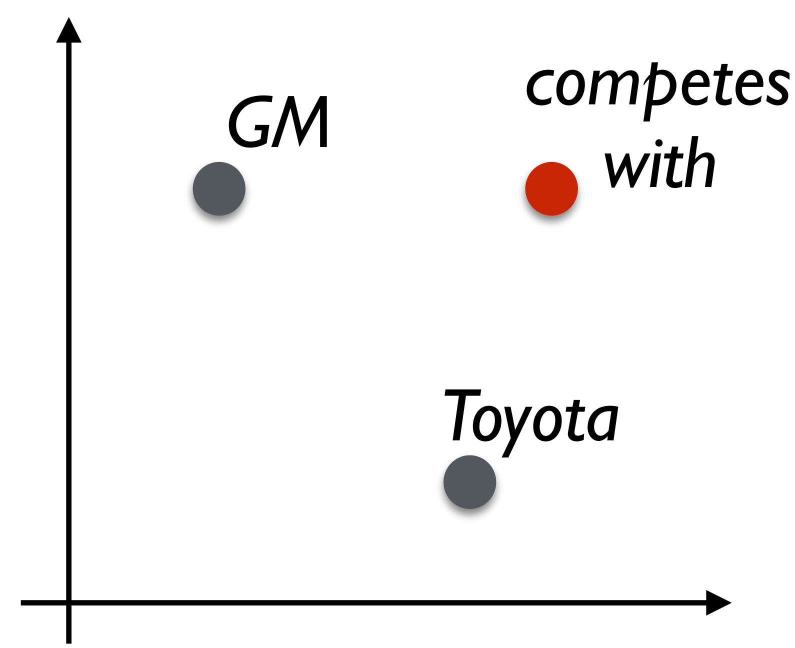
Part 3: Use KG

Read Better

Two Views of Knowledge



Knowledge Graph

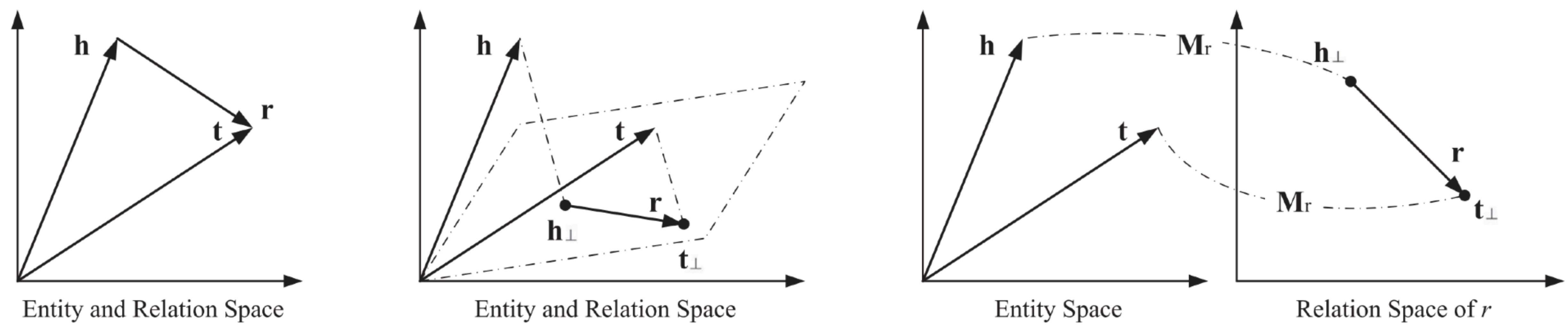


Dense Representations

Knowledge Graph Embedding

[Surveys: Wang et al., TKDE 2017, ThuNLP]

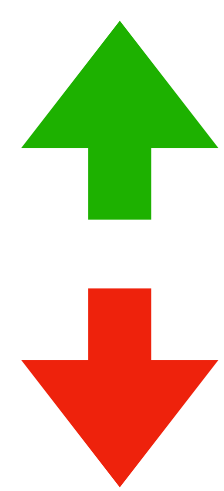
(h, r, t) = (Barack Obama, presidentOf, USA)



(a) TransE.   (b) TransH.   (c) TransR.

$h + r \approx t$

Triple scoring function: $f_r(h, t)$



Positive triples

Negative triples

Knowledge Graph Embedding

[Surveys: Wang et al., TKDE 2017, ThuNLP]

Method	Ent. embedding	Rel. embedding	Scoring function $f_r(h,t)$	Constraints/Regularization
TransE [14]	$\mathbf{h},\mathbf{t}\in\mathbb{R}^d$	$\mathbf{r}\in\mathbb{R}^d$	$-\Vert\mathbf{h}+\mathbf{r}-\mathbf{t}\Vert_{1/2}$	$\Vert\mathbf{h}\Vert_2=1, \Vert\mathbf{t}\Vert_2=1$
TransH [15]	$\mathbf{h},\mathbf{t}\in\mathbb{R}^d$	$\mathbf{r},\mathbf{w}_r\in\mathbb{R}^d$	$-\Vert(\mathbf{h}-\mathbf{w}_r^\top\mathbf{h}\mathbf{w}_r)+\mathbf{r}-(\mathbf{t}-\mathbf{w}_r^\top\mathbf{t}\mathbf{w}_r)\Vert_2^2$	$\Vert\mathbf{h}\Vert_2\le 1, \Vert\mathbf{t}\Vert_2\le 1$
TransR [16]	$\mathbf{h},\mathbf{t}\in\mathbb{R}^d$	$\mathbf{r}\in\mathbb{R}^k, \mathbf{M}_r\in\mathbb{R}^{k\times d}$	$-\Vert\mathbf{M}_r\mathbf{h}+\mathbf{r}-\mathbf{M}_r\mathbf{t}\Vert_2^2$	$\vert\mathbf{w}_r^\top\mathbf{r}\vert/\Vert\mathbf{r}\Vert_2\le\epsilon, \Vert\mathbf{w}_r\Vert_2=1$ $\Vert\mathbf{h}\Vert_2\le 1, \Vert\mathbf{t}\Vert_2\le 1, \Vert\mathbf{r}\Vert_2\le 1$
TransD [50]	$\mathbf{h},\mathbf{w}_h\in\mathbb{R}^d$ $\mathbf{t},\mathbf{w}_t\in\mathbb{R}^d$	$\mathbf{r},\mathbf{w}_r\in\mathbb{R}^k$	$-\Vert(\mathbf{w}_r\mathbf{w}_h^\top+\mathbf{I})\mathbf{h}+\mathbf{r}-(\mathbf{w}_r\mathbf{w}_t^\top+\mathbf{I})\mathbf{t}\Vert_2^2$	$\Vert\mathbf{M}_r\mathbf{h}\Vert_2\le 1, \Vert\mathbf{M}_r\mathbf{t}\Vert_2\le 1$ $\Vert\mathbf{h}\Vert_2\le 1, \Vert\mathbf{t}\Vert_2\le 1, \Vert\mathbf{r}\Vert_2\le 1$ $\Vert(\mathbf{w}_r\mathbf{w}_h^\top+\mathbf{I})\mathbf{h}\Vert_2\le 1$ $\Vert(\mathbf{w}_r\mathbf{w}_t^\top+\mathbf{I})\mathbf{t}\Vert_2\le 1$
TranSparse [51]	$\mathbf{h},\mathbf{t}\in\mathbb{R}^d$	$\mathbf{r}\in\mathbb{R}^k, \mathbf{M}_r(\theta_r)\in\mathbb{R}^{k\times d}$ $\mathbf{M}_r^1(\theta_r^1), \mathbf{M}_r^2(\theta_r^2)\in\mathbb{R}^{k\times d}$	$-\Vert\mathbf{M}_r(\theta_r)\mathbf{h}+\mathbf{r}-\mathbf{M}_r(\theta_r)\mathbf{t}\Vert_{1/2}^2$ $-\Vert\mathbf{M}_r^1(\theta_r^1)\mathbf{h}+\mathbf{r}-\mathbf{M}_r^2(\theta_r^2)\mathbf{t}\Vert_{1/2}^2$	$\Vert\mathbf{h}\Vert_2\le 1, \Vert\mathbf{t}\Vert_2\le 1, \Vert\mathbf{r}\Vert_2\le 1$ $\Vert\mathbf{M}_r(\theta_r)\mathbf{h}\Vert_2\le 1, \Vert\mathbf{M}_r(\theta_r)\mathbf{t}\Vert_2\le 1$ $\Vert\mathbf{M}_r^1(\theta_r^1)\mathbf{h}\Vert_2\le 1, \Vert\mathbf{M}_r^2(\theta_r^2)\mathbf{t}\Vert_2\le 1$
TransM [52]	$\mathbf{h},\mathbf{t}\in\mathbb{R}^d$	$\mathbf{r}\in\mathbb{R}^d$	$-\theta_r\Vert\mathbf{h}+\mathbf{r}-\mathbf{t}\Vert_{1/2}$	$\Vert\mathbf{h}\Vert_2=1, \Vert\mathbf{t}\Vert_2=1$
ManifoldE [53]	$\mathbf{h},\mathbf{t}\in\mathbb{R}^d$	$\mathbf{r}\in\mathbb{R}^d$	$-(\Vert\mathbf{h}+\mathbf{r}-\mathbf{t}\Vert_2^2-\theta_r^2)^2$	$\Vert\mathbf{h}\Vert_2\le 1, \Vert\mathbf{t}\Vert_2\le 1, \Vert\mathbf{r}\Vert_2\le 1$
TransF [54]	$\mathbf{h},\mathbf{t}\in\mathbb{R}^d$	$\mathbf{r}\in\mathbb{R}^d$	$(\mathbf{h}+\mathbf{r})^\top\mathbf{t}+(\mathbf{t}-\mathbf{r})^\top\mathbf{h}$	$\Vert\mathbf{h}\Vert_2\le 1, \Vert\mathbf{t}\Vert_2\le 1, \Vert\mathbf{r}\Vert_2\le 1$
TransA [55]	$\mathbf{h},\mathbf{t}\in\mathbb{R}^d$	$\mathbf{r}\in\mathbb{R}^d, \mathbf{M}_r\in\mathbb{R}^{d\times d}$	$-(\vert\mathbf{h}+\mathbf{r}-\mathbf{t}\vert)^\top\mathbf{M}_r(\vert\mathbf{h}+\mathbf{r}-\mathbf{t}\vert)$	$\Vert\mathbf{h}\Vert_2\le 1, \Vert\mathbf{t}\Vert_2\le 1, \Vert\mathbf{r}\Vert_2\le 1$
KG2E [45]	$\mathbf{h}\sim\mathcal{N}(\boldsymbol{\mu}_h,\Sigma_h)$ $\mathbf{t}\sim\mathcal{N}(\boldsymbol{\mu}_t,\Sigma_t)$ $\boldsymbol{\mu}_h,\boldsymbol{\mu}_t\in\mathbb{R}^d$ $\Sigma_h,\Sigma_t\in\mathbb{R}^{d\times d}$	$\mathbf{r}\sim\mathcal{N}(\boldsymbol{\mu}_r,\Sigma_r)$ $\boldsymbol{\mu}_r\in\mathbb{R}^d, \Sigma_r\in\mathbb{R}^{d\times d}$	$-\mathrm{tr}(\Sigma_r^{-1}(\Sigma_h+\Sigma_t))-\boldsymbol{\mu}^\top\Sigma_r^{-1}\boldsymbol{\mu}-\ln\frac{\det(\Sigma_r)}{\det(\Sigma_h+\Sigma_t)}$ $-\boldsymbol{\mu}^\top\Sigma^{-1}\boldsymbol{\mu}-\ln(\det(\Sigma))$ $\boldsymbol{\mu}=\boldsymbol{\mu}_h+\boldsymbol{\mu}_r-\boldsymbol{\mu}_t$ $\Sigma=\Sigma_h+\Sigma_r+\Sigma_t$	$\Vert\mathbf{M}_r\Vert_F\le 1, [\mathbf{M}_r]_{ij}=[\mathbf{M}_r]_{ji}\ge 0$ $\Vert\boldsymbol{\mu}_h\Vert_2\le 1, \Vert\boldsymbol{\mu}_t\Vert_2\le 1, \Vert\boldsymbol{\mu}_r\Vert_2\le 1$ $c_{min}\mathbf{I}\le\Sigma_h\le c_{max}\mathbf{I}$ $c_{min}\mathbf{I}\le\Sigma_t\le c_{max}\mathbf{I}$ $c_{min}\mathbf{I}\le\Sigma_r\le c_{max}\mathbf{I}$
TransG [46]	$\mathbf{h}\sim\mathcal{N}(\boldsymbol{\mu}_h,\sigma_h^2\mathbf{I})$ $\mathbf{t}\sim\mathcal{N}(\boldsymbol{\mu}_t,\sigma_t^2\mathbf{I})$ $\boldsymbol{\mu}_h,\boldsymbol{\mu}_t\in\mathbb{R}^d$	$\boldsymbol{\mu}_r^i\sim\mathcal{N}(\boldsymbol{\mu}_t-\boldsymbol{\mu}_h,(\sigma_h^2+\sigma_t^2)\mathbf{I})$ $\mathbf{r}=\sum_i\pi_r^i\boldsymbol{\mu}_r^i\in\mathbb{R}^d$	$\sum_i\pi_r^i\exp\left(-\frac{\Vert\boldsymbol{\mu}_h+\boldsymbol{\mu}_r^i-\boldsymbol{\mu}_t\Vert_2^2}{\sigma_h^2+\sigma_t^2}\right)$	$\Vert\boldsymbol{\mu}_h\Vert_2\le 1, \Vert\boldsymbol{\mu}_t\Vert_2\le 1, \Vert\boldsymbol{\mu}_r^i\Vert_2\le 1$
UM [56]	$\mathbf{h},\mathbf{t}\in\mathbb{R}^d$	—	$-\Vert\mathbf{h}-\mathbf{t}\Vert_2^2$	$\Vert\mathbf{h}\Vert_2=1, \Vert\mathbf{t}\Vert_2=1$
SE [57]	$\mathbf{h},\mathbf{t}\in\mathbb{R}^d$	$\mathbf{M}_r^1,\mathbf{M}_r^2\in\mathbb{R}^{d\times d}$	$-\Vert\mathbf{M}_r^1\mathbf{h}-\mathbf{M}_r^2\mathbf{t}\Vert_1$	$\Vert\mathbf{h}\Vert_2=1, \Vert\mathbf{t}\Vert_2=1$

CESI: Canonicalization

[WWW 2018]


Shikhar


Prince

Barack Obama, Mr. Obama, George Bush, Mumbai, Bombay, Madrid

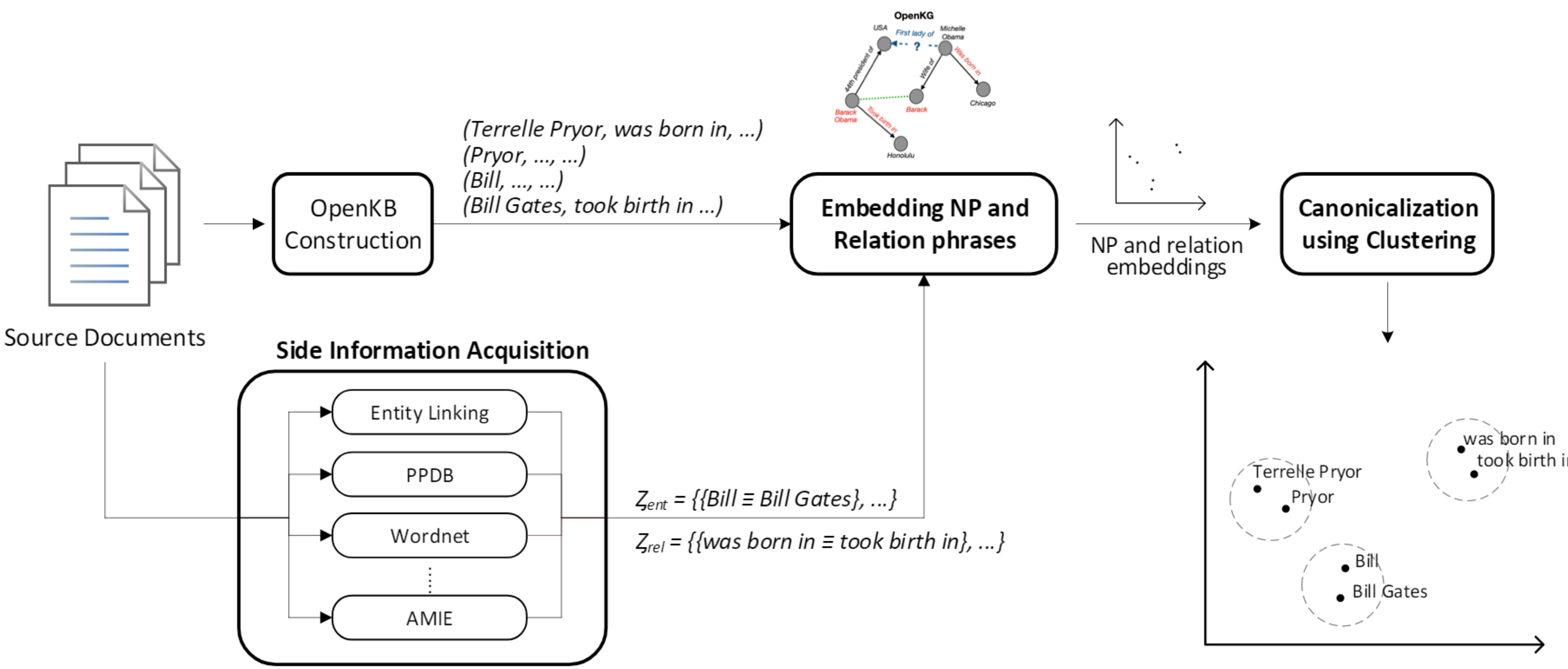
Barack Obama
Mr. Obama

George Bush

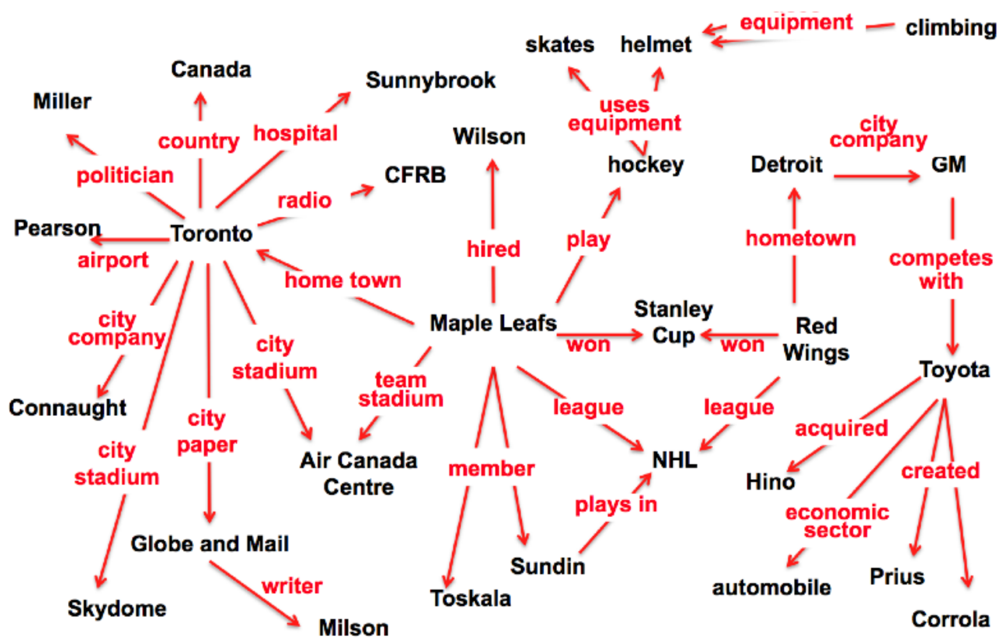
Madrid

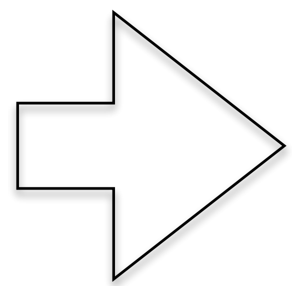
Mumbai
Bombay

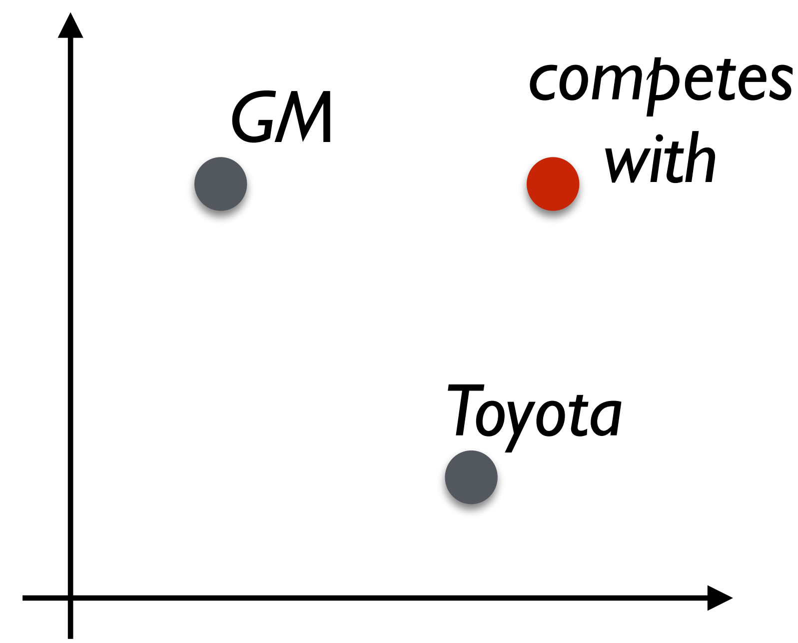
CESI Architecture



KG Embedding



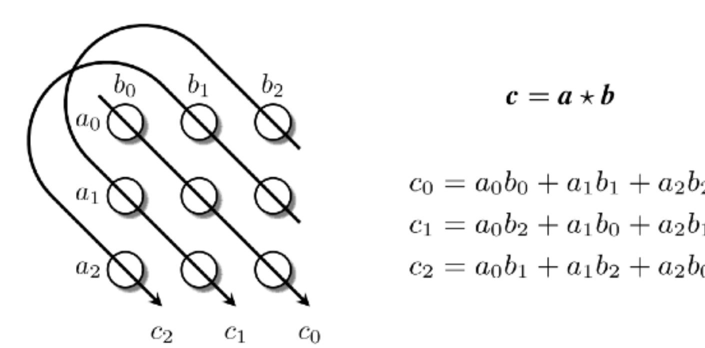




Many methods in literature, we use Holographic Embedding (HolE)

HolE assigns a score η to each triple $< v, r, v' > \in KB$

$$\eta = e_r^T(e_v \star e_{v'})$$



(a) Circular Correlation

Side Information

- Entity Linking
  - Identify entity mention and link to KBs like Wikipedia
  - *US -> United_States, America -> United_States*
- PPDB (Paraphrase database)
  - Large collection of paraphrases in English
  - *management* ≡ *administration, head of* ≡ *chief of*
- WordNet with Word-sense disambiguation
  - Identify synsets for NPs
  - *picture* and *image* can be identified as equivalent
- IDF Token Overlap
  - NPs sharing infrequent terms give strong indication of equivalence
  - *Warren Buffett* and *Mr. Buffett* refers to the same person

CESI Optimization Objective

$$\min_{\Theta} \quad \lambda_{str} \boxed{\sum_{i \in D_+} \sum_{j \in D_-} \max(0, \gamma + \sigma(\eta_j) - \sigma(\eta_i))}$$

KG Embedding

$$+ \boxed{\sum_{\theta \in \mathscr{C}_{\text{ent}}} \frac{\lambda_{\text{ent},\theta}}{|\mathcal{Z}_{\text{ent},\theta}|} \sum_{v,v' \in \mathcal{Z}_{\text{ent},\theta}} \|e_v - e_{v'}\|^2}$$

NP Side Info

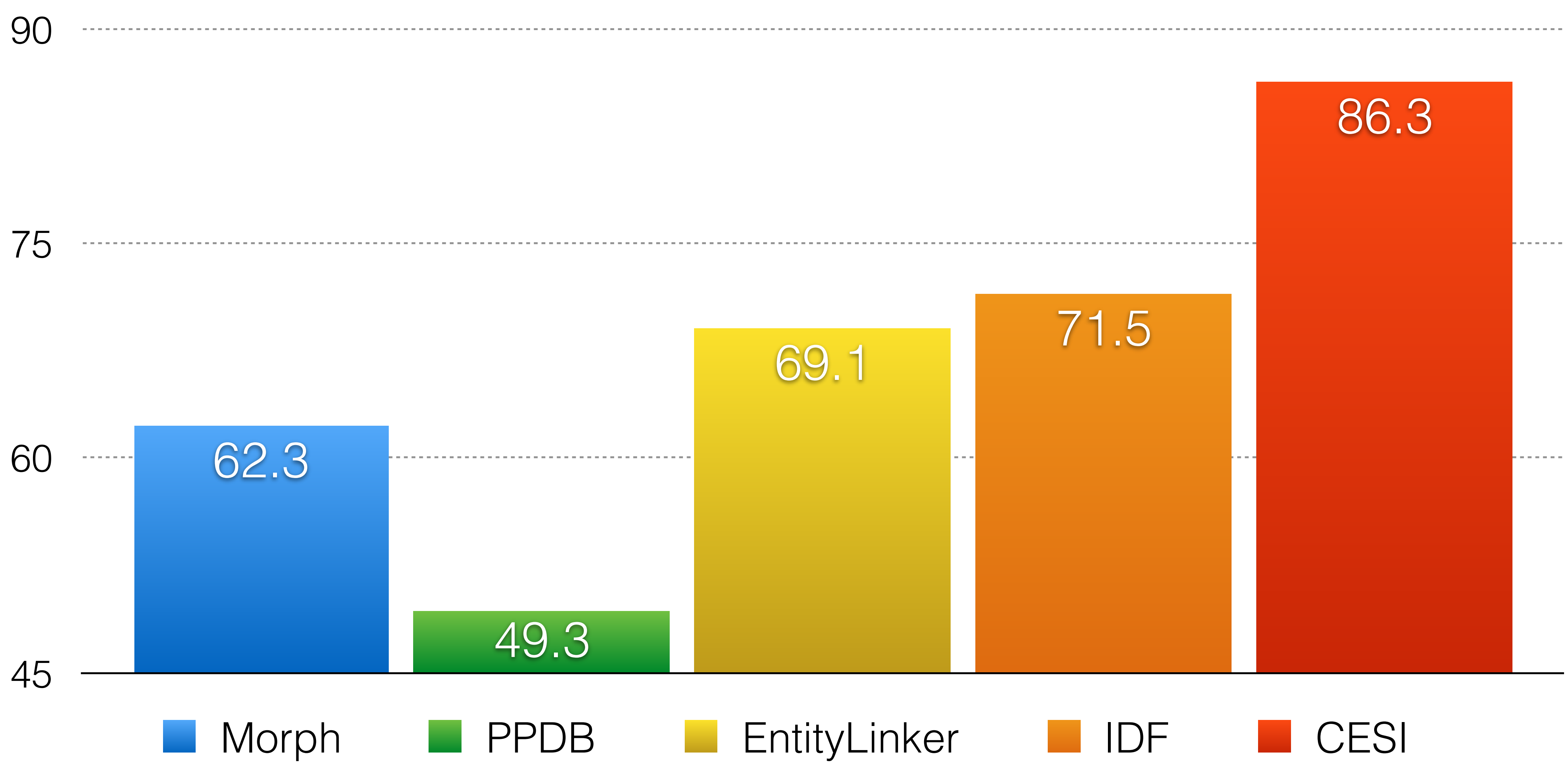
$$+ \boxed{\sum_{\phi \in \mathscr{C}_{\text{rel}}} \frac{\lambda_{\text{rel},\phi}}{|\mathcal{Z}_{\text{rel},\phi}|} \sum_{u,u' \in \mathcal{Z}_{\text{rel},\phi}} \|r_u - r_{u'}\|^2}$$

Relation Side Info

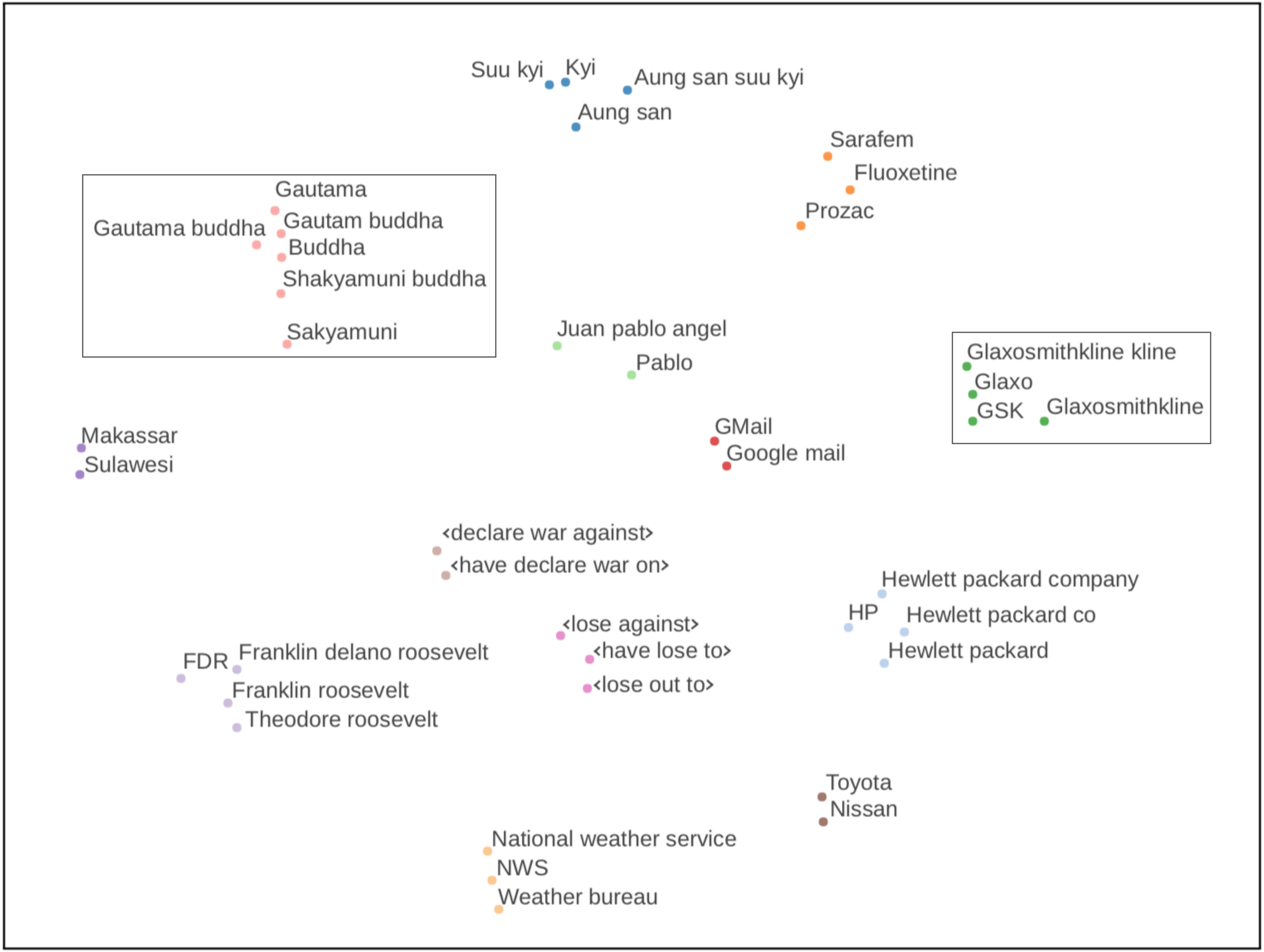
$$+ \boxed{\lambda_{\text{reg}} \left(\sum_{v \in V} \|e_v\|^2 + \sum_{r \in R} \|e_r\|^2 \right)}.$$

Regularize

Results: NP Canonicalization



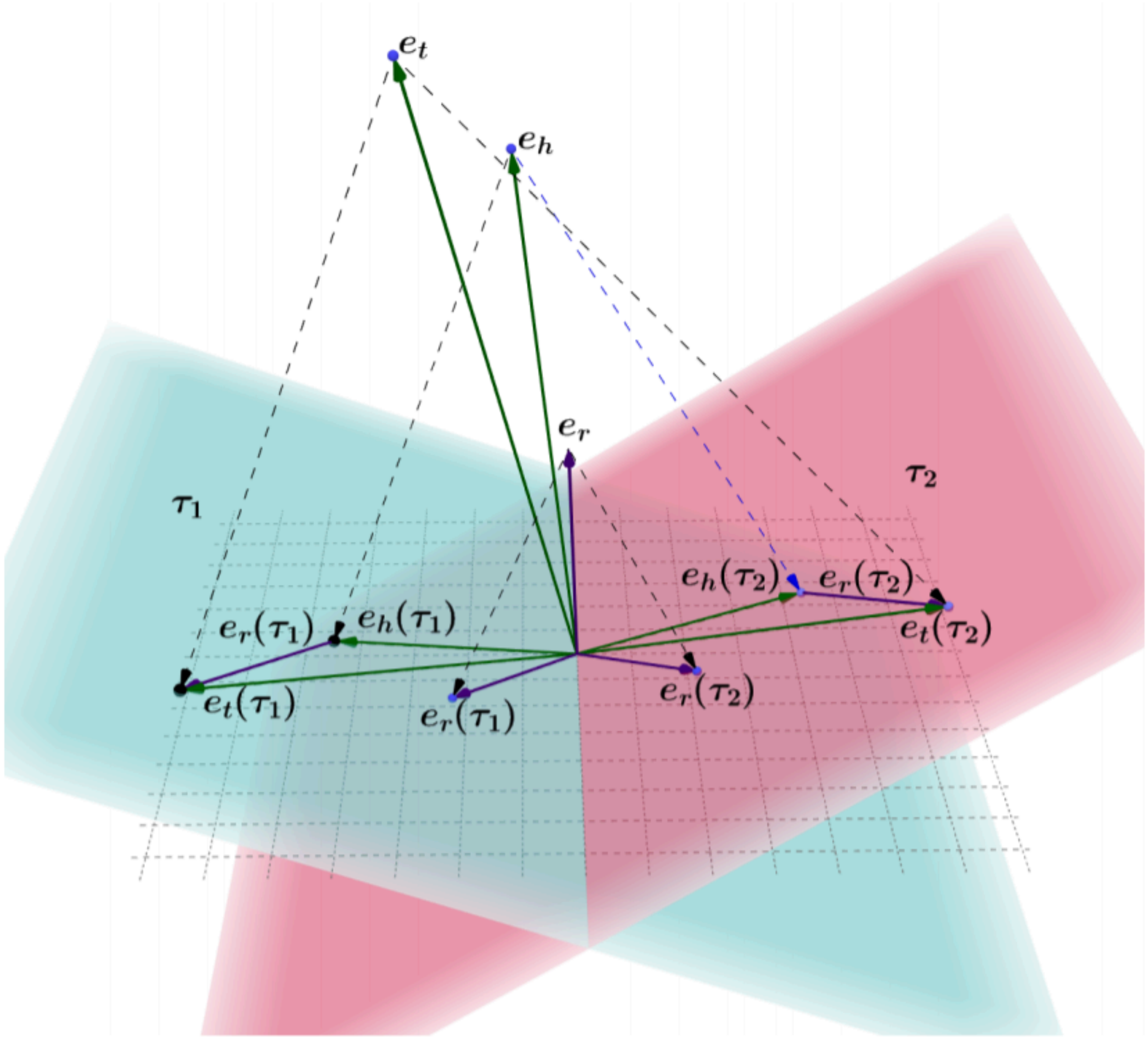
Averaged over three datasets


Suu kyi
Kyi
Aung san suu kyi
Aung san
Sarafem
Fluoxetine
Prozac
Gautama
Gautama buddha
Gautam buddha
Buddha
Shakyamuni buddha
Sakyamuni
Juan pablo angel
Pablo
Glaxosmithkline kline
Glaxo
GSK
Glaxosmithkline
Makassar
Sulawesi
GMail
Google mail
<declare war against>
<have declare war on>
Hewlett packard company
HP
Hewlett packard co
Hewlett packard
<lose against>
<have lose to>
<lose out to>
FDR
Franklin delano roosevelt
Franklin roosevelt
Theodore roosevelt
Toyota
Nissan
National weather service
NWS
Weather bureau

HyTE [EMNLP 2018]

Shib

Swayambhu



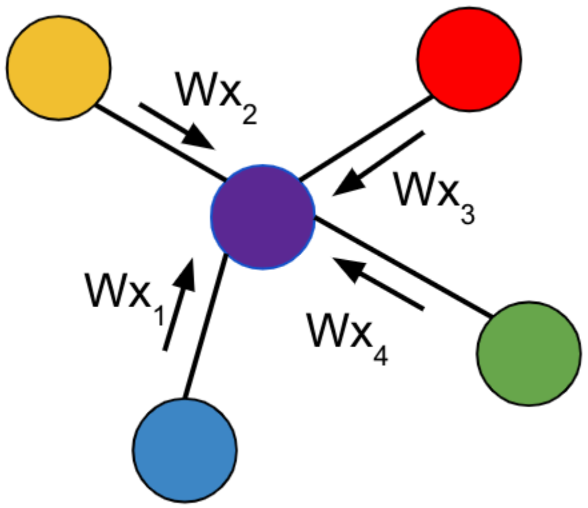
Incorporating **time** into KG embedding

Figure 1: In the figure, the vectors e_h, e_r and e_t correspond to the triple (h, r, t) that is valid at time τ_1 and τ_2. $e_h(\tau_1), e_r(\tau_1)$ and $e_t(\tau_1)$ are the projections of this triple on the hyperplane corresponding to time τ_1 (similarly for time τ_2). Our method HyTE minimizes the translational distance, $\sum_i \|e_t(\tau_i) + e_r(\tau_i) - e_t(\tau_i)\|_1$ in order to learn the temporal representations of entities and relations this triple.

HyTE: Results

Dataset	YAGO11K				Wikidata12K			
Metric	Mean Rank		Hits@10(%)		Mean Rank		Hits@10(%)	
	tail	head	tail	head	tail	head	tail	head
Trans-E	504	2020	4.4	1.2	520	740	11.0	6.0
TransH(Wang et al., 2014)	354	1808	5.8	1.5	423	648	23.7	11.8
HolE(Nickel et al., 2016b)	1828	1953	29.4	13.7	734	808	25.0	12.3
t-TransE (Jiang et al., 2016)	292	1692	6.2	1.3	283	413	24.5	14.5
HyTE	**107**	**1069**	**38.4**	**16.0**	**179**	**237**	**41.6**	**25.0**

GCN Limitations for KG



GCN First-order approximation

(Kipf et. al. 2016)

$$h_v = f\left(\frac{1}{|\mathcal{N}(v)|}\sum_{u\in\mathcal{N}(v)} W x_u + b\right), \quad \forall v \in \mathcal{V}.$$

Parameter explosion

$$\boldsymbol{h}_v = f\left(\sum_{(u,r)\in\mathcal{N}(v)} \boldsymbol{W}_r \boldsymbol{h}_u\right),$$

Directed-GCN $$h_v^{k+1} = \sigma\left(\sum_{u\in\mathcal{N}(v)} W^k_{dir(u,v)} h^k_u + b^k_{r(u,v)}\right)$$

Relational-GCN $$h_v^{k+1} = \sigma\left(\sum_{u\in\mathcal{N}(v)} W^k_{r(u,v)} h^k_u + b^k_{r(u,v)}\right)$$

GCN Limitations for KG

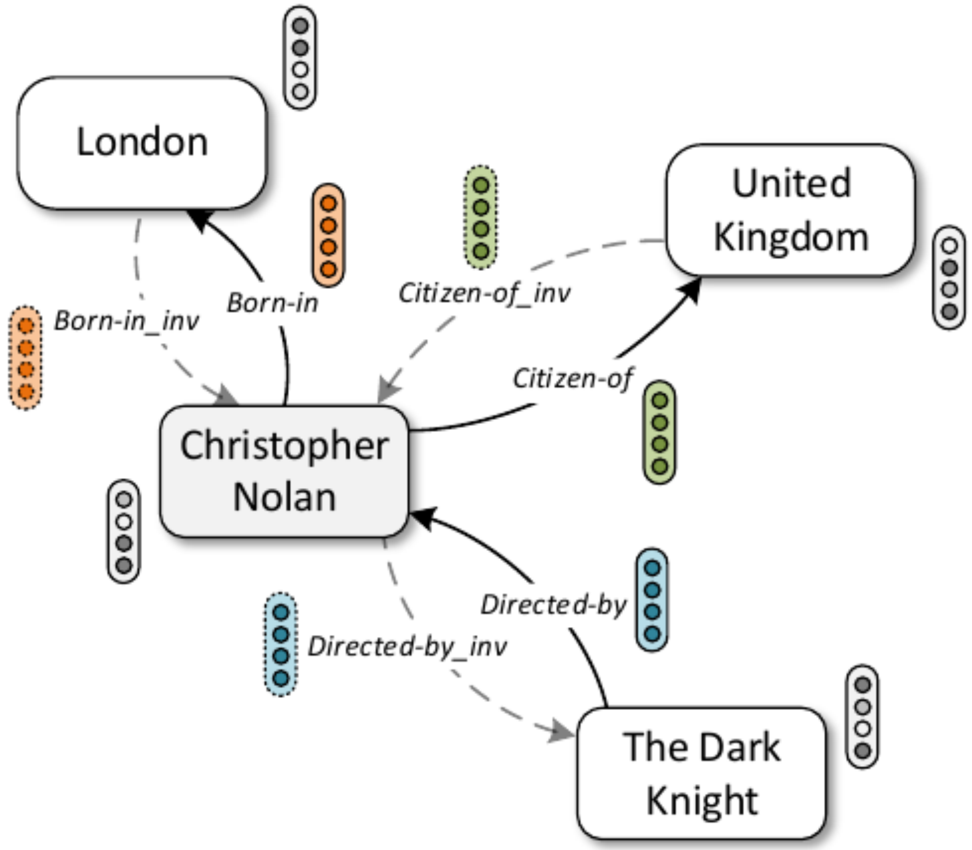
Methods	Node Embeddings	Directions	Relations	Relation Embeddings	Number of Parameters
GCN Kipf & Welling (2016)	✓				$\mathcal{O}(Kd^2)$
Directed-GCN Marcheggiani & Titov (2017)	✓	✓			$\mathcal{O}(Kd^2)$
Weighted-GCN Shang et al. (2019)	✓		✓		$\mathcal{O}(Kd^2 + K\lvert\mathcal{R}\rvert)$
Relational-GCN Schlichtkrull et al. (2017)	✓	✓	✓		$\mathcal{O}(\mathcal{B}Kd^2 + \mathcal{B}K\lvert\mathcal{R}\rvert)$
COMPGCN (Proposed Method)	✓	✓	✓	✓	$\mathcal{O}(Kd^2 + \mathcal{B}d + \mathcal{B}\lvert\mathcal{R}\rvert)$

K: number of GCN layers, d: embedding dimension, B: number of bases, |R|: number of relations

- Directed-GCN: only considers direction of the relations
- Weighted-GCN: learns a scalar weight for each relation
- Relational-GCN: assigns a matrix for each relation
- CompGCN: learns a relation vector embedding for each relation

CompGCN

GCN for Multi-relational Graphs [ICLR 2020]



Relational Graph with Embeddings

$$\boldsymbol{h}_v^{k+1} = f\left(\sum_{(u,r)\in\mathcal{N}(v)} \boldsymbol{W}_{g(r)}^k \phi(\boldsymbol{h}_u^k, \boldsymbol{h}_r^k)\right) \qquad \boldsymbol{W}_{g(r)} = \begin{cases} \boldsymbol{W}_O, & r \in \mathcal{R} \\ \boldsymbol{W}_I, & r \in \mathcal{R}_{inv} \\ \boldsymbol{W}_S, & r = \top \end{cases}$$

$$\boldsymbol{h}_r^{k+1} = \boldsymbol{W}_{rel}^k \, \boldsymbol{h}_r^k$$

CompGCN Results

	FB15k-237					WN18RR				
	MRR	MR	H@10	H@3	H@1	MRR	MR	H@10	H@3	H@1
TransE (Bordes et al., 2013)	.294	357	.465	-	-	.226	3384	.501	-	-
DistMult (Yang et al., 2014)	.241	254	.419	.263	.155	.43	5110	.49	.44	.39
ComplEx (Trouillon et al., 2016)	.247	339	.428	.275	.158	.44	5261	.51	.46	.41
R-GCN (Schlichtkrull et al., 2017)	.248	-	.417		.151	-	-	-		-
KBGAN (Cai & Wang, 2018)	.278	-	.458		-	.214	-	.472	-	-
ConvE (Dettmers et al., 2018)	.325	244	.501	.356	.237	.43	4187	.52	.44	.40
ConvKB (Nguyen et al., 2018)	.243	311	.421	.371	.155	.249	3324	.524	.417	.057
SACN (Shang et al., 2019)	.35	-	.54	.39	.26	.47	-	.54	.48	.43
HypER (Balažević et al., 2019)	.341	250	.520	.376	.252	.465	5798	.522	.477	.436
RotatE (Sun et al., 2019)	.338	**177**	.533	.375	.241	.476	**3340**	**.571**	.492	.428
ConvR (Jiang et al., 2019)	.350	-	.528	.385	.261	.475	-	.537	.489	**.443**
VR-GCN (Ye et al., 2019)	.248	-	.432	.272	.159	-	-	-	-	-
CompGCN (Proposed Method)	**.355**	197	**.535**	**.390**	**.264**	**.479**	3533	.546	**.494**	**.443**

Link Prediction

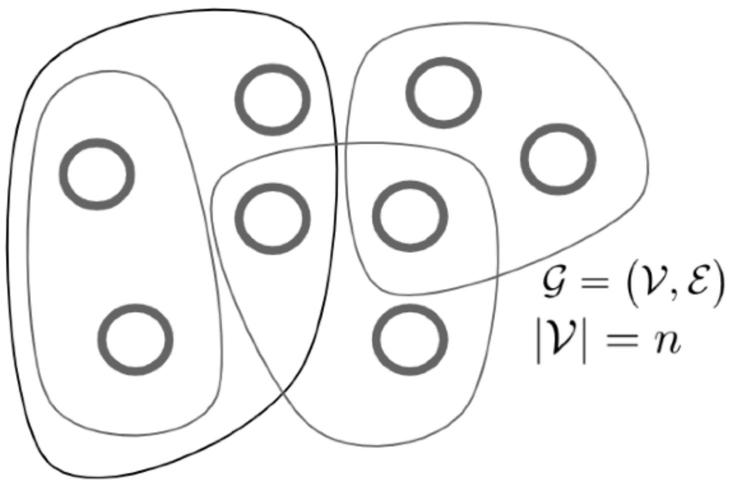
Performance

	MUTAG (Node)	AM
Feat*	77.9	66.7
WL*	80.9	87.4
RDF2Vec*	67.2	88.3
R-GCN*	73.2	89.3
SynGCN	74.8 ± 5.5	86.2 ± 1.9
WGCN	77.9 ± 3.2	90.2 ± 0.9
CompGCN	$\mathbf{85.3 \pm 1.2}$	$\mathbf{90.6 \pm 0.2}$

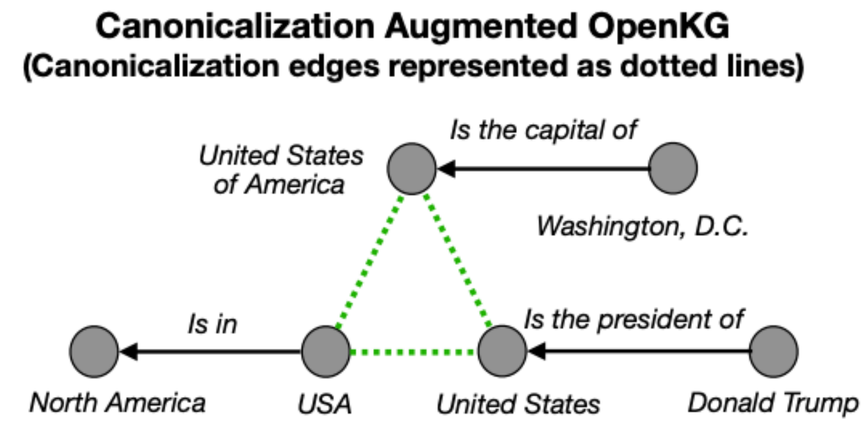
Node classification

Performance

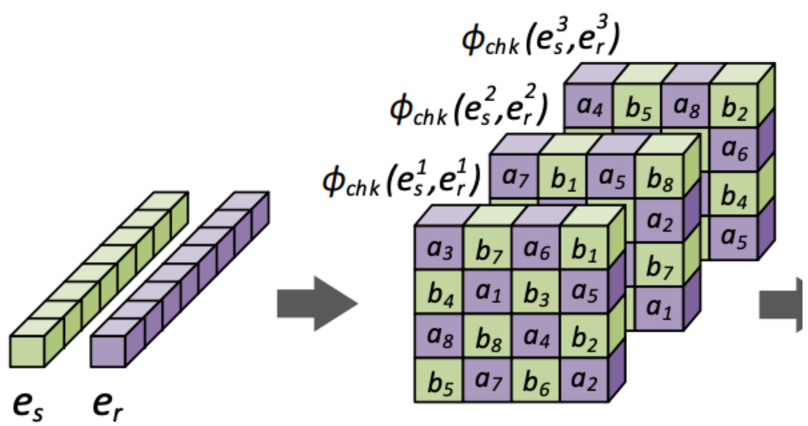
Other Related Works



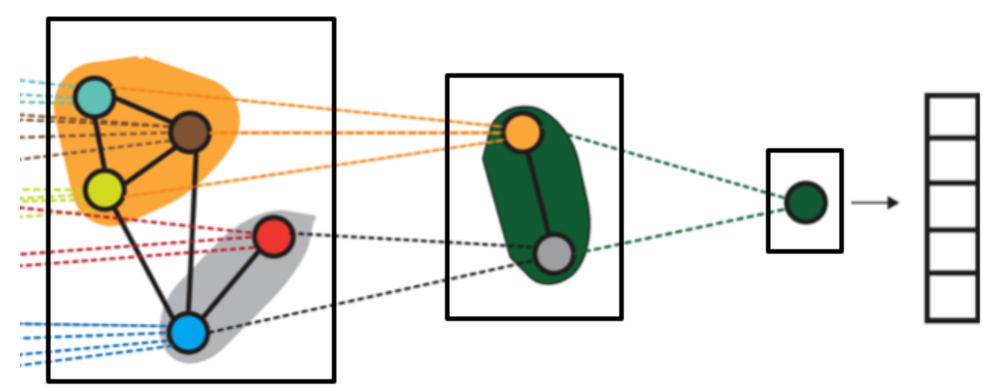
HyperGCN: GCN for HyperGraphs
(NeurIPS 19)



CaRE: Open KG Embedding
(EMNLP 19)



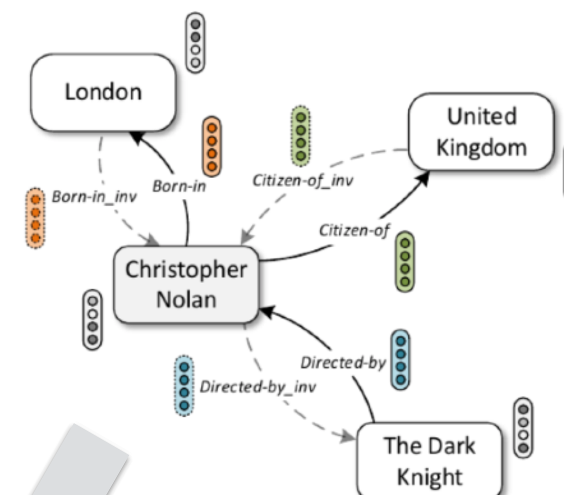
InteractE: More Feature Interaction
(AAAI 2020)



ASAP: Graph Pooling
(AAAI 2020)

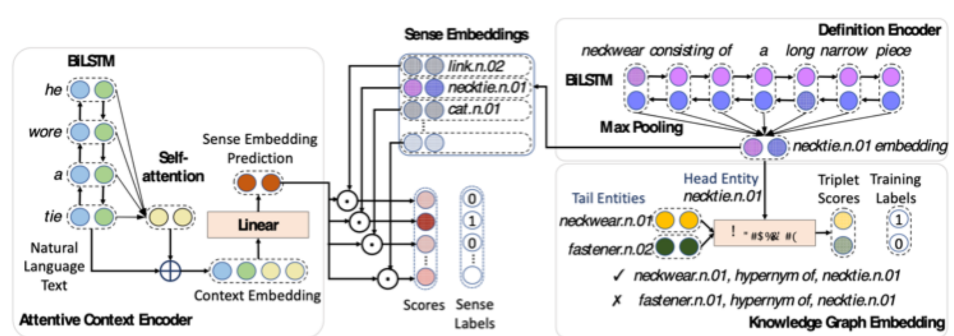
Part 1: Build KG

Part 2: Embed KG



Part 3: Use KG

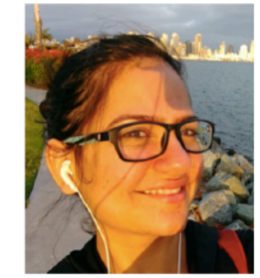
Read Better



EWISE: Zero-shot WSD + KG

[ACL 2019 Outstanding Paper Award]


Sawan Kumar[1]


Sharmistha Jat[1]


Karan Saxena[2]

"tie"

Word

Test sentence

"he wore a vest and tie"

WSD

tie_1 (necktie)

Sense

tie_1 (necktie) neckwear of a narrow …

tie_2 (equality) the finish of a contest …

tie_3 (connect) make a bond or …

Candidate senses

Unseen word

Seen word

Seen sense

Unseen sense

Previous WSD Approaches

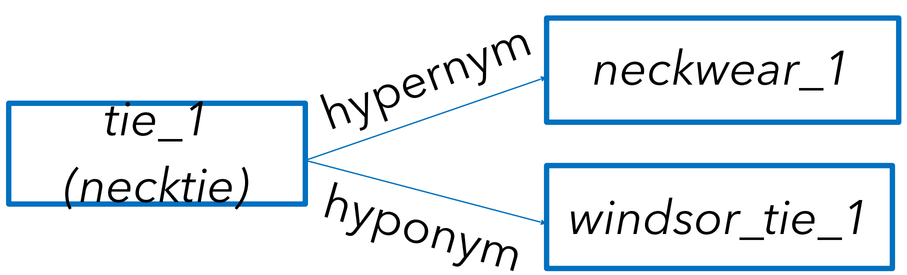
Sense Definition Based

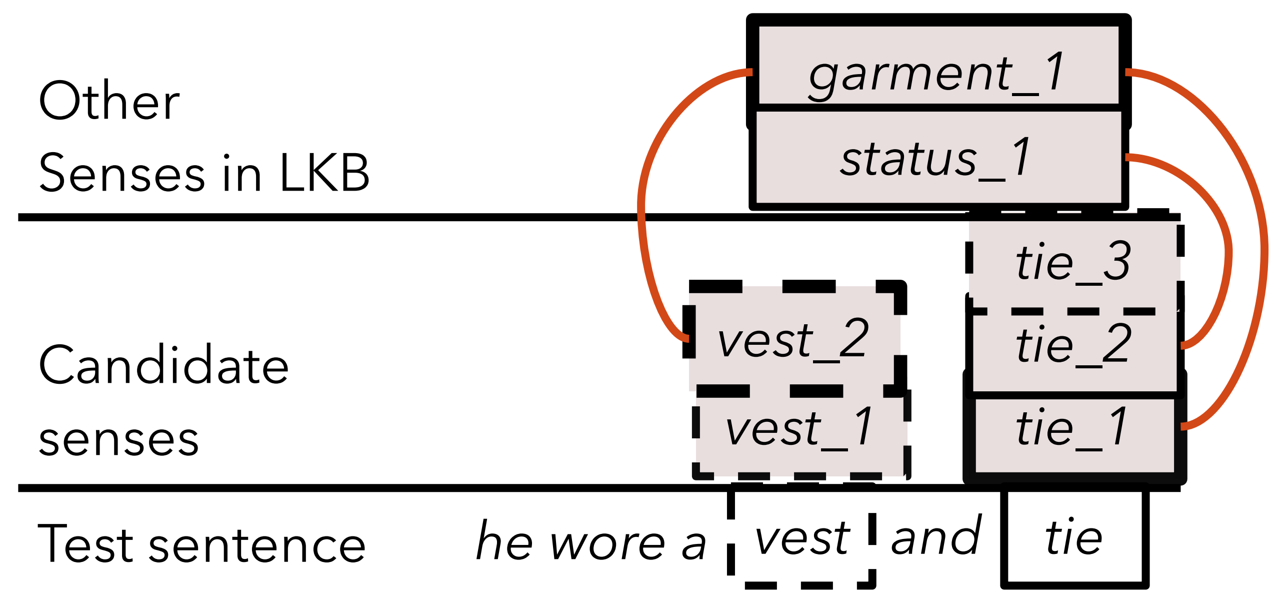
Context:
he wore a tie …

Definition (of the candidate sense):
neckwear consisting of a long narrow piece of material worn under a collar …

Knowledge Based

WordNet: Lexical knowledge base (LKB)





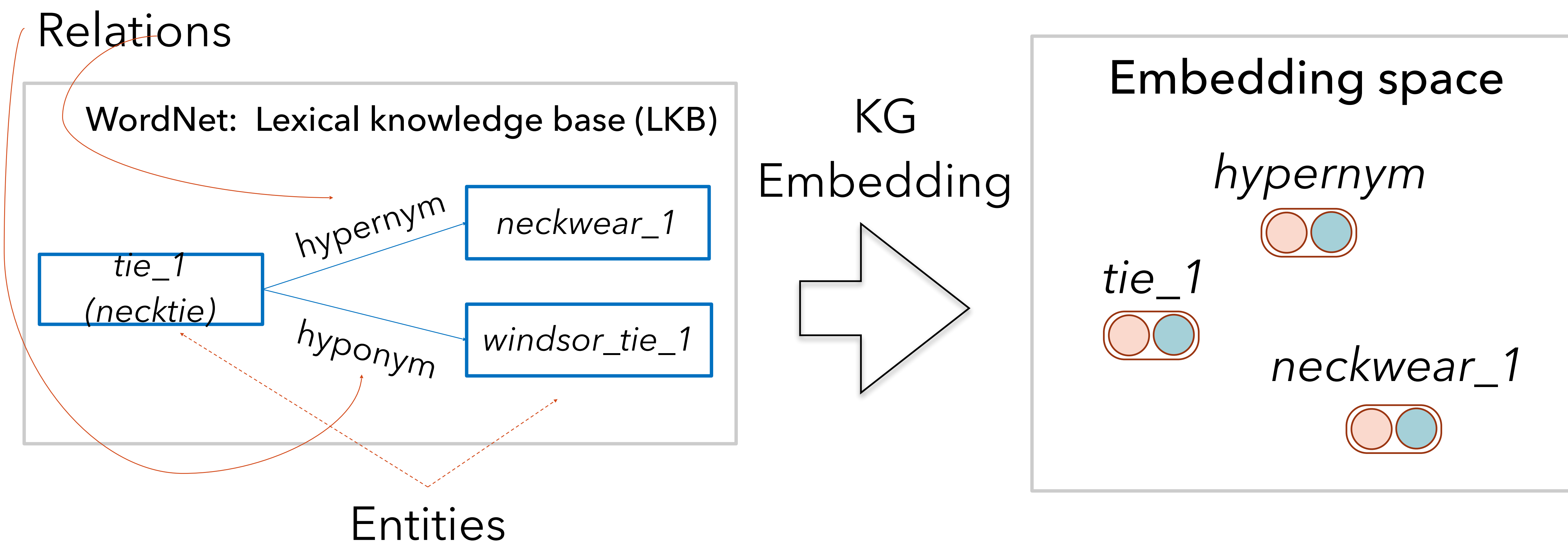
- "Embeddings for word sense disambiguation: An evaluation study.", Iacobacci et al, 2016
- "Neural sequence learning models for word sense disambiguation", Raganato et al, 2017 …
- "An enhanced Lesk word sense disambiguation algorithm through a distributional semantic model.", Basile et al, 2014
- "Random walks for knowledge-based word sense disambiguation", Agirre et al, 2014.
- "Entity linking meets word sense disambiguation: a unified approach.", Moro et al, 2014 …

EWISE vs Previous WSD Approaches

EWISE: Extended WSD Incorporating Sense Embeddings

	Handle unseen words?	Handle unseen senses?	Trainable?
Supervised (DL)	✘	✘	✓
Definition-based	✓	✓	✘
Knowledge-based	✓	✓	✘
EWISE	✓	✓	✓

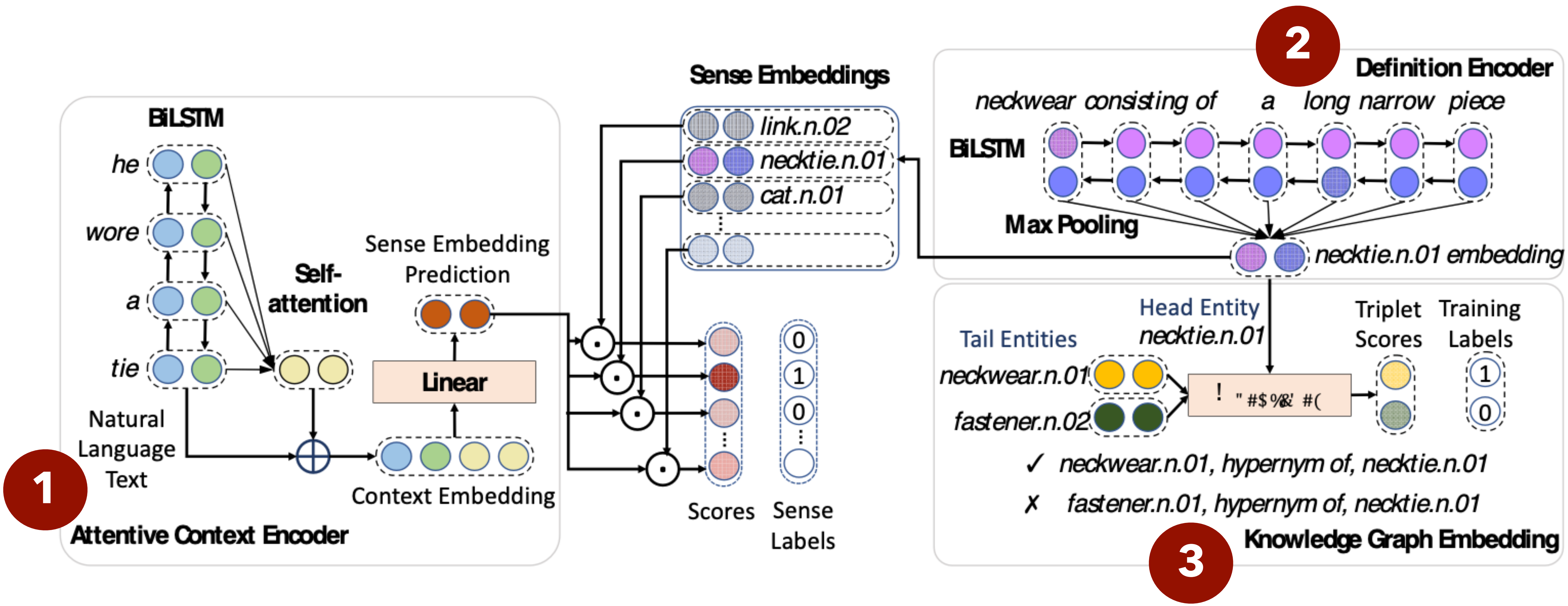
Embedding Word Senses



- "Translating embeddings for modeling multi- relational data.", Bordes et al, 2013
- "Convolutional 2d knowledge graph embeddings ", Dettmers et al, 2018 …

EWISE: Zero-shot WSD + KG

Sense embeddings, rather than discrete sense labels

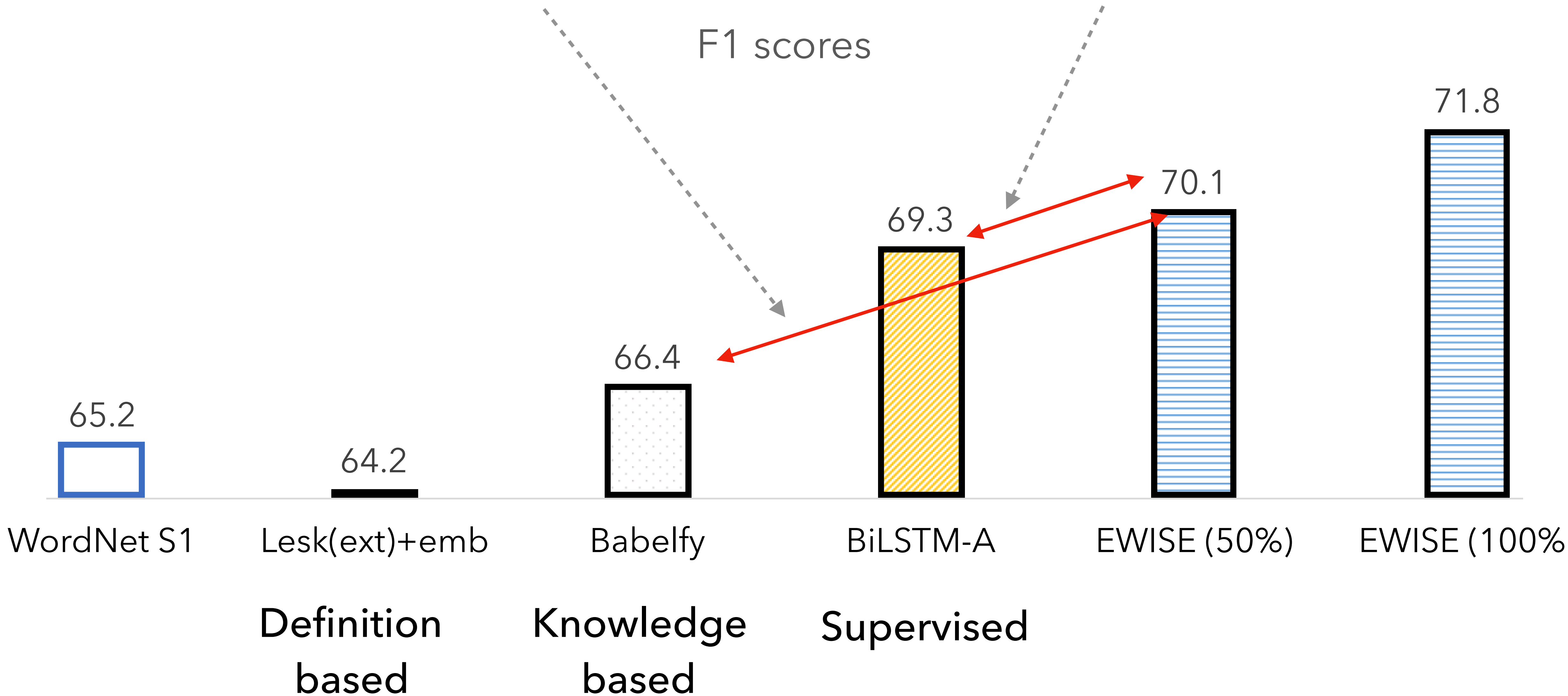


https:// github.com/malllabiisc/EWISE

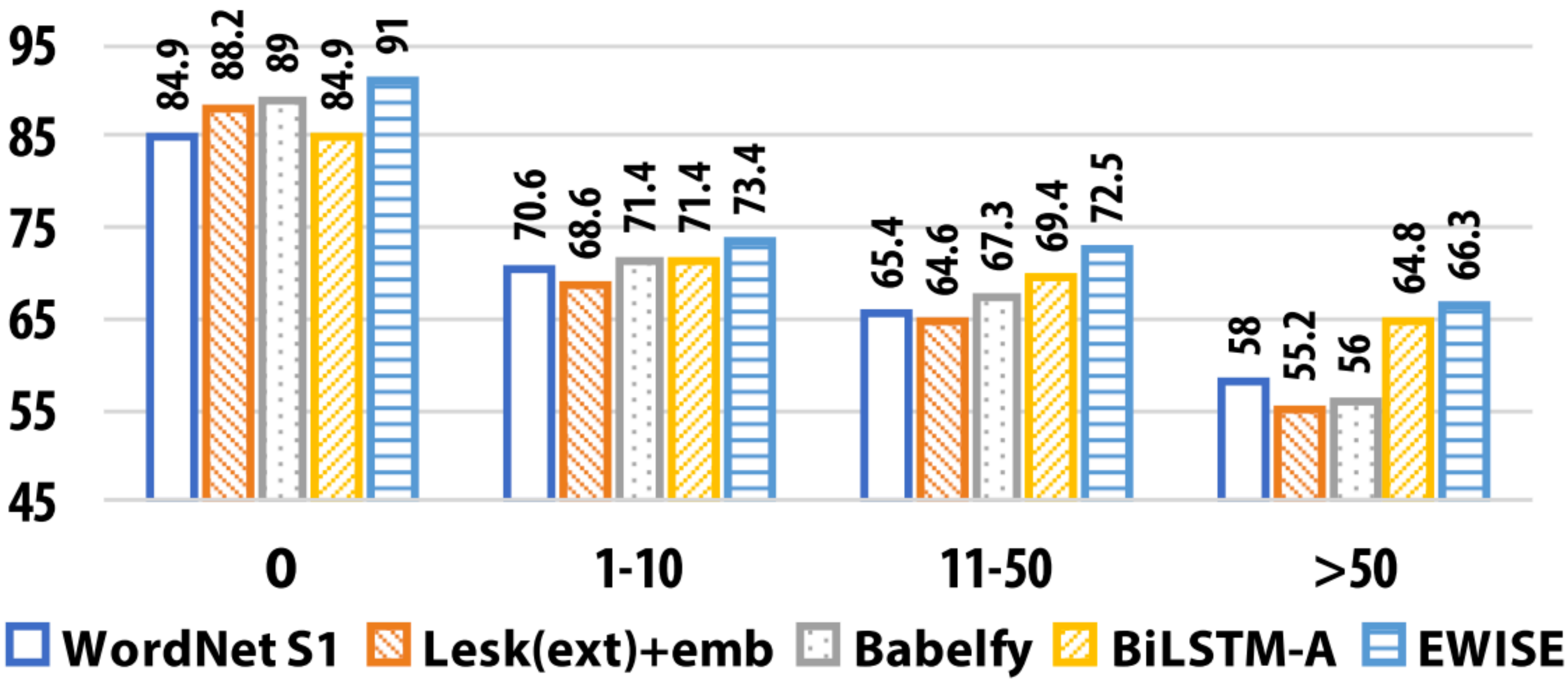
EWISE Results

EWISE provides significant gains over Knowledge based methods with limited data

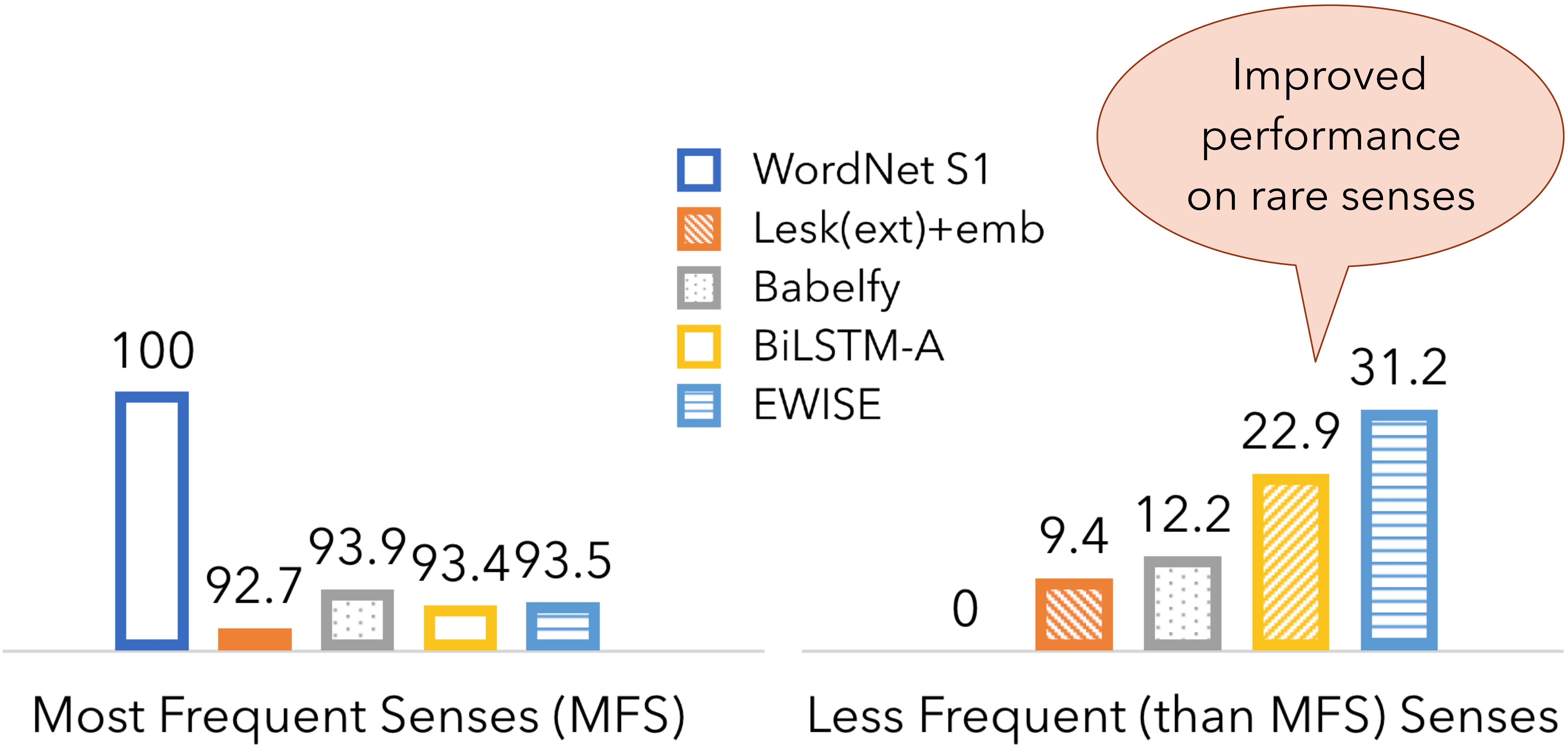
EWISE outperforms an ablated supervised version with 50% data



Performance on Unseen Words

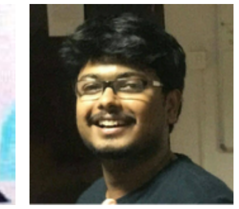


Performance on Rare Senses



MFS : instances labeled with the most frequent sense of the word

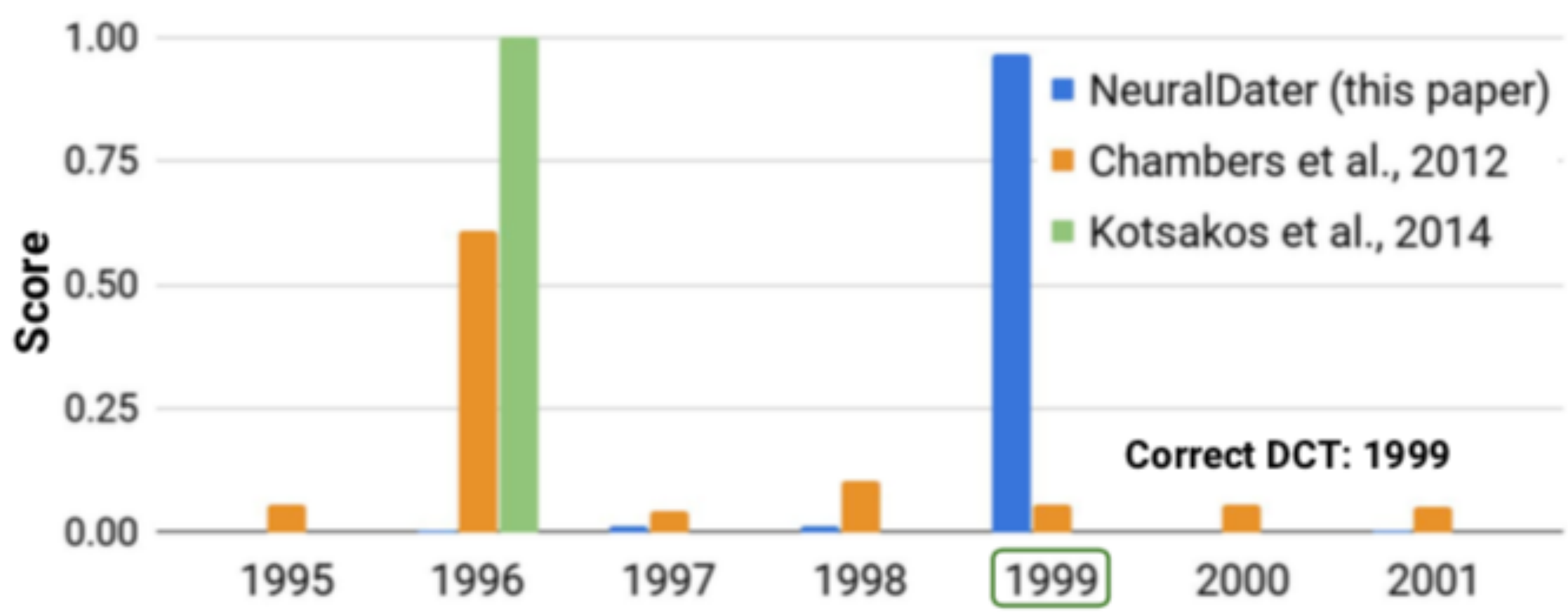
Timestamping Document using GCNs [ACL 2018]

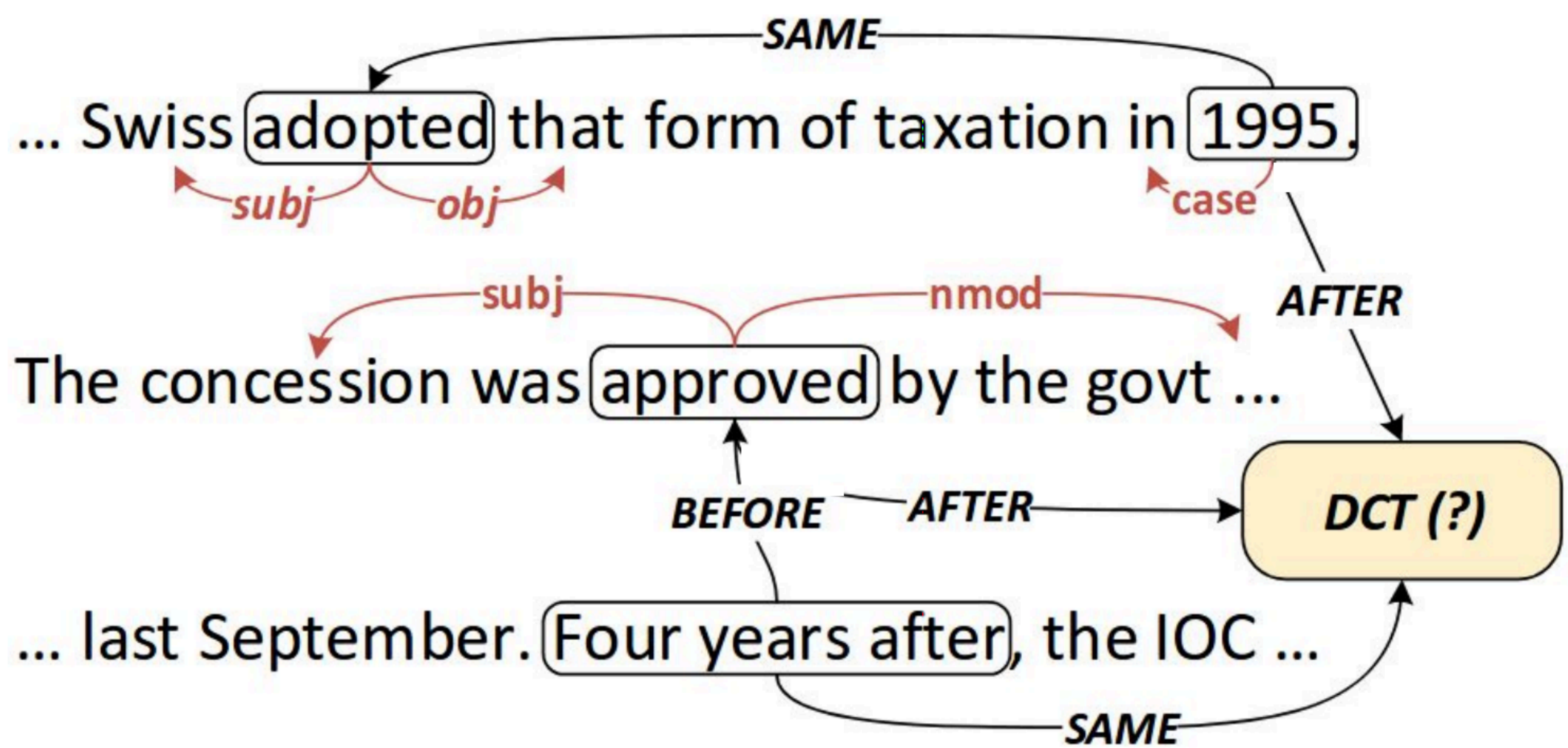

Shikhar Shib Swayambhu

... Swiss adopted that form of taxation in 1995.
The concession was approved by the govt ...
... last September. Four years after, the IOC approved ...

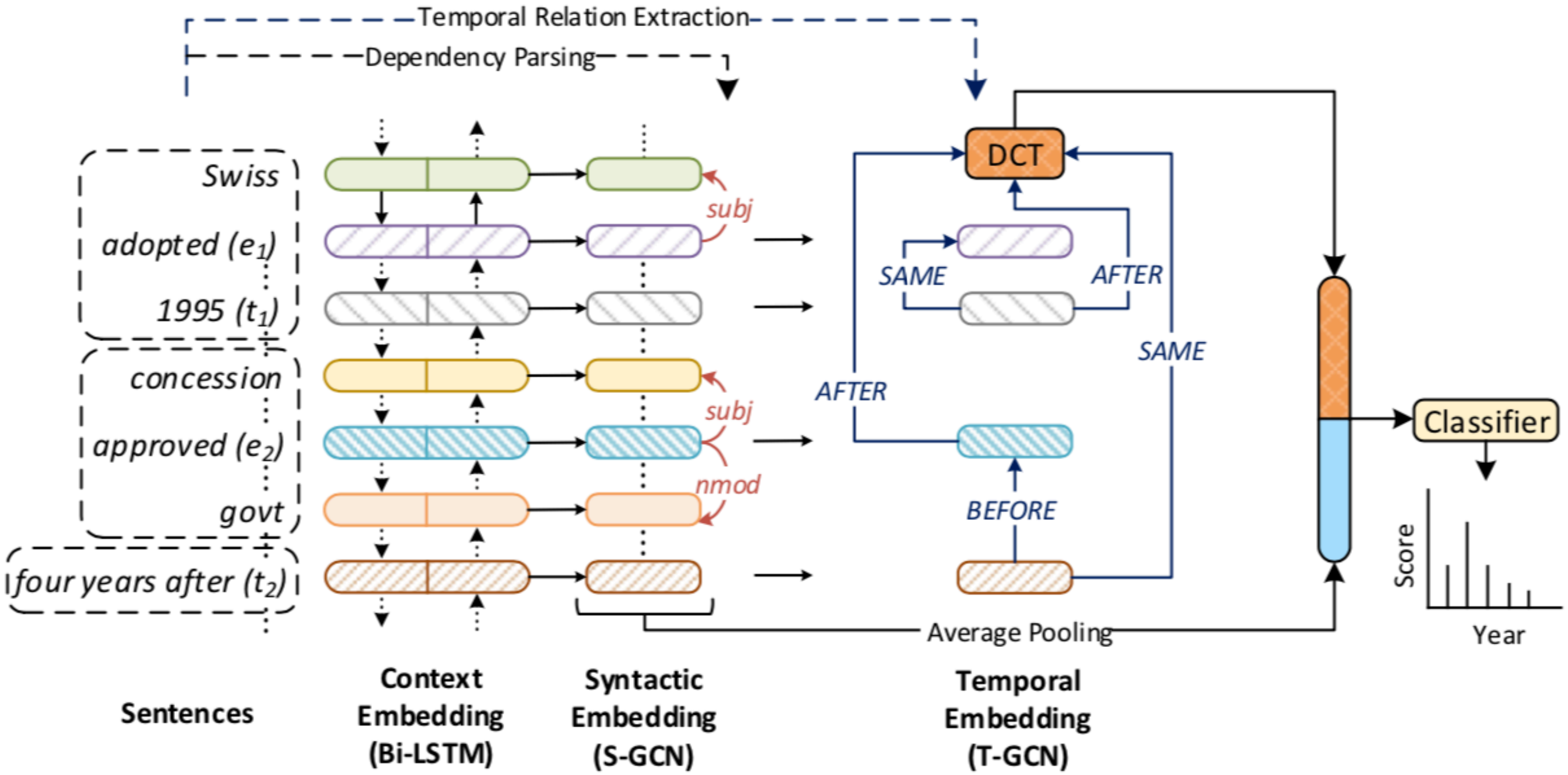
1999



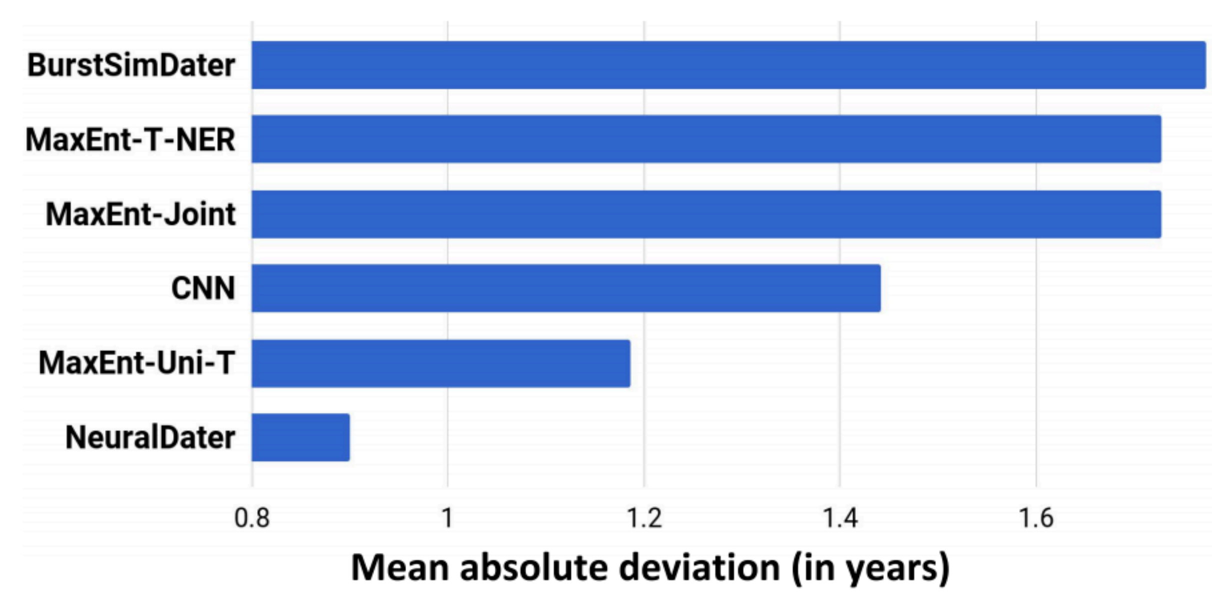
Timestamping Document using GCNs [ACL 2018, EMNLP 2018]



NeuralDater [ACL 2018, EMNLP 2018]



Method	APW	NYT
BurstySimDater	45.9	38.5
MaxEnt-Time+NER	52.5	42.3
MaxEnt-Joint	52.5	42.5
CNN	56.3	50.4
MaxEnt-Unigram	57.5	50.5
NeuralDater	**64.3**	**59.1**



Incorporating Syntactic and Semantic Information in Word Embeddings using GCNs [ACL 2019]


Shikhar


Manik


Prateek

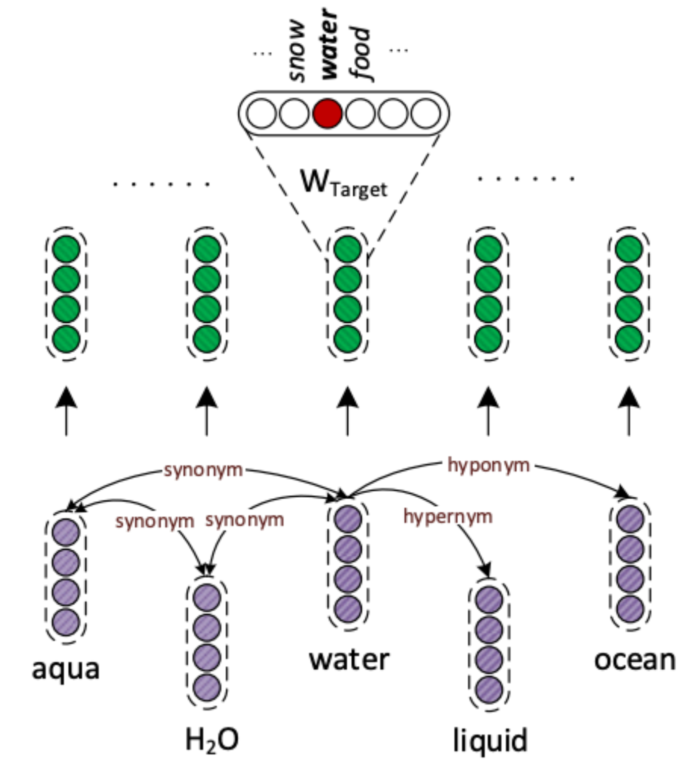

Piyush


Chiranjib

- **Graph Convolution** based method for learning word embeddings which utilizes **syntactic context** instead of linear context.

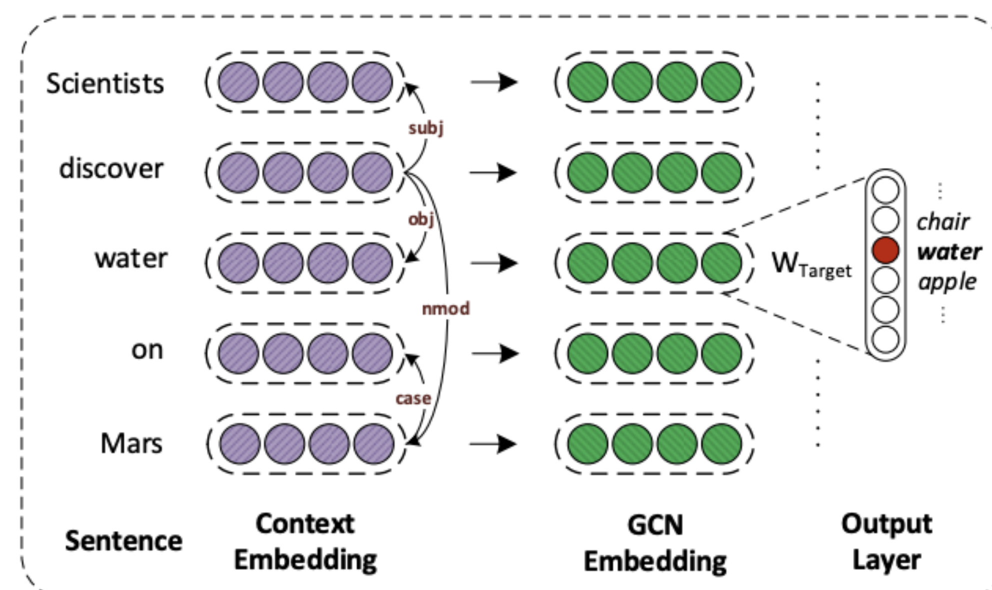
- Also allows for **incorporating diverse semantic knowledge** e.g. synonyms, antonyms, hypernyms.

- Code: **github.com/malllabiisc/WordGCN**



SemGCN (Corpus-level)

Method	POS	SQuAD	NER	Coref
X = SynGCN	95.4±0.1	79.6±0.2	**89.5±0.1**	65.8±0.1
Retro-fit (X,1)	94.8±0.1	79.6±0.1	88.8±0.1	66.0±0.2
Counter-fit (X,2)	94.7±0.1	79.8±0.1	88.3±0.3	65.7±0.3
JointReps (X,4)	95.4±0.1	79.4±0.3	89.1±0.3	65.6±0.1
SemGCN (X,4)	**95.5±0.1**	**80.4±0.1**	**89.5±0.1**	**66.1±0.1**



SynGCN (Sentence-level)

Diverse Paraphraser using Sub-modularity (DiPS) [NAACL 2019]



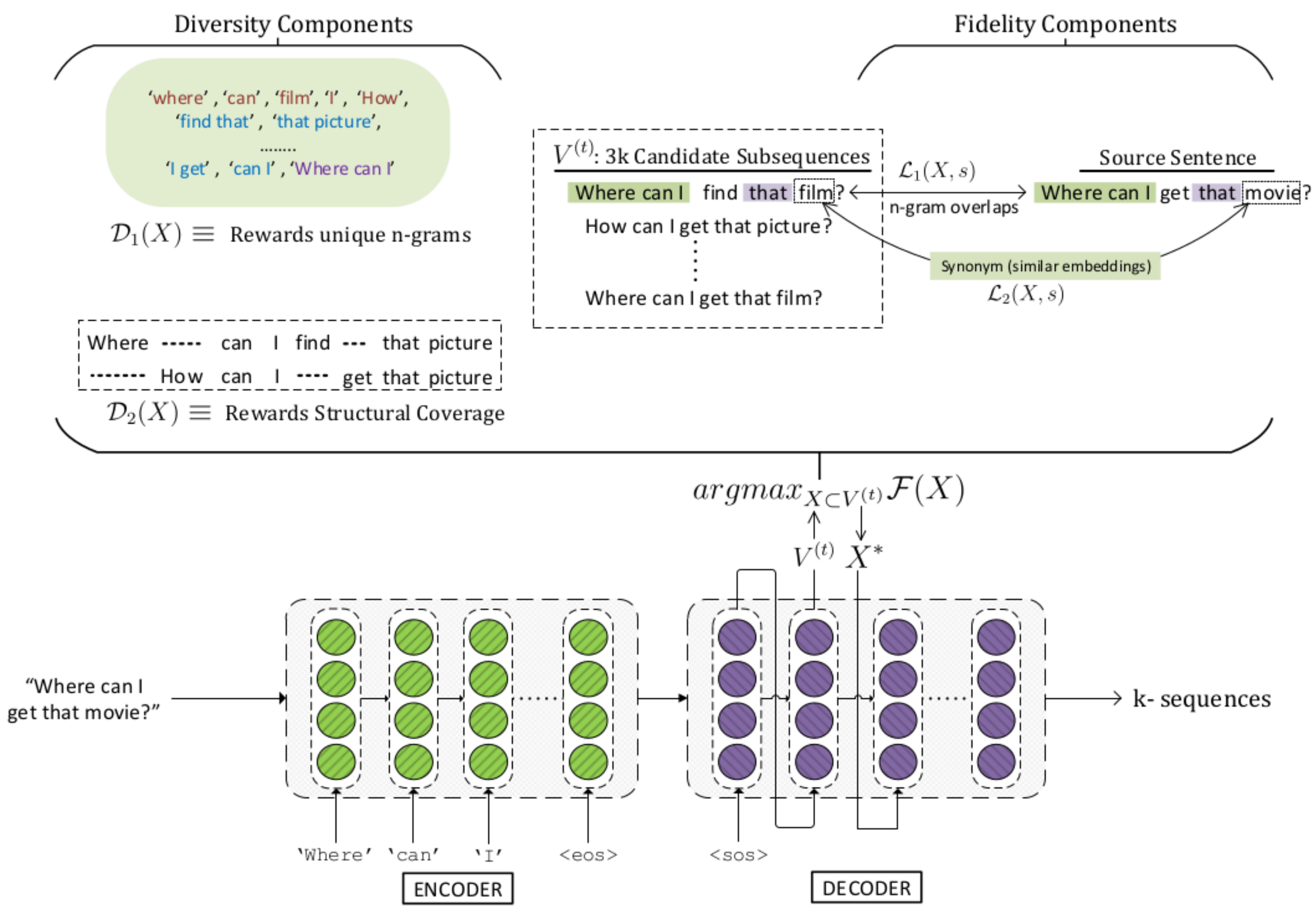
Ashutosh Satwik Manik

Source: *how do i increase body height ?*

Current Paraphrases:

1. *how do i increase my height ?*
2. *how do i increase my body height ?*
3. *how do i increase the height ?*

DiPS

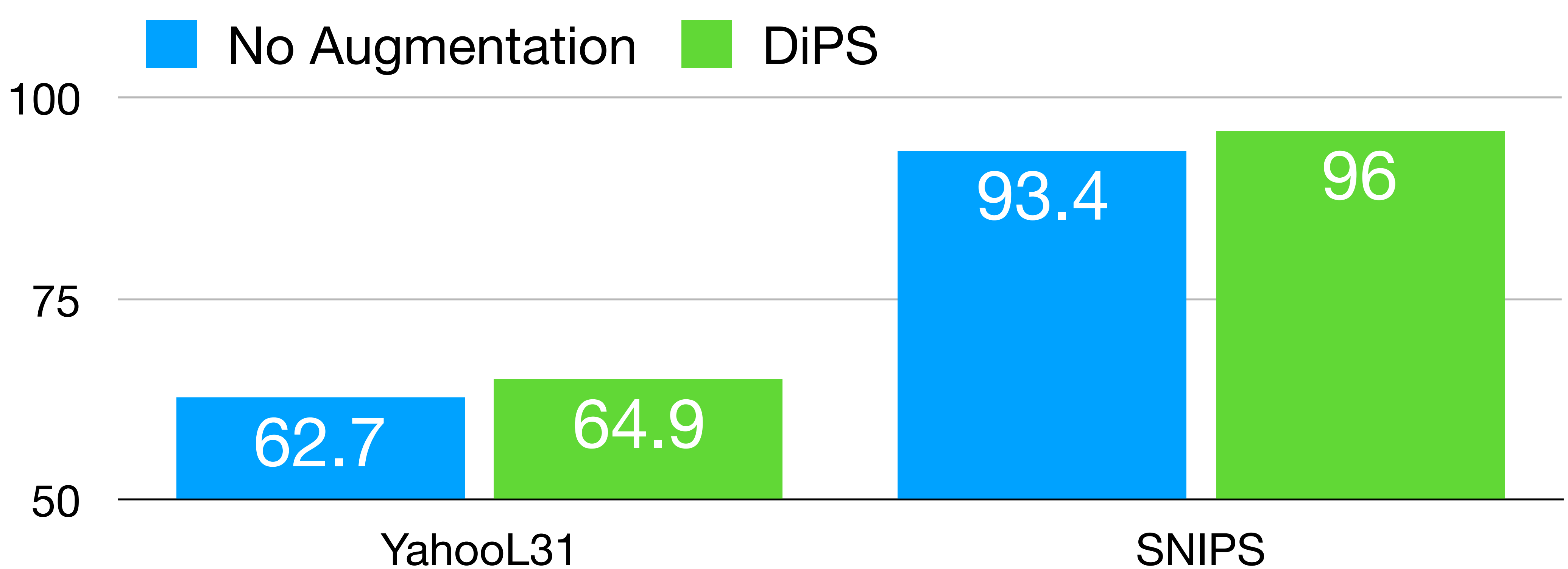


DiPS induces diversity during submodular decoding

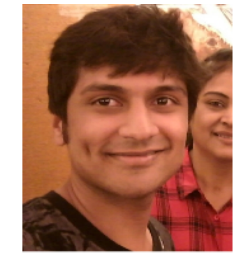
Results: Data Augmentation for Intent Classification

DiPS Paraphrases:

1. how could i increase my height?
2. what should i do to increase my height?
3. what are the fastest ways to increase my height?
4. is there any proven method to increase height?

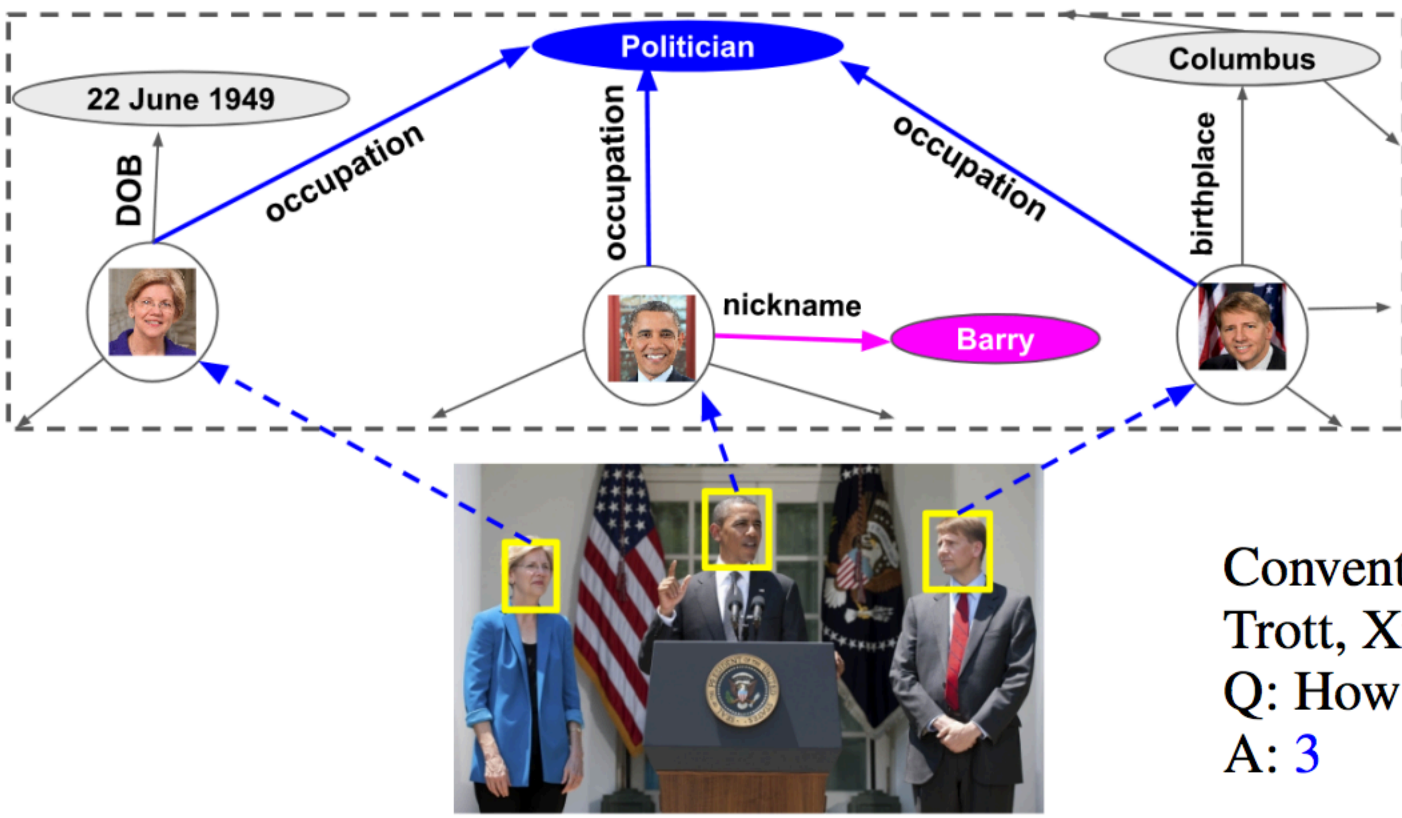


KG + Vision [AAAI 2019]


Sanket


Anand


Naganand



KVQA

[bit.ly/iisc-kvqa]

New Dataset for Knowledge-aware Computer Vision

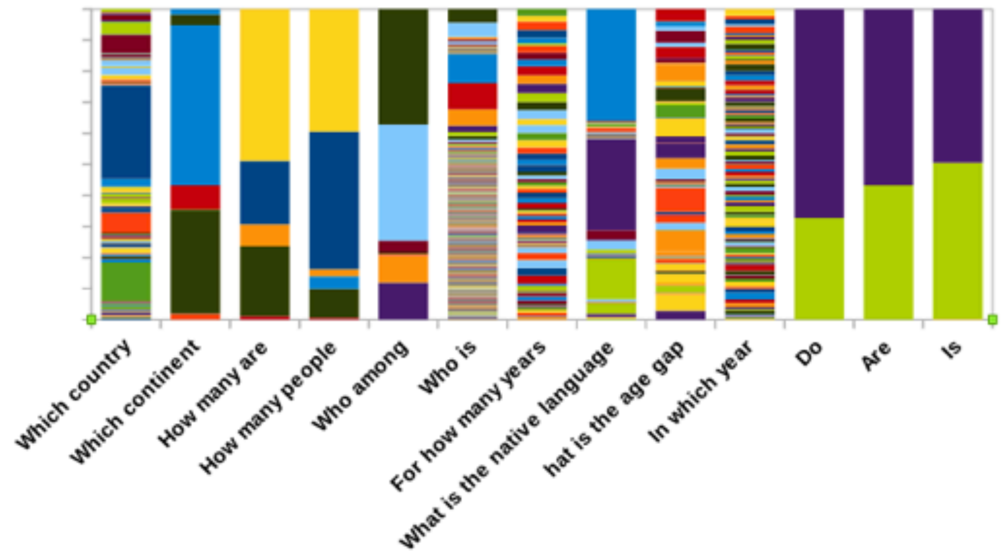
Conventional VQA (Antol et al. 2015; Goyal et al. 2017; Trott, Xiong, and Socher 2018)
Q: How many people are there in the image?
A: 3

World knowledge-enabled VQA (this paper):
Q: Who is to the left of Barack Obama?
A: Richard Cordray
Q: Do people in the image have common occupation?
A: Yes
Q: Who among the people in the image is called by the nick-name Barry?
A: Person in the center

KVQA: in Summary

bit.ly/iisc-kvqa

24,602 images and 18800 persons

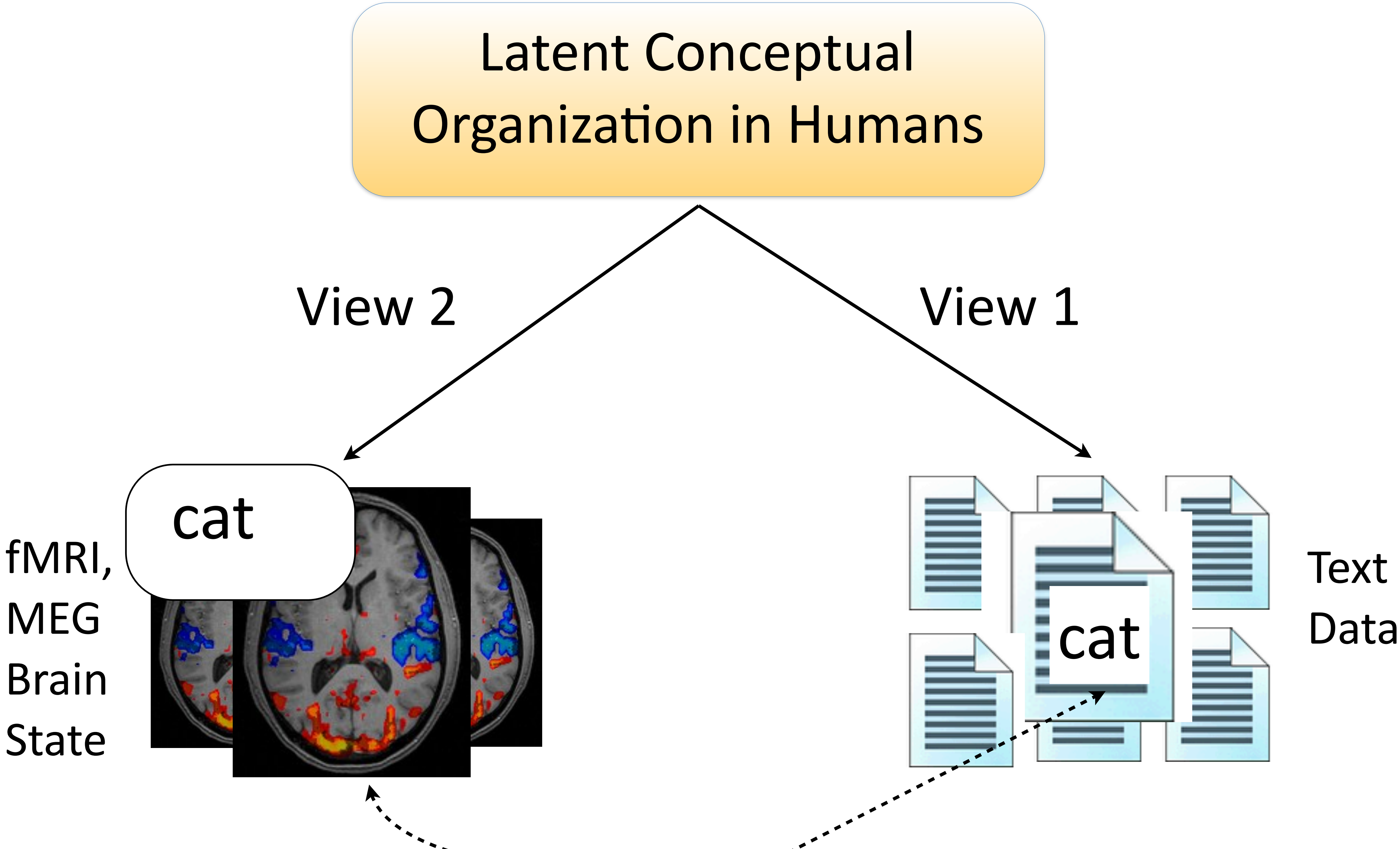


183,007 QA pairs about images

19,571 unique answers

Reference images of 69K persons with their Wikidata ID

Broader Interest: Latent Conceptual Organization in Brains and Text



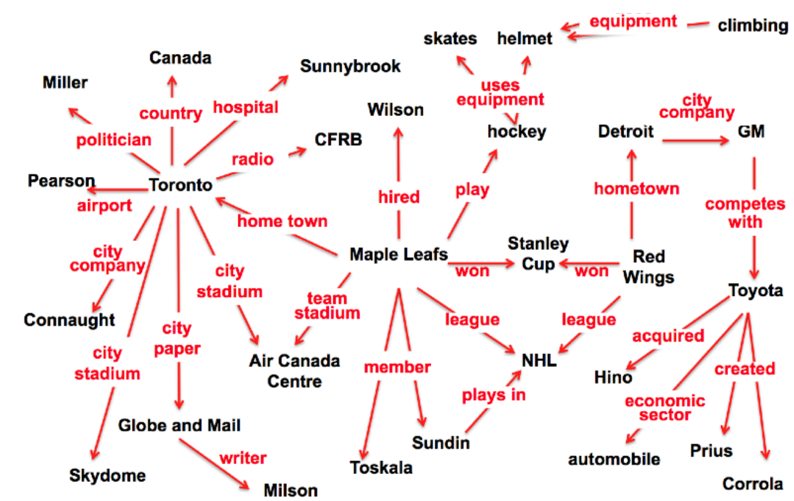
StarSem 12, COLING 12, CoNLL 13, KDD 14, SDM 14, ACL 14, PLoS ONE, ACL 19]

From Strings to Things and Beyond

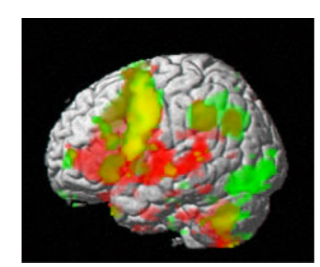
Construct

Maintain

New Data



Apply



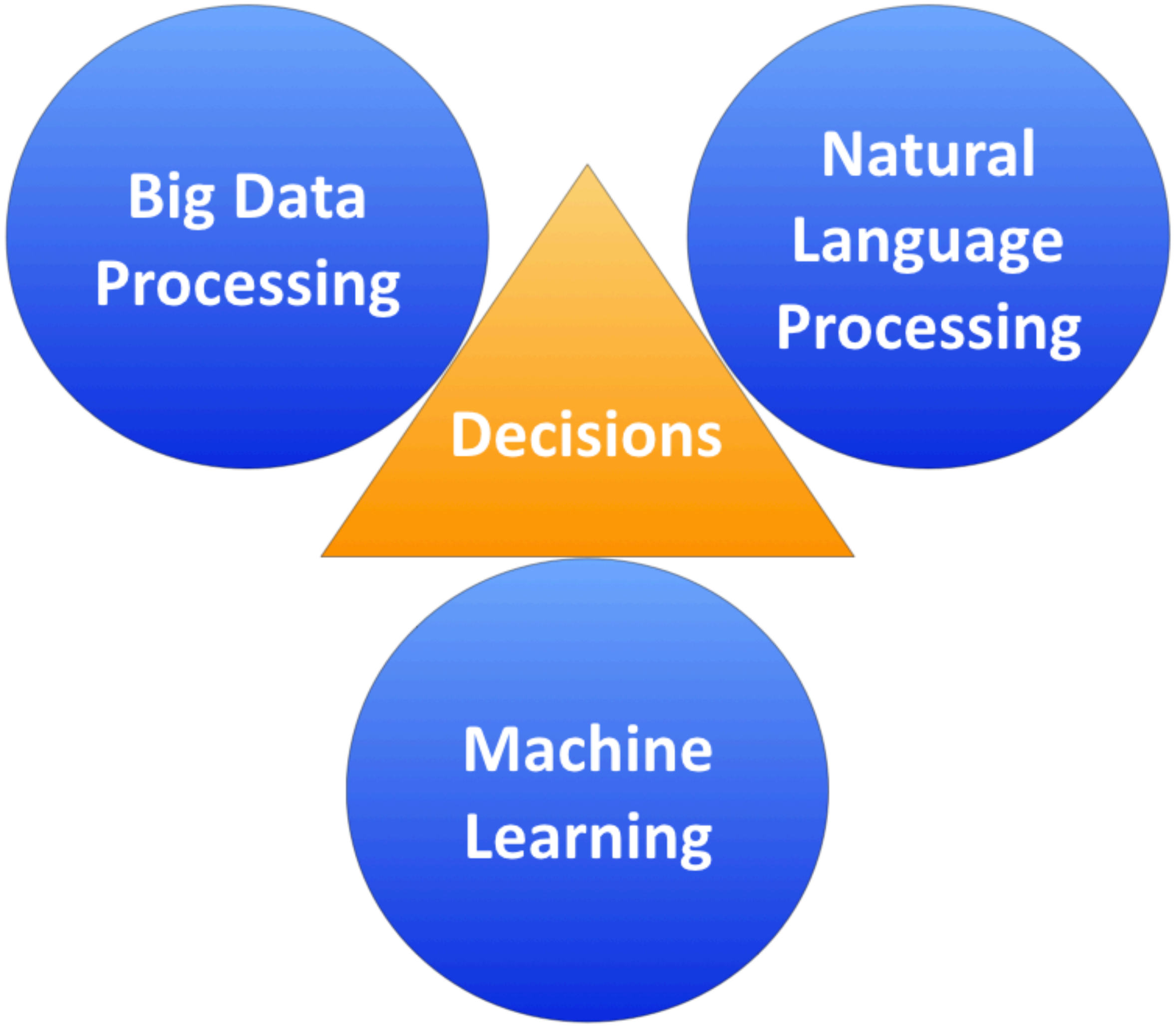
 or  ?

Ongoing Research at MALL Lab

- Large-scale Interpretable Representation Learning [Chandrahas]
- Learning with Explanations [Sawan]
- Deep Learning over Hypergraphs [Naganand]
- Knowledge Acquisition from Technical Literature [Harshita, Akash]
- Data Augmentation and Controlled Generation [Ashutosh]
- Deep Learning for Material Science [Soumya]
- Question answering over Knowledge Graphs [Apoorv]

7 PhD, 3 Masters,
4 Research Assistants, 4 Interns

Alumni @ Google, Facebook, Samsung R&D, Microsoft, Adobe, Amazon, Flipkart, Columbia, CMU, Stanford, UIUC, UMass, UT Austin



talukdar.net

malllabiisc.github.io