Incorporating Structured World Knowledge into Unstructured Documents via **Heterogeneous Information Networks**

Yangqiu Song

香 港 科 技 大 學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Collaborators

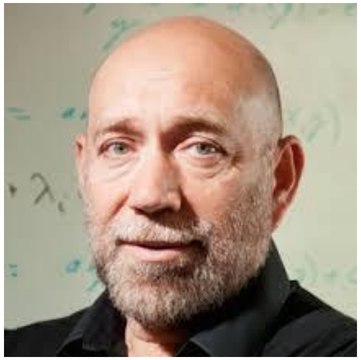
Chenguang Wang

Ming Zhang

Yizhou Sun

Jiawei Han

Dan Roth



Slides Credit: Chenguang Wang

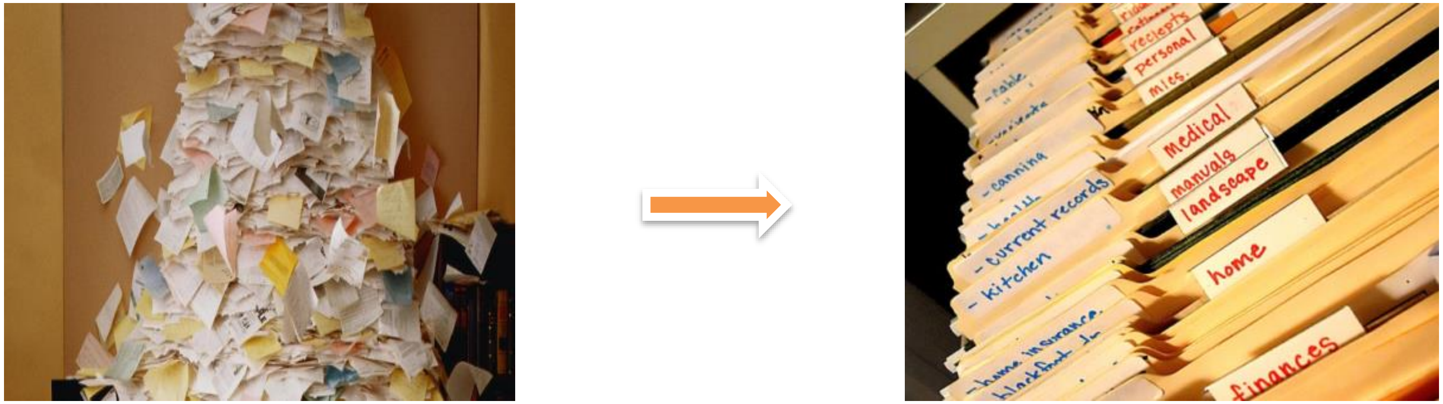
Outline

- Text Analytics: Motivation
  - Two Challenges
    - Representation
    - Labels

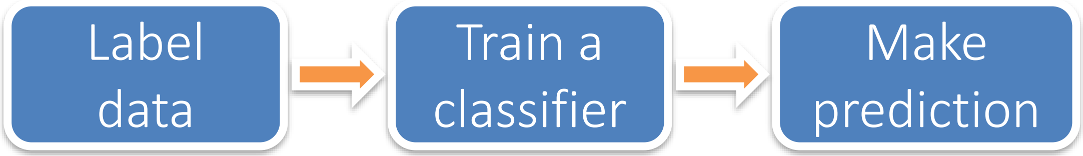
- Text Categorization via HIN
  - HIN construction from texts
  - From HIN similarity to clustering and classification
  - World knowledge indirect supervision

- Conclusions and future work

Text Categorization: Two Challenges



- Impacts many applications!
  - ✓ Social network analysis, health care, machine reading …
- Traditional approach:

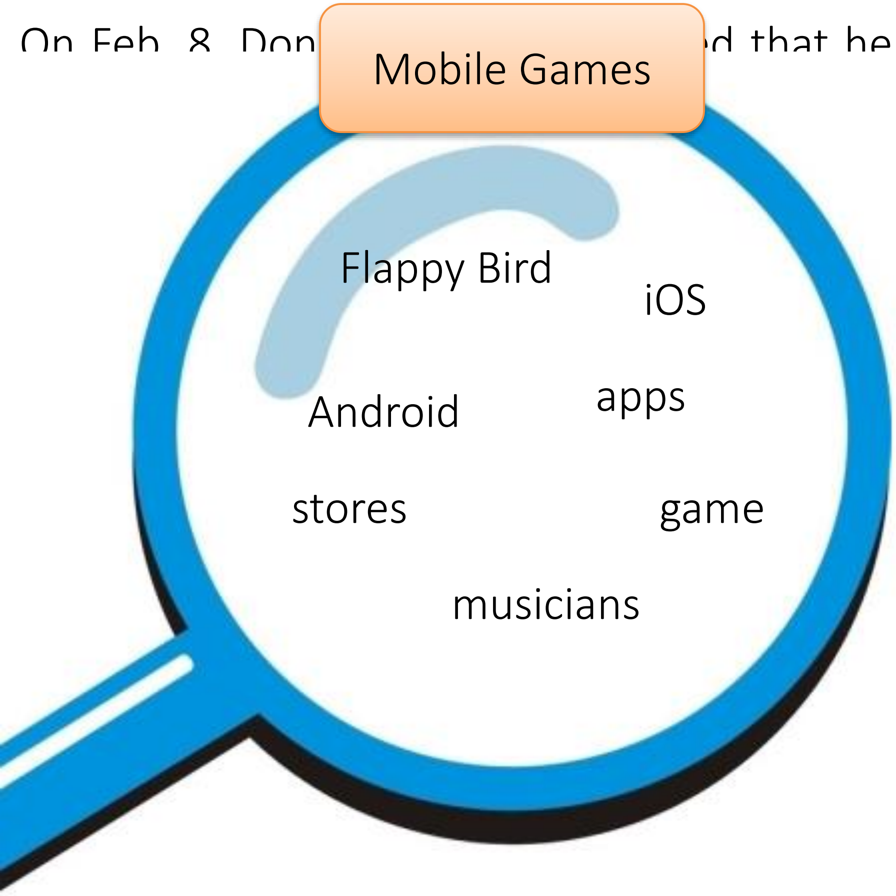


Label data → Train a classifier → Make prediction

- Two challenges:
  - ✓ Representation
  - ✓ Labels

Representation: Bag-of-words

On Feb. 8, Don d that he



Mobile Games

Flappy Bird

iOS

Android

apps

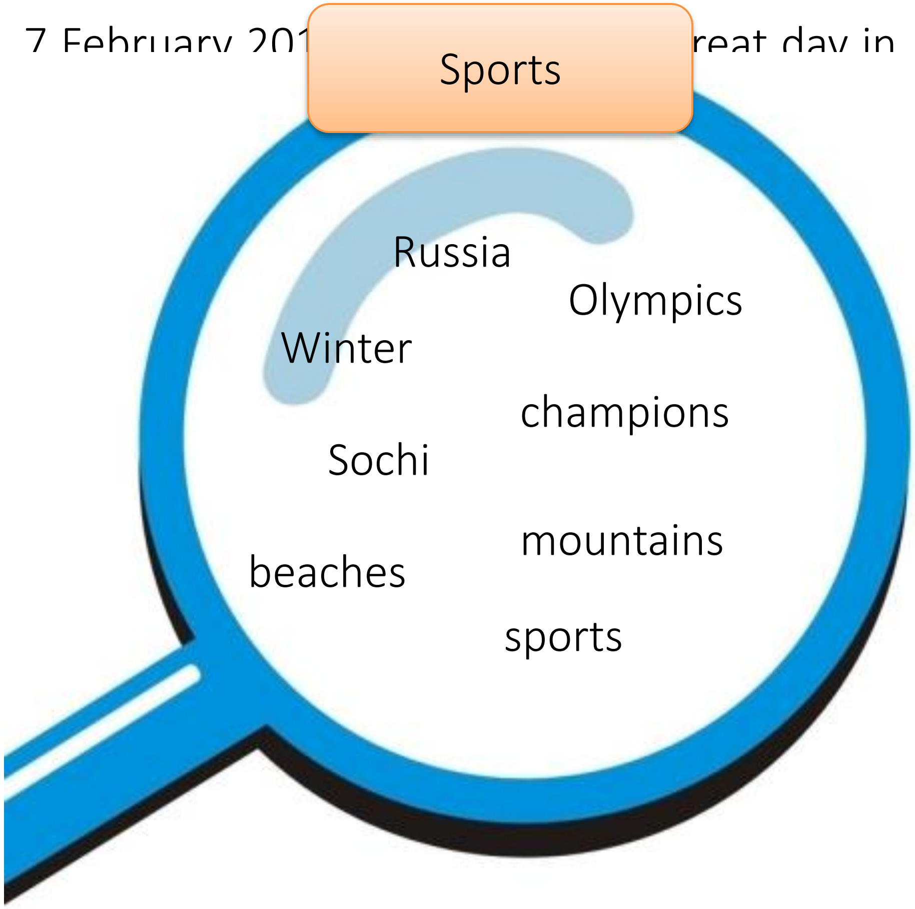
stores

game

musicians

Internet trolls."

7 February 201 reat day in



Sports

Russia

Olympics

Winter

champions

Sochi

mountains

beaches

sports

from 7 to 23 February 2014.

Context: Topic Models and Word Embeddings

- Topic Modeling (Blei et al., 2003)

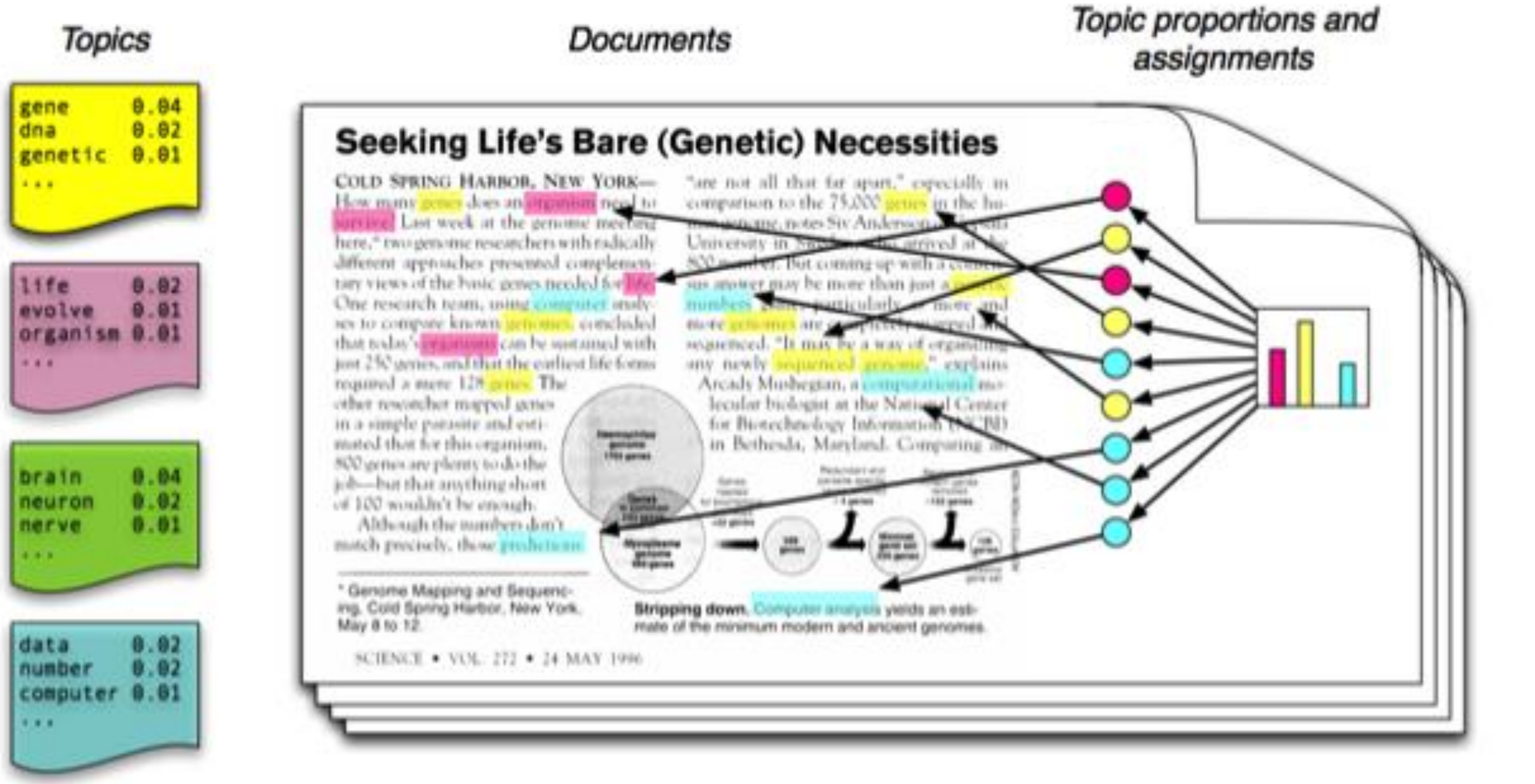
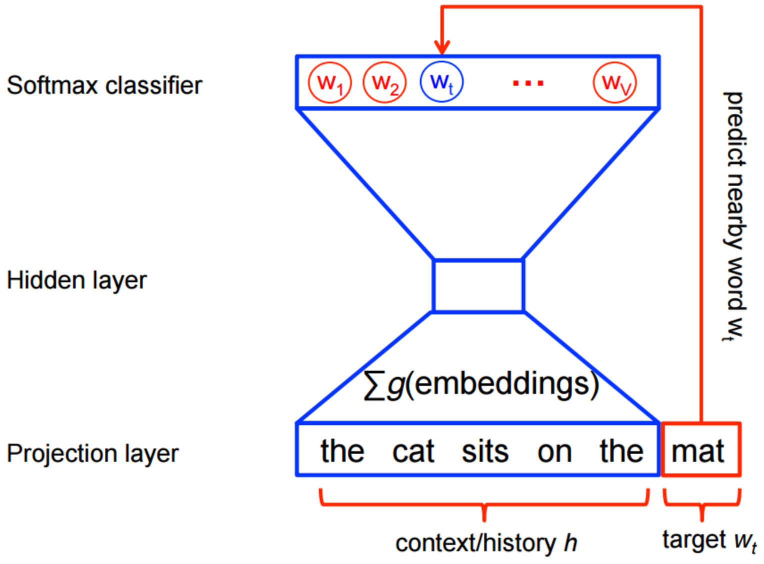
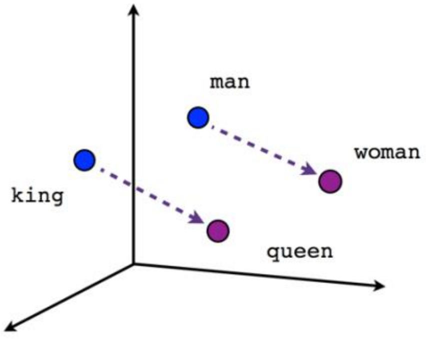

Topics
Documents
Topic proportions and assignments
gene 0.04
dna 0.02
genetic 0.01
...
life 0.02
evolve 0.01
organism 0.01
...
brain 0.04
neuron 0.02
nerve 0.01
...
data 0.02
number 0.02
computer 0.01
...
Seeking Life's Bare (Genetic) Necessities
COLD SPRING HARBOR, NEW YORK—
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.
SCIENCE • VOL. 272 • 24 MAY 1996

Figure source: Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77-84.

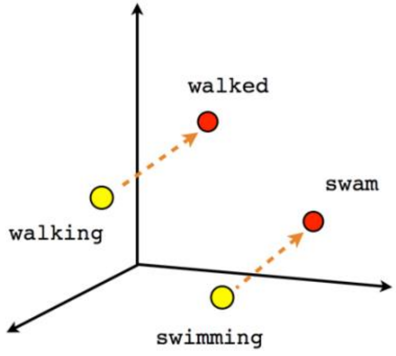
Context: Topic Models and Word Embeddings

- Word embedding
  - Word2vec (Mikolov et al., 13)
  - Glove (Pennington et al., 14)
  - Matrix factorization (Deerwester’90;Levy et al., 15)
  - …

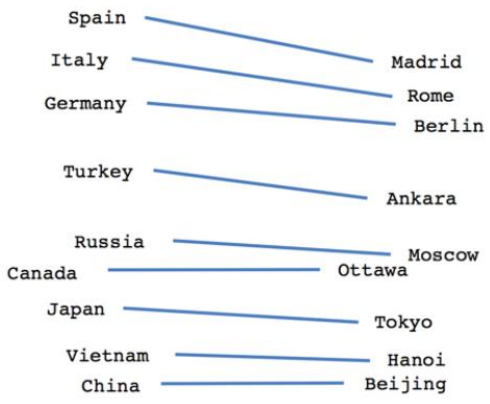




Male-Female



Verb tense

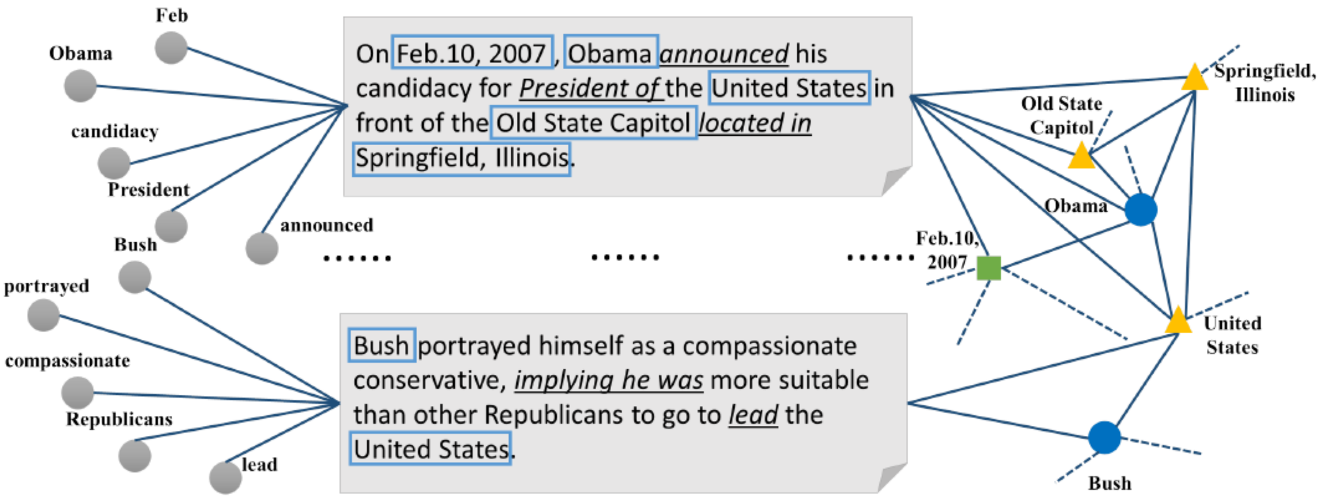


Country-Capital

https://www.tensorflow.org/versions/r0.7/tutorials/word2vec/index.html

What’s Missing?

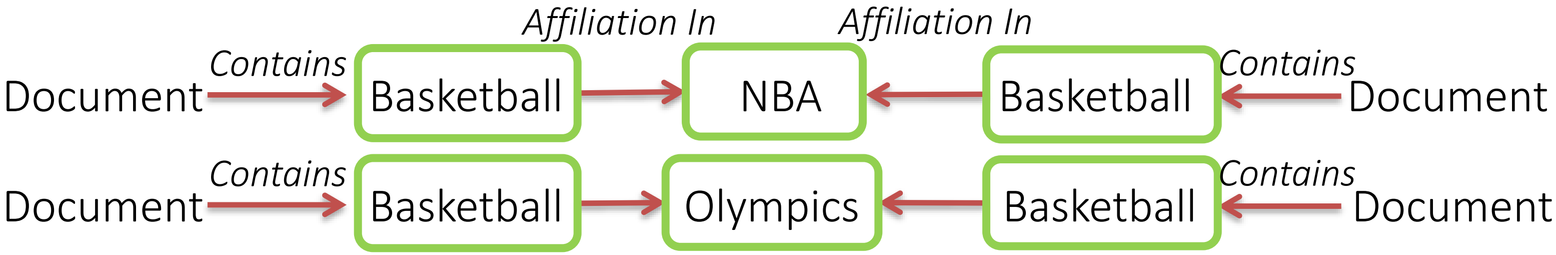
- The semantics of entities and their relations



On Feb.10, 2007, Obama *announced* his candidacy for *President of* the United States in front of the Old State Capitol *located in* Springfield, Illinois.

Bush portrayed himself as a compassionate conservative, *implying he was* more suitable than other Republicans to go to *lead* the United States.

- What can context cover? ``New York'' vs. ``New York Times''
- What cannot? ``George Washington'' vs. ``Washington''
  - Higher order relations



Document *Contains* Basketball *Affiliation In* NBA *Affiliation In* Basketball *Contains* Document

Document *Contains* Basketball Olympics Basketball *Contains* Document

Outline

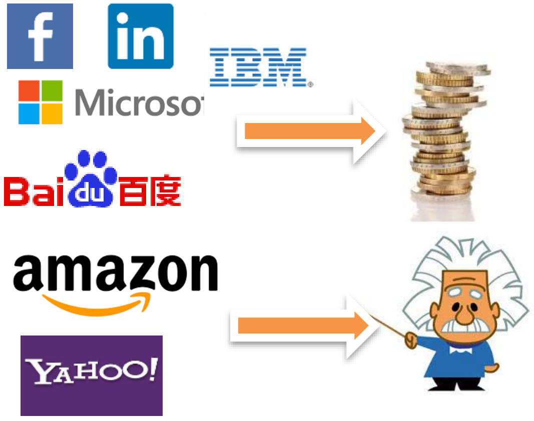
- Text Analytics: Motivation
  - Two Challenges
    - Representation
    - Labels

- Text Categorization via HIN
  - HIN construction from texts
  - From HIN similarity to clustering and classification
  - World knowledge indirect supervision

- Conclusions and future work

Acquire Labeled Data

Expert Annotation



Only big companies can hire a lot of experts

Costly

Crowdsourcing

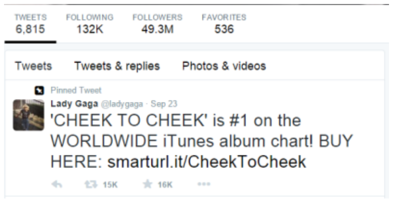
Simple tasks

Low quality

Still costly

Semi-supervised /transfer learning

Fast changing domains



TWEETS 6,815 FOLLOWING 132K FOLLOWERS 49.3M FAVORITES 536

Tweets Tweets & replies Photos & videos

Pinned Tweet

Lady Gaga @ladygaga · Sep 23

'CHEEK TO CHEEK' is #1 on the WORLDWIDE iTunes album chart! BUY HERE: smarturl.it/CheekToCheek

15K 16K

KDD 2015

KDD 2015, 10-13 August 2015, Sydney

21th ACM SIGKDD Conference on Knowledge Discovery and Data Mining homepage... KDD 2015 is a premier conference that b...

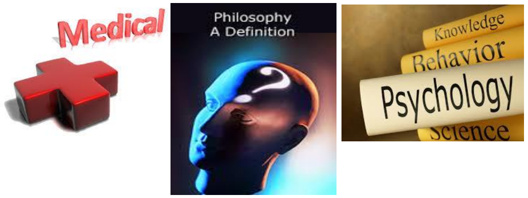
KDD CUP 2015

Yao-Yuan Yang, Sheng-Chi You, Sz-Han Yu, Hsuan-Tien Lin, and Shou-De Lin

[illegible] Aggregation of Gradient Boosting Machines for KDD Cup 2015...

www.kddcup2015.com/

Many diverse domains



Domain dependent

Our Solution

- World Knowledge enabled learning
  - – Millions of entities and concepts
  - – Billions of relationships

WIKIPEDIA
The Free Encyclopedia

Freebase Labs

NELL

yago select knowledge

DBpedia

ProBase

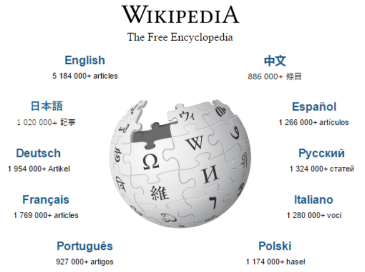
The Knowledge Graph

Learn more about one of the key breakthroughs behind the future of search.

- Grounding texts to knowledge bases

Classification without Supervision

- Label names carry a lot of information
  - We can use world knowledge as features
  - Classify document to English labels
  - 179 languages with Wikipedia
- July 15 08:30–09:55:
  - Machine Learning19: Classification2



Wikipedia

The Free Encyclopedia

English
5 184 000+ articles

中文
886 000+ 條目

日本語
1 020 000+ 記事

Español
1 266 000+ artículos

Deutsch
1 954 000+ Artikel

Русский
1 324 000+ статей

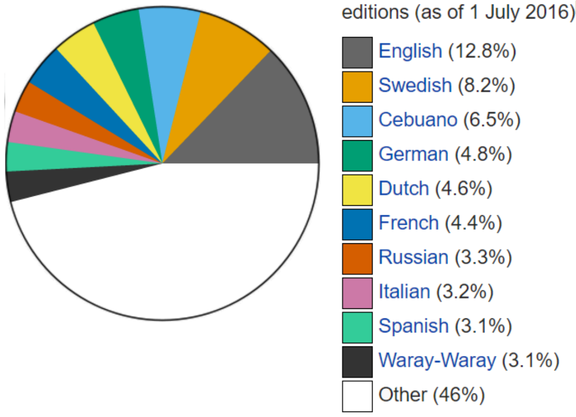
Français
1 769 000+ articles

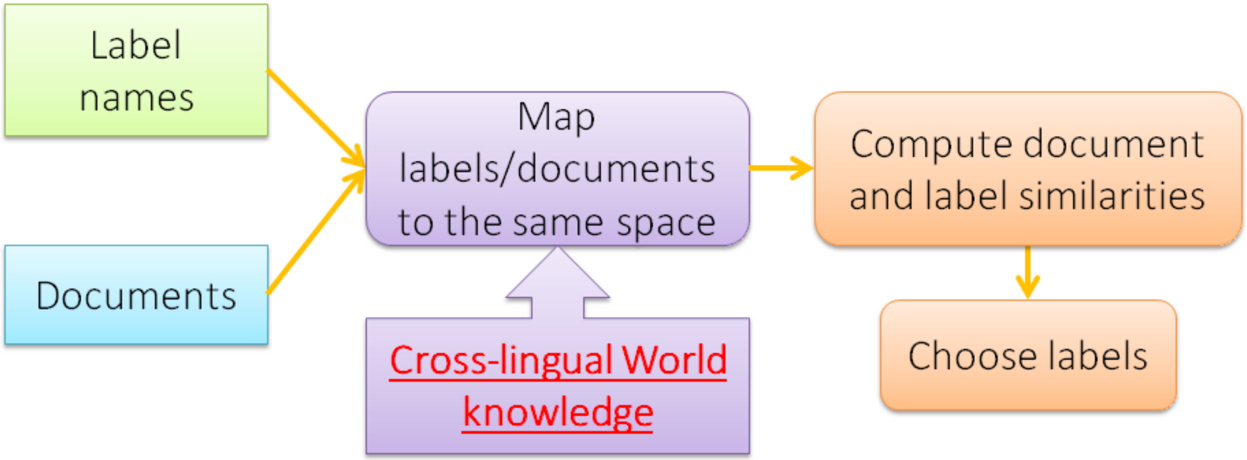
Italiano
1 280 000+ voci

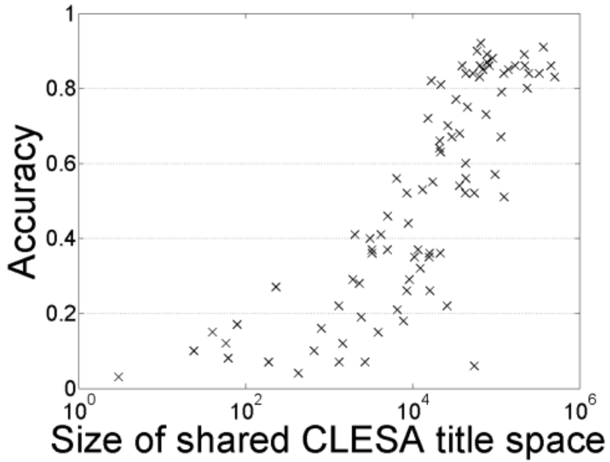
Português
927 000+ artigos

Polski
1 174 000+ haseł

Distribution of the 40,633,831 articles in different language editions (as of 1 July 2016)[1]




Label names
Documents
Map labels/documents to the same space
Cross-lingual World knowledge
Compute document and label similarities
Choose labels



M. Chang, L. Ratinov, D. Roth, V. Srikumar: Importance of Semantic Representation: Dataless Classification. AAAI'08.
Y. Song, D. Roth: On dataless hierarchical text classification. AAAI'14.
Y. Song, D. Roth: Unsupervised Sparse Vector Densification for Short Text Similarity. HLT-NAACL'15.

This Talk: Structured World Knowledge Enabled Learning and Text Mining

Different domains

tweets, blogs, websites, medical, psychology

+

Structured world knowledge bases

Wikipedia
The Free Encyclopedia

NELL

Freebase Labs

ProBase

The Knowledge Graph
Learn more about one of the key breakthroughs behind the future of search.

DBpedia

yago
select knowledge

With help of machine learning algorithms

=

More general and effective machine learning/ data mining

[Document similarity in ICDM'15]
[Document clustering in KDD'15]
[Document classification in AAAI'16]
[Item recommendation, ongoing]

[Relation clustering in IJCAI'15]
[Similarity search in SDM'16]
[Paraphrasing in ACL'13]
[Data type refinement, ongoing]

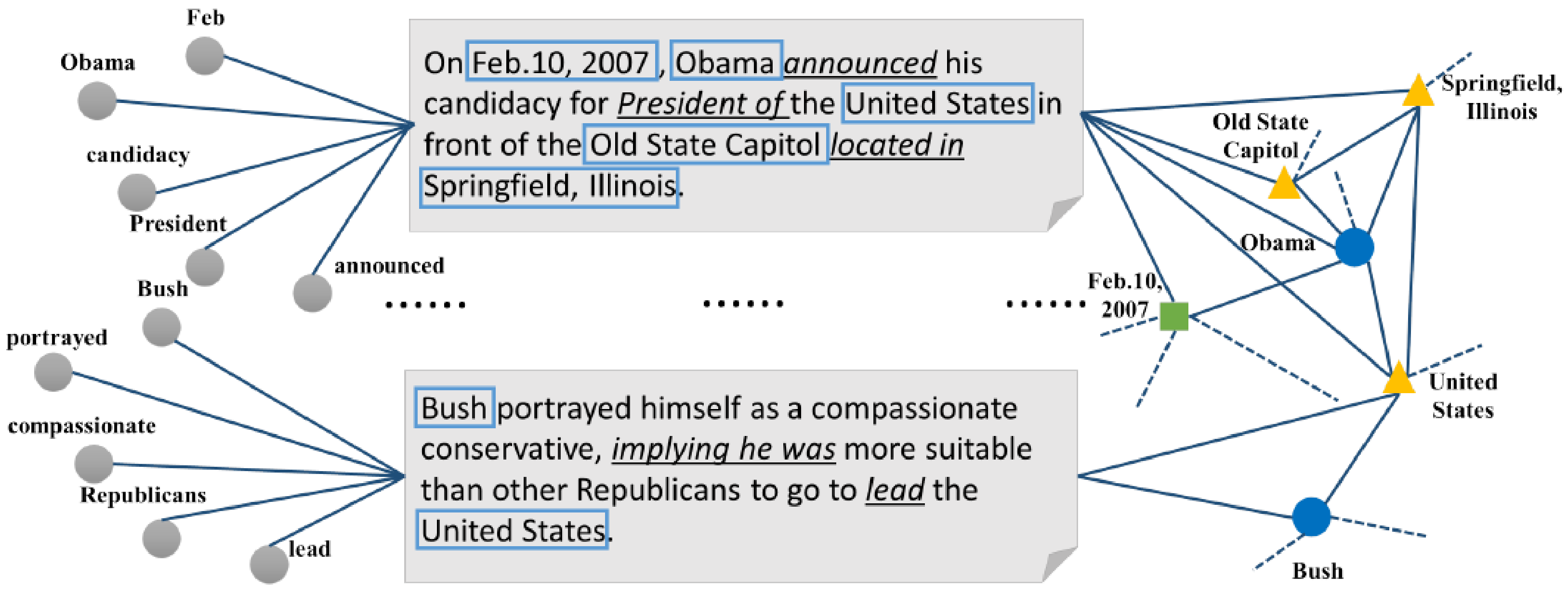
Outline

- Motivation
  - Two Challenges
    - Representation
    - Labels

- Text Categorization via HIN
  - HIN construction from texts
  - From HIN similarity to clustering and classification
  - World knowledge indirect supervision

- Conclusions and future work

Text Categorization via HIN



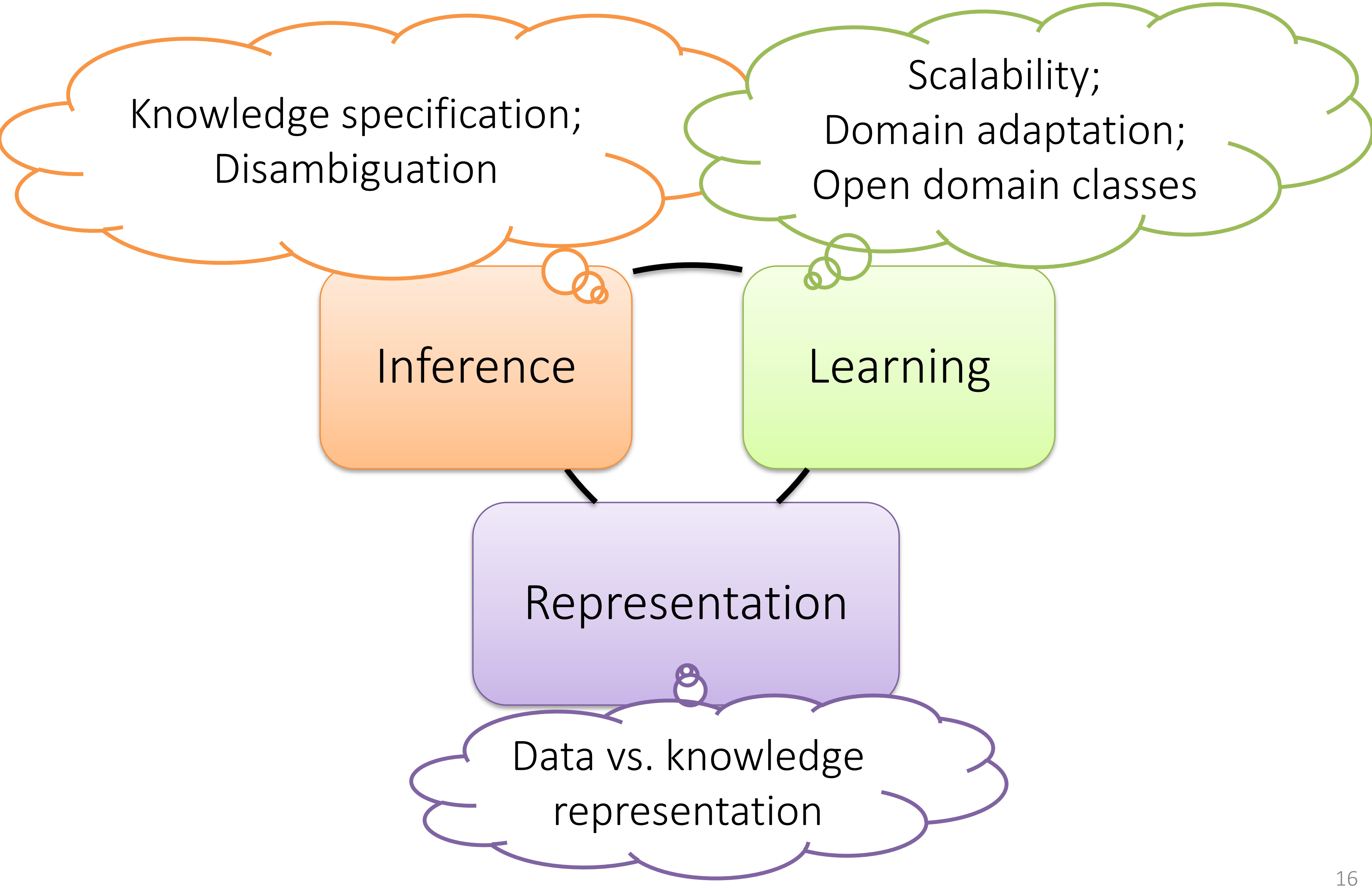
On Feb.10, 2007, Obama *announced* his candidacy for *President of* the United States in front of the Old State Capitol *located in* Springfield, Illinois.

Bush portrayed himself as a compassionate conservative, *implying he was* more suitable than other Republicans to go to *lead* the United States.

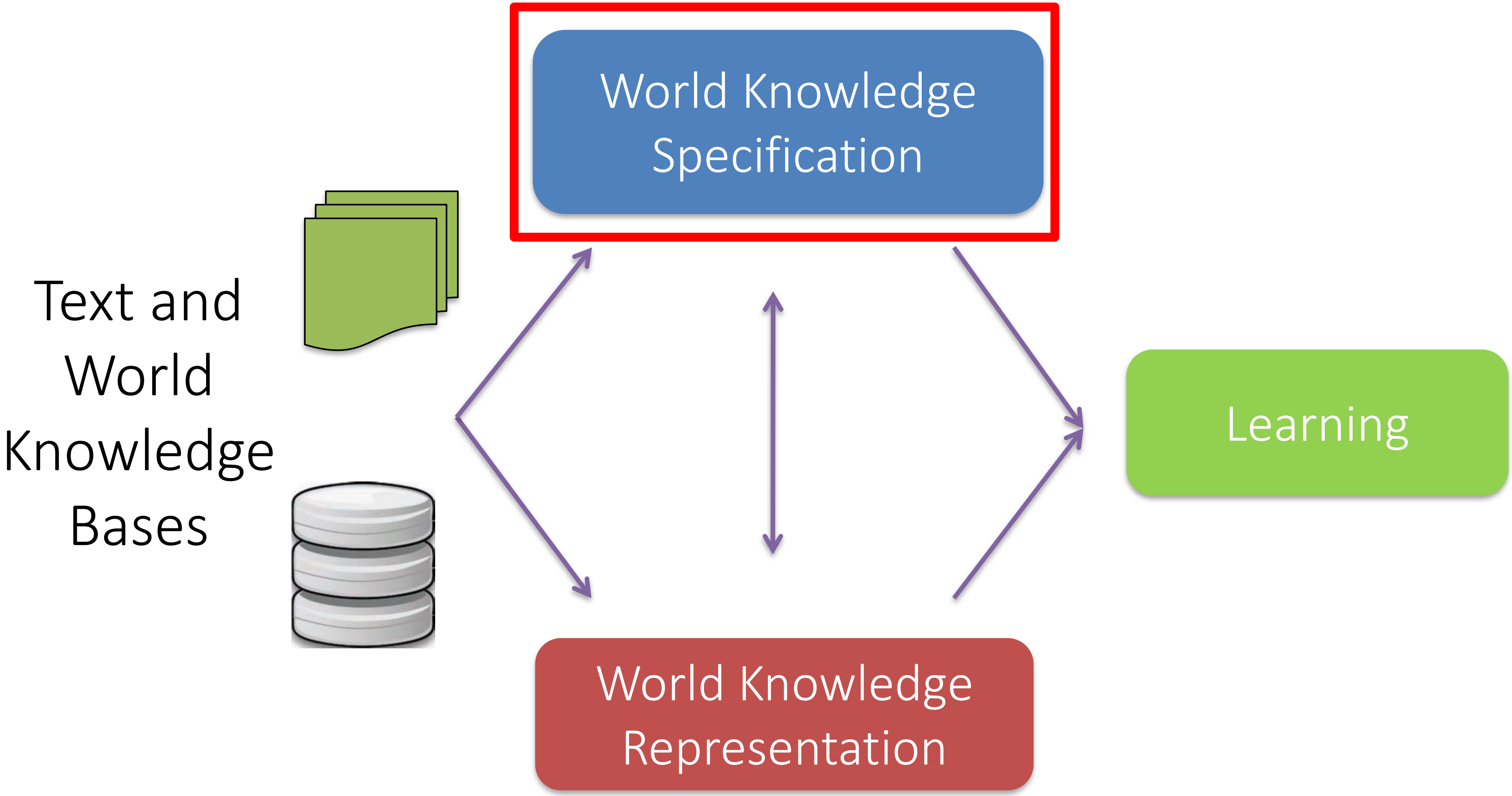
- How to convert unstructured texts to HINs?

- What can we do with the HINs?

Challenges of Using World Knowledge



Networked Text Analysis Framework


World Knowledge Specification
Text and World Knowledge Bases
Learning
World Knowledge Representation

Wang et al., Incorporating World Knowledge to Document Clustering via Heterogeneous Information Networks. KDD’15.
Wang et al. World knowledge as indirect supervision for document clustering. TKDD’16.

World Knowledge Specification: Unsupervised Semantic Parsing for Documents

Document Obama is the president of the United States of America

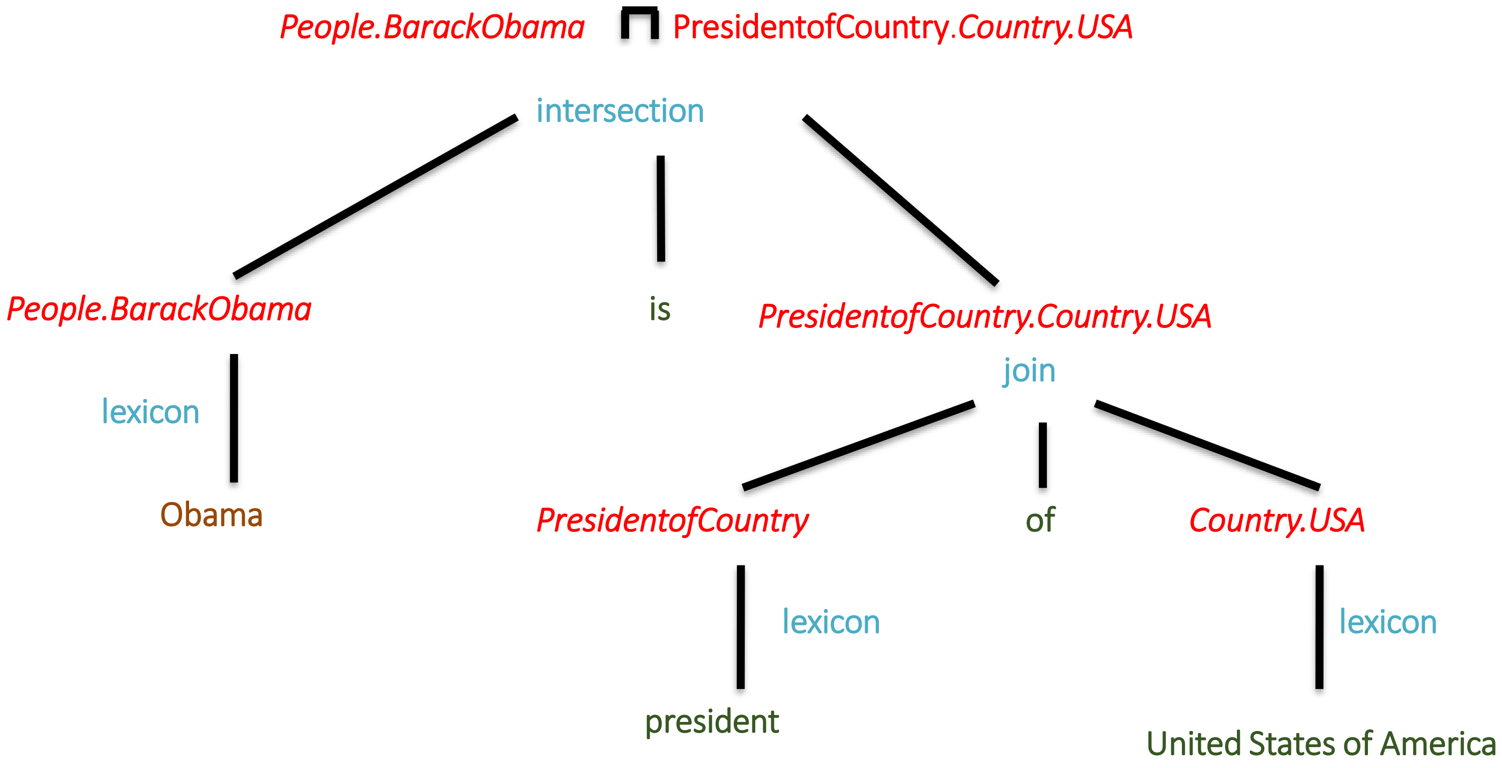
Semantic parsing is the task of mapping a piece of natural language text to a formal meaning representation.

Logic form *People.BarackObama* ⊓ PresidentofCountry.*Country.USA*

- Motivation: [Berant et al. EMNLP'13] aim to train a parser from question/answer pairs on a large knowledge-base Freebase
  - Existing semantic parsing approaches, that require expert annotation
  - Scales to large scale knowledge-bases, supervised by the QA pairs
- No such training data for the document dataset.

World Knowledge Specification:
Unsupervised Semantic Parsing for Documents

Document Obama is the president of the United States of America



People.BarackObama ⊓ PresidentofCountry.Country.USA

intersection

People.BarackObama

is

PresidentofCountry.Country.USA

lexicon

join

Obama

PresidentofCountry

of

Country.USA

lexicon

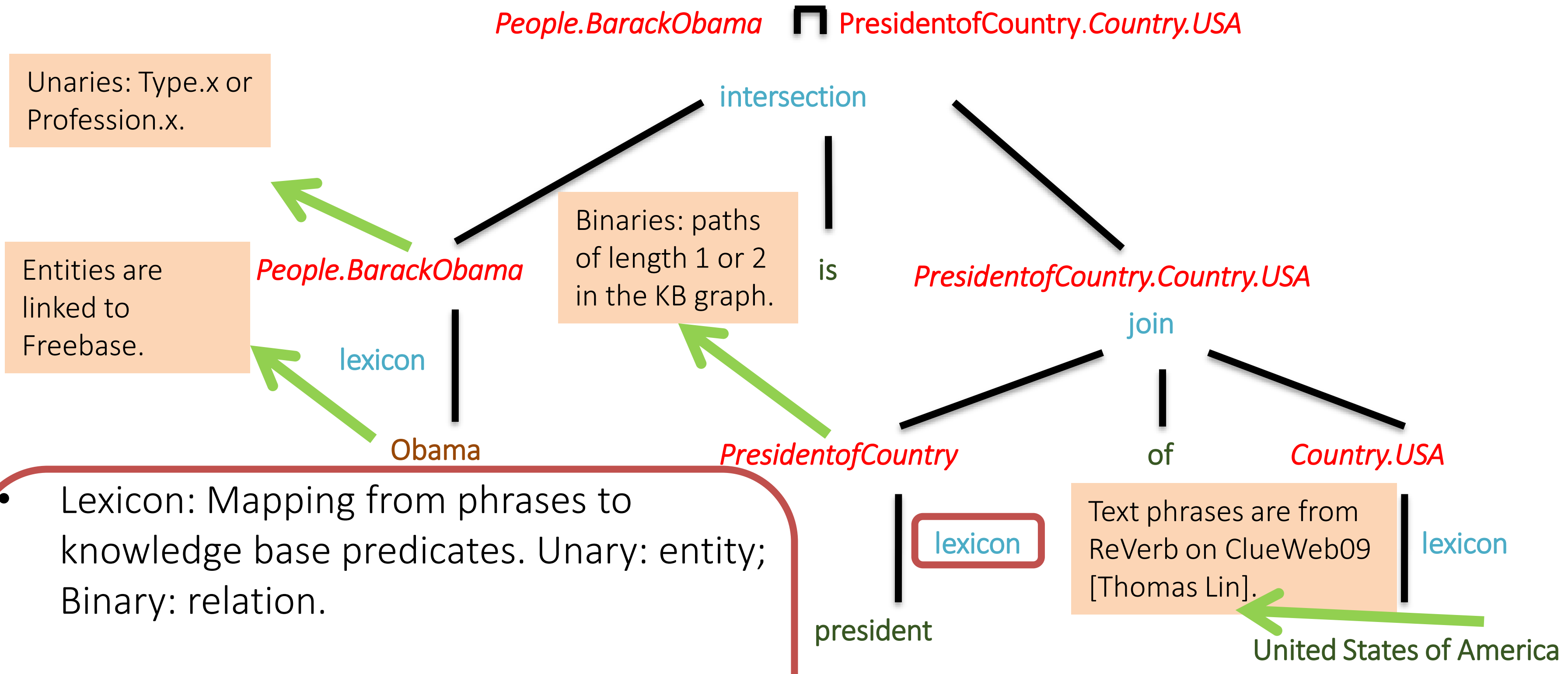
lexicon

president

United States of America

World Knowledge Specification: Unsupervised Semantic Parsing for Documents

Document Obama is the president of the United States of America



Unaries: Type.x or Profession.x.

Entities are linked to Freebase.

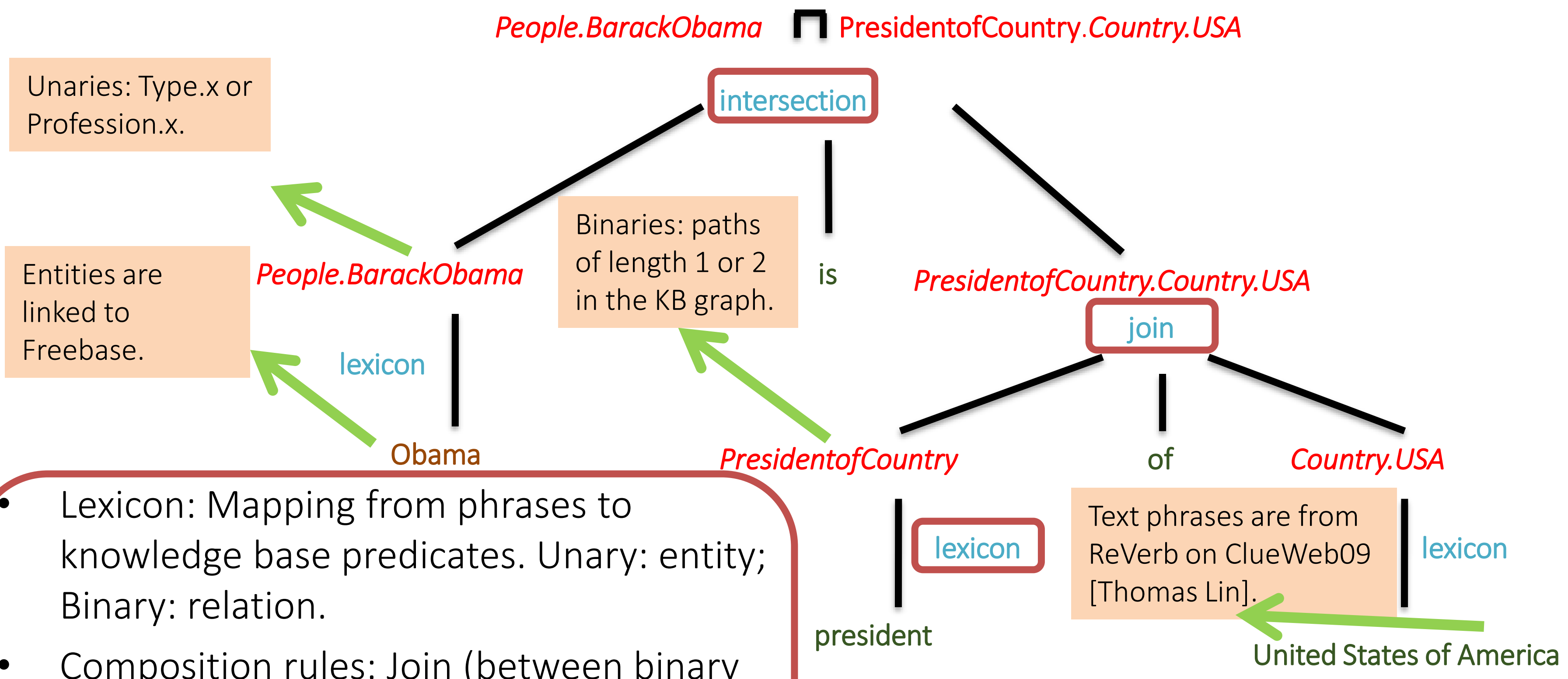
Binaries: paths of length 1 or 2 in the KB graph.

Text phrases are from ReVerb on ClueWeb09 [Thomas Lin].

- Lexicon: Mapping from phrases to knowledge base predicates. Unary: entity; Binary: relation.

World Knowledge Specification: Unsupervised Semantic Parsing for Documents

Document   Obama is the president of the United States of America



Unaries: Type.x or Profession.x.

Entities are linked to Freebase.

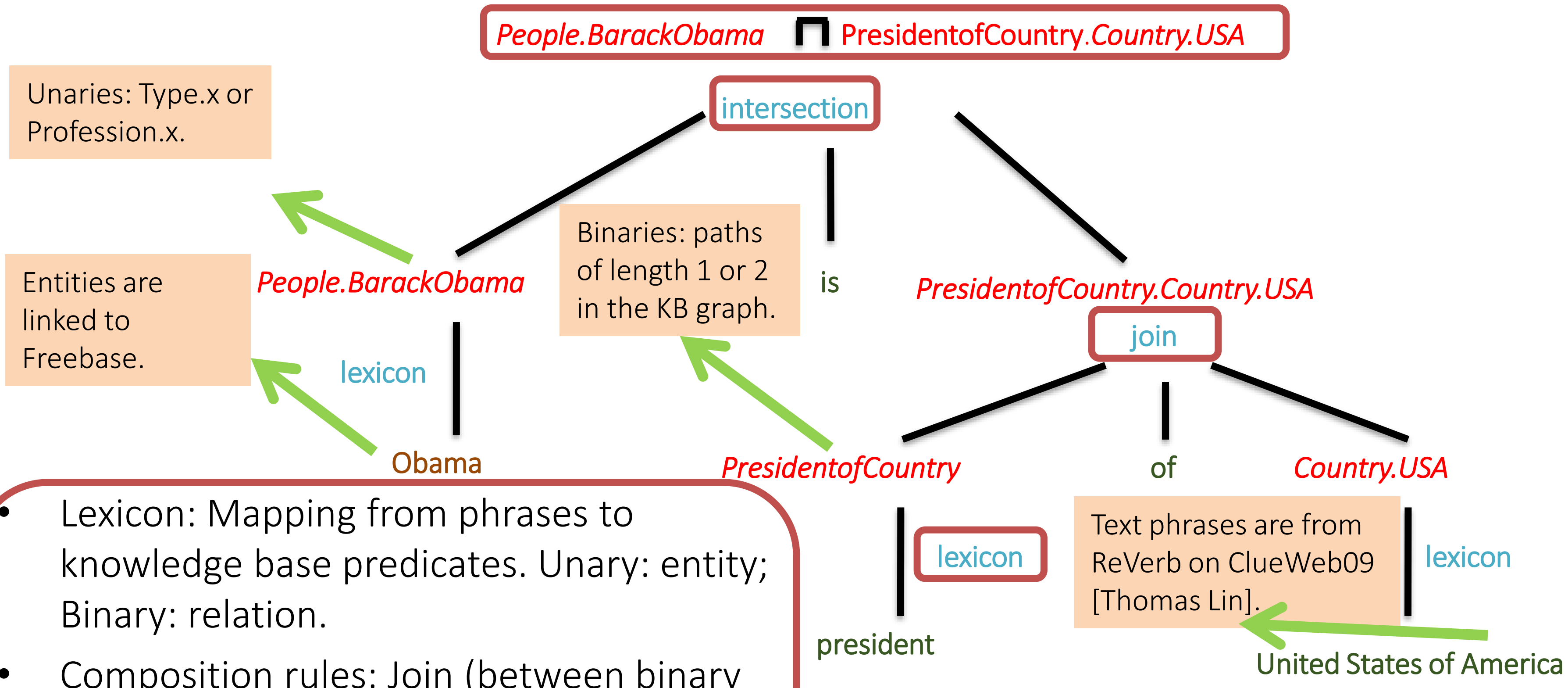
Binaries: paths of length 1 or 2 in the KB graph.

Text phrases are from ReVerb on ClueWeb09 [Thomas Lin].

- Lexicon: Mapping from phrases to knowledge base predicates. Unary: entity; Binary: relation.
- Composition rules: Join (between binary and unary); Intersection (between unary and unary).

World Knowledge Specification: Unsupervised Semantic Parsing for Documents

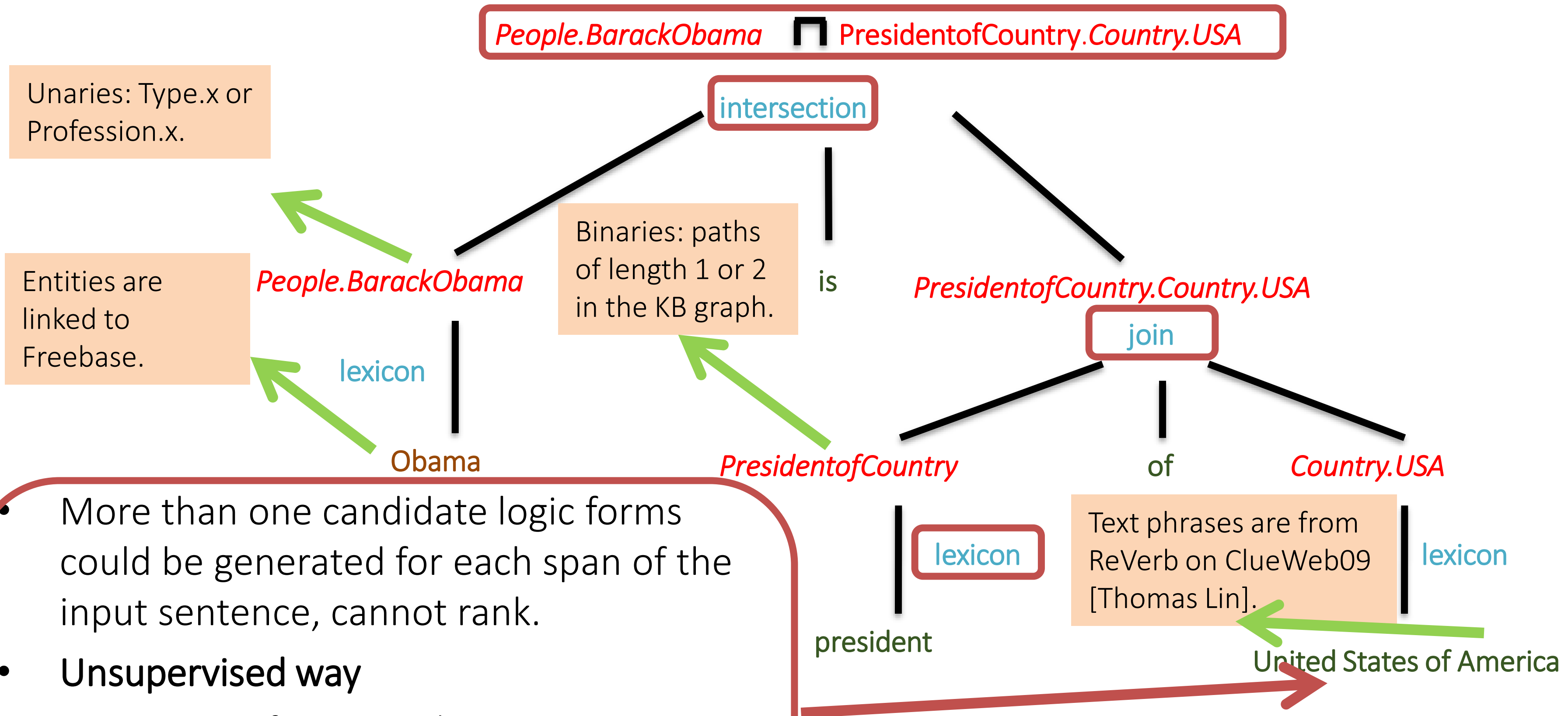
Document Obama is the president of the United States of America


People.BarackObama ⊓ PresidentofCountry.Country.USA
intersection
Unaries: Type.x or Profession.x.
People.BarackObama
Binaries: paths of length 1 or 2 in the KB graph.
is
PresidentofCountry.Country.USA
join
Entities are linked to Freebase.
lexicon
Obama
PresidentofCountry
of
Country.USA
lexicon
Text phrases are from ReVerb on ClueWeb09 [Thomas Lin].
lexicon
president
United States of America

- Lexicon: Mapping from phrases to knowledge base predicates. Unary: entity; Binary: relation.
- Composition rules: Join (between binary and unary); Intersection (between unary and unary).
- Logic form construction: based on lexicon and composition rules recursively.

World Knowledge Specification: Unsupervised Semantic Parsing for Documents

Document Obama is the president of the United States of America



People.BarackObama ⊓ PresidentofCountry.*Country.USA*

intersection

Unaries: Type.x or Profession.x.

Entities are linked to Freebase.

People.BarackObama

Binaries: paths of length 1 or 2 in the KB graph.

is

PresidentofCountry.Country.USA

join

lexicon

Obama

PresidentofCountry

of

Country.USA

lexicon

Text phrases are from ReVerb on ClueWeb09 [Thomas Lin].

lexicon

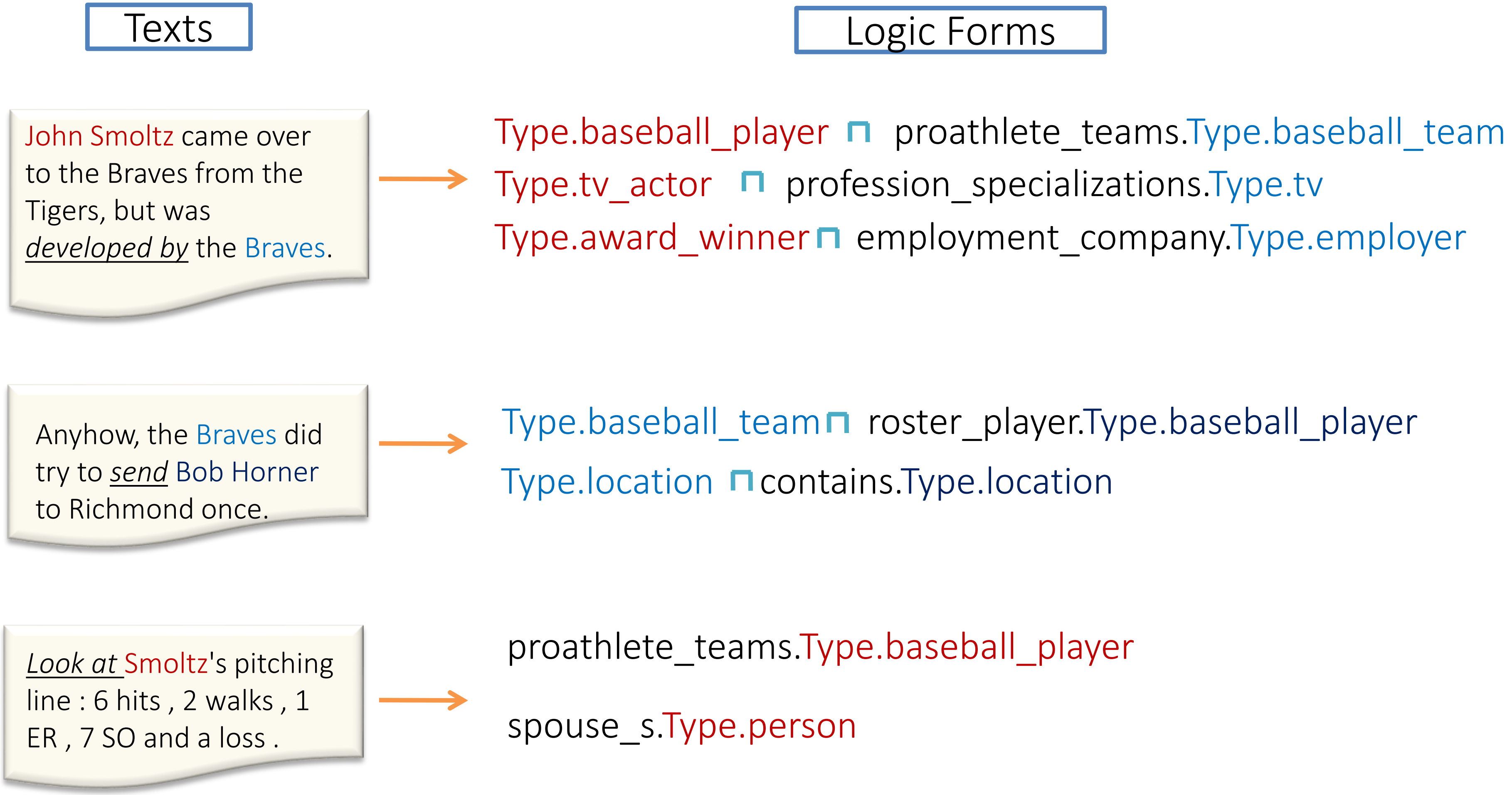
president

United States of America

- More than one candidate logic forms could be generated for each span of the input sentence, cannot rank.
- **Unsupervised way**
  - A state-of-art named entity recognition tool [L. Ratinov et al. CoNLL 2009] is used to find only maximum spanning phrase.
  - Only generate partial immediate logic form based on the maximum spanning phrase.

NOT ``America’’ or ``United States’’

Examples of Semantic Parsing on 20-NG



Texts

Logic Forms

John Smoltz came over to the Braves from the Tigers, but was *developed by* the Braves.

→

Type.baseball_player ⊓ proathlete_teams.Type.baseball_team

Type.tv_actor ⊓ profession_specializations.Type.tv

Type.award_winner ⊓ employment_company.Type.employer

Anyhow, the Braves did try to *send* Bob Horner to Richmond once.

→

Type.baseball_team ⊓ roster_player.Type.baseball_player

Type.location ⊓ contains.Type.location

Look at Smoltz's pitching line : 6 hits , 2 walks , 1 ER , 7 SO and a loss .

→

proathlete_teams.Type.baseball_player

spouse_s.Type.person

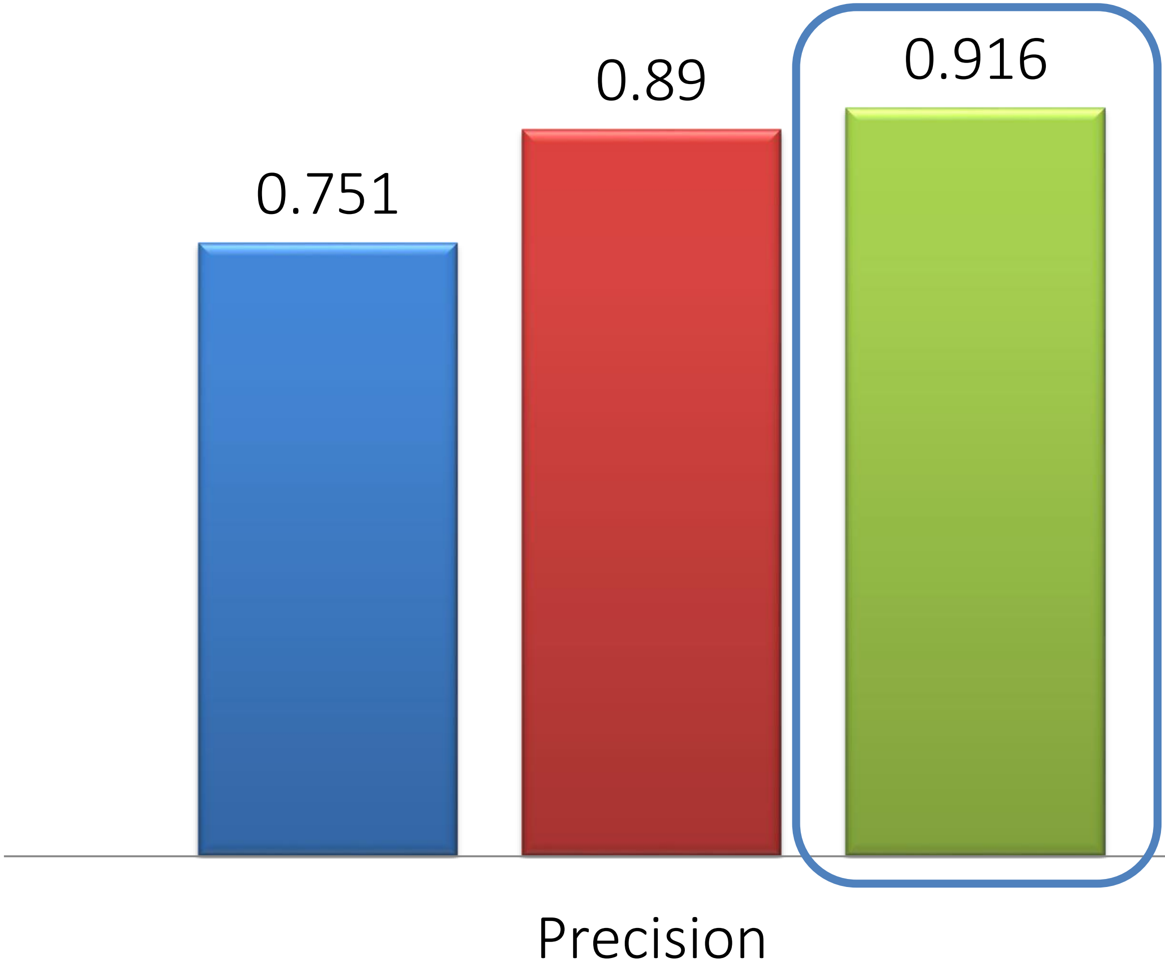
Some of the forms are not noisy results

World Knowledge Specification: Semantic Filtering

- Term frequency based semantic filtering (FBSF)
  - How many times a type appearing **in a document**
- Document frequency based semantic filtering (DFBSF)
  - How many documents a type appearing in, **in a corpus**
- Conceptualization based semantic filter (CBSF)
  - **Clustering the same entity** (with different mentions) based on their types
  - In each cluster, use the most frequent type for the mentions

Song et al., Open Domain Short Text Conceptualization: A Generative + Descriptive Modeling Approach. IJCAI'15.
Song et al., Short Text Conceptualization using a Probabilistic Knowledgebase. IJCAI'11.

Precision of Different Semantic Filtering



🟦 FBSF

Frequency based semantic filter.
Type is decided by the counts in one document.

🟥 DFBSF

Document frequency based semantic filter.
Type is decided by the counts in whole document set.

🟩 CBSF

Conceptualization based semantic filter.
Type is decided by the context in whole document set.

Wang et al., Incorporating World Knowledge to Document Clustering via Heterogeneous Information Networks. KDD'15.
Wang et al., World knowledge as indirect supervision for document clustering. TKDD'16.

Examples of Semantic Filtering on 20NG

John Smoltz came over to the Braves from the Tigers, but was *developed by* the Braves.

→

Type.baseball_player $\sqcap$ proathlete_teams.Type.baseball_team

Type.tv_actor $\sqcap$ profession_specializations.Type.tv

Type.award_winner $\sqcap$ employment_company.Type.employer

Anyhow, the Braves did try to *send* Bob Horner to Richmond once.

→

Type.baseball_team $\sqcap$ roster_player.Type.baseball_player

Type.location $\sqcap$ contains.Type.location

Look at Smoltz's pitching line : 6 hits , 2 walks , 1 ER , 7 SO and a loss .

→

proathlete_teams.Type.baseball_player

spouse_s.Type.person

John Smoltz:**Type.baseball_player**

Braves:**Type.baseball_team**

Error Analysis of Semantic Filtering

Type of error	Example sentence	Number and percentage of errors		
		FBSF (805)	DFBSF (359)	CBSF (272)
Entity Recognition	“Einstein ’s theory of relativity explained mercury ’s motion.”	179 (22.2%)	129 (35.9%)	105 (38.6%)
Entity Disambiguation	“Bill said all this to make the point that Christianity is eminently.”	537 (66.7%)	182 (50.7%)	130 (47.8%)
Subordinate Clause	“Bruce S. Winters, worked at United States Technologies Research Center, bought a Ford.”	89 (11.1%)	48 (13.4%)	37 (13.6%)

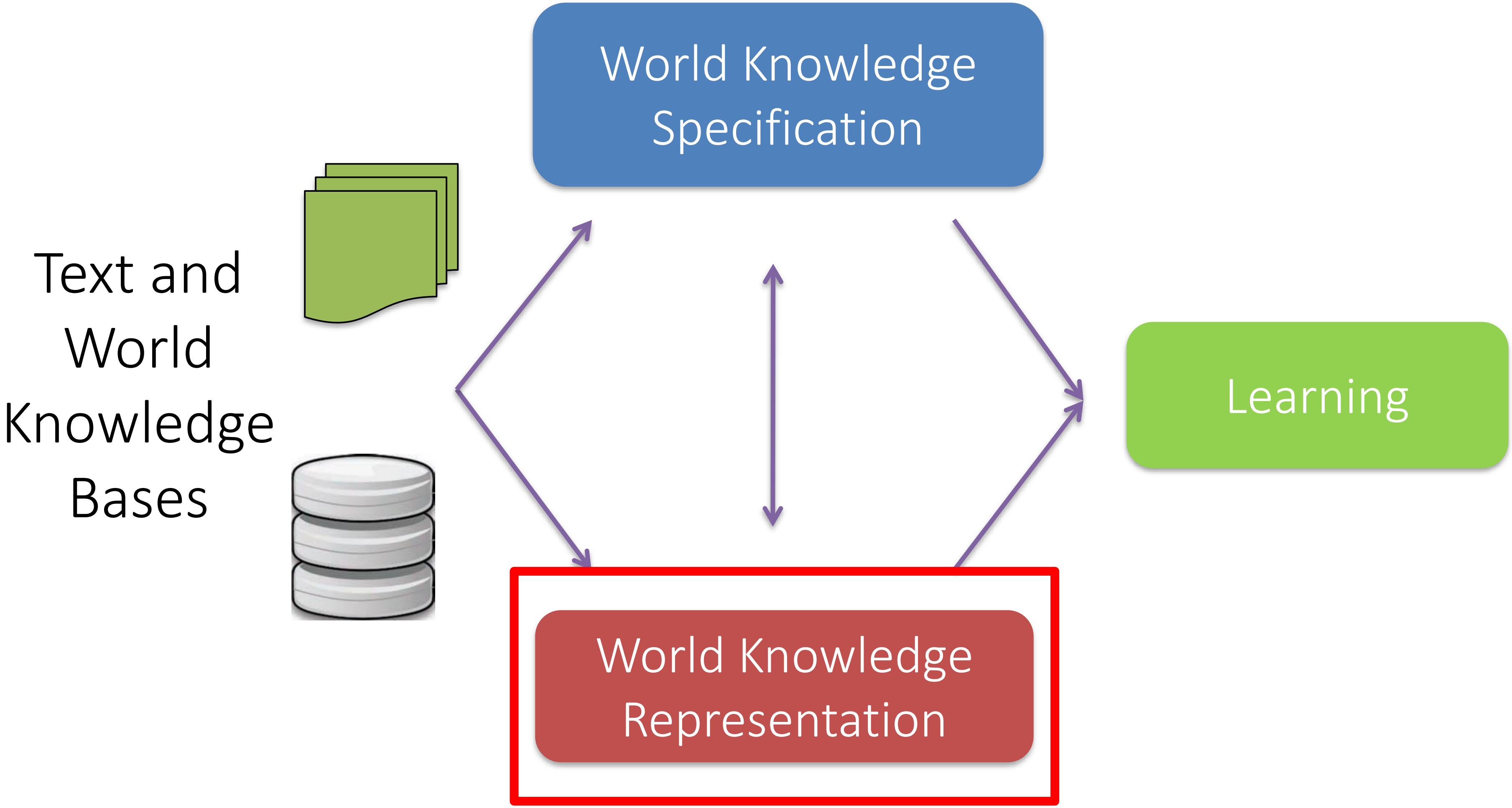
Finding #1: Entity disambiguation is the major error factor.

Entity disambiguation is a tough research problem in NLP community. The type information of relations are not sufficient to further prune out mismatching entities during semantic filtering process.

Finding #2: CBSF performs the best.

For example, by using context, the number of incorrect entities caused by disambiguation can be dramatically reduced.

Networked Text Analysis Framework



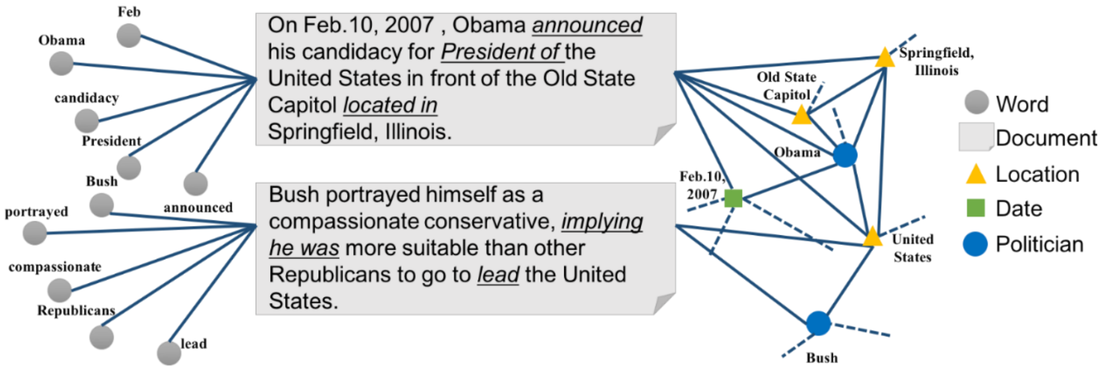
Text and World Knowledge Bases

World Knowledge Specification

World Knowledge Representation

Learning

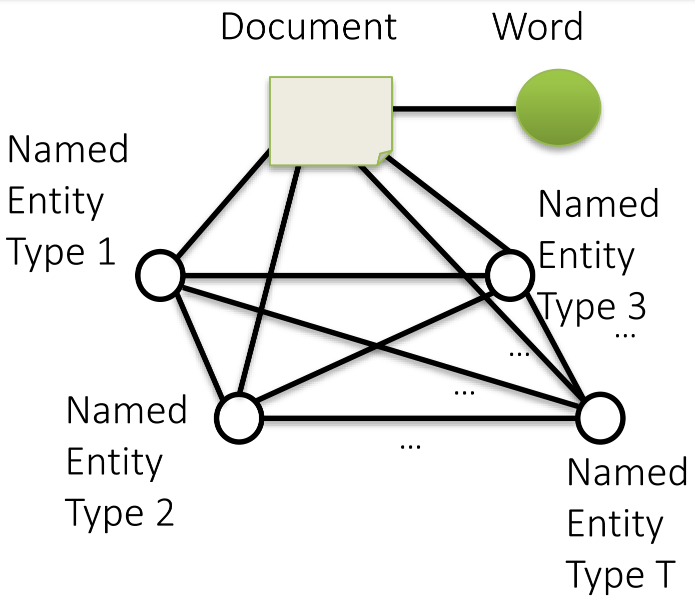
World Knowledge Representation： Heterogeneous Information Network (HIN)

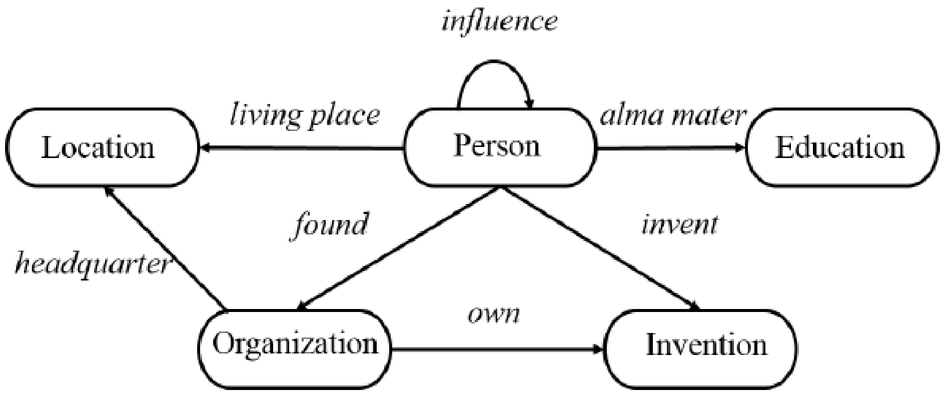


On Feb.10, 2007 , Obama *announced* his candidacy for *President of* the United States in front of the Old State Capitol *located in* Springfield, Illinois.

Bush portrayed himself as a compassionate conservative, *implying he was* more suitable than other Republicans to go to *lead* the United States.

HIN **network-schema**: network with multiple object types and/or multiple link types.



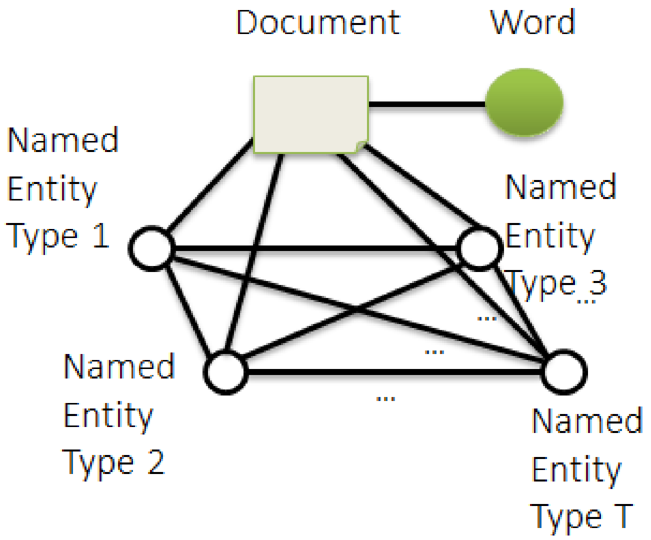


Outline

- Motivation
  - Two Challenges
    - Representation
    - Labels
- Text Categorization via HIN
  - HIN construction from texts
  - From HIN similarity to clustering and classification
  - World knowledge indirect supervision
- Conclusions and future work

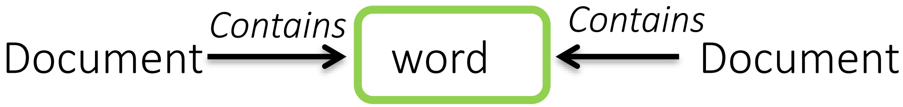
Meta-path, Commuting Matrix, and PathSim

- Meta-path path defined over the network schema.
  - [Sun et al., 2011]



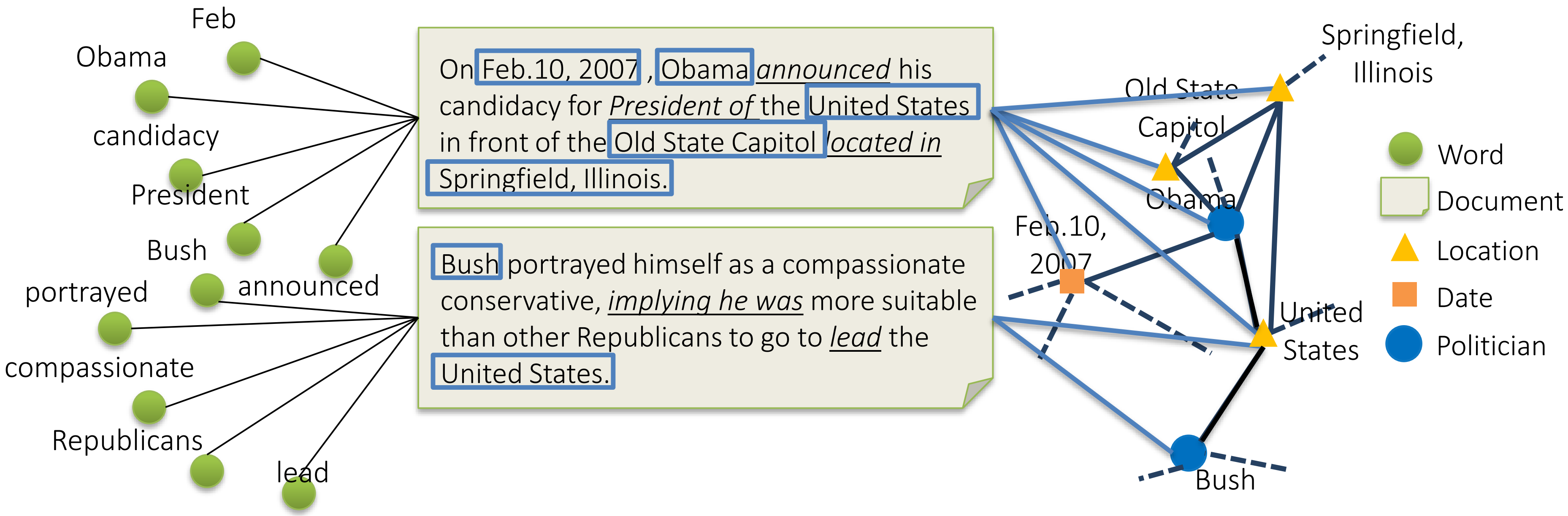
- Commuting matrix:
  - e.g., document->word binary occurrence matrix: W

- PathSim



  - e.g.,
  - $W^T W$: dot product

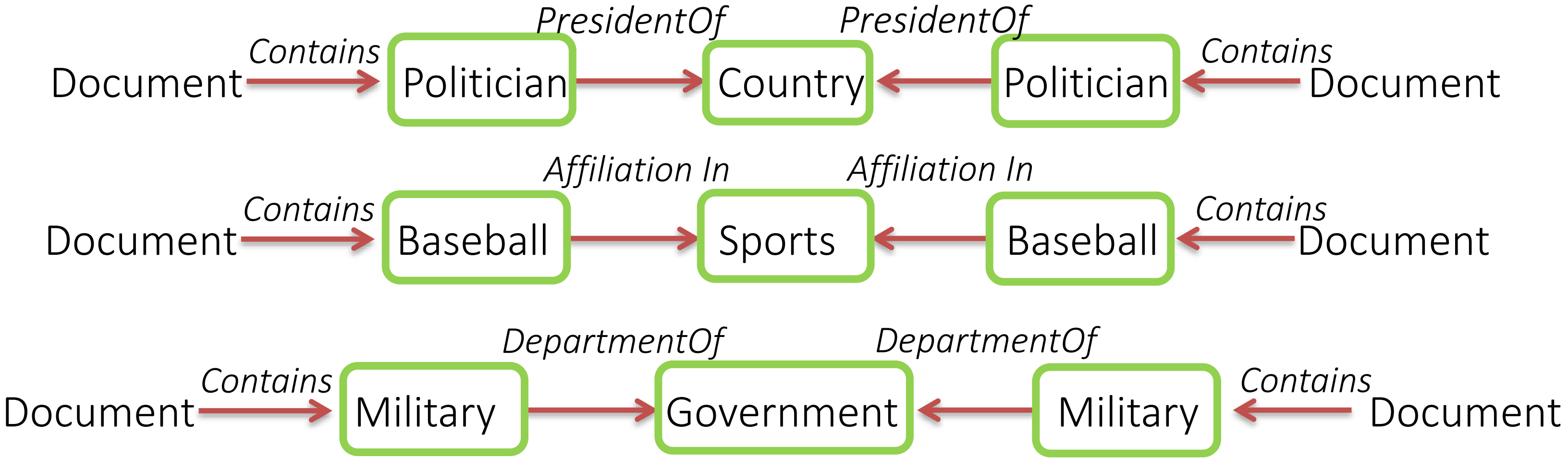
Other Meta-paths in Text HIN



On Feb.10, 2007 , Obama *announced* his candidacy for *President of* the United States in front of the Old State Capitol *located in* Springfield, Illinois.

Bush portrayed himself as a compassionate conservative, *implying he was* more suitable than other Republicans to go to *lead* the United States.

Capturing higher-order relations



Document $\xrightarrow{Contains}$ Politician $\xrightarrow{PresidentOf}$ Country $\xleftarrow{PresidentOf}$ Politician $\xleftarrow{Contains}$ Document

Document $\xrightarrow{Contains}$ Baseball $\xrightarrow{Affiliation\ In}$ Sports $\xleftarrow{Affiliation\ In}$ Baseball $\xleftarrow{Contains}$ Document

Document $\xrightarrow{Contains}$ Military $\xrightarrow{DepartmentOf}$ Government $\xleftarrow{DepartmentOf}$ Military $\xleftarrow{Contains}$ Document

KnowSim

An ensemble of similarity measures defined on structured HIN.

Semantic overlap: the number of meta-paths between two documents.

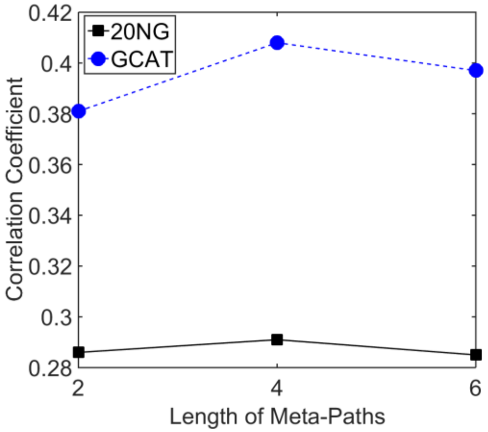
$$KS(d_i,d_j)=\frac{2\times\boxed{\sum_m^{M'} w_m\,|\{p_{i\to j}\in P_m\}|}}{\boxed{\sum_m^{M'} w_m\,|\{p_{i\to i}\in P_m\}|+\sum_m^{M'} w_m\,|\{p_{j\to j}\in P_m\}|}}$$

Semantic broadness: the number of total meta-paths between themselves.

- Intuition: The larger number of highly weighted meta-paths between two documents, the more similar these documents are, which is further normalized by the semantic broadness.

- KnowSim is computed in nearly linear time.

Wang et al., KnowSim: A Document Similarity Measure on Structured Heterogeneous Information Networks. ICDM’15.



Challenges

# of meta-paths:
20NG (325) GCAT (1,682)

Number of meta-paths could be very large.

$$KS(d_i, d_j) = \frac{2 \times \sum_m^{\boxed{M'}} w_m \,|\{p_{i\to j} \in P_m\}|}{\sum_m^{M'} \boxed{w_m}\,|\{p_{i\to i} \in P_m\}| + \sum_m^{M'} w_m \,|\{p_{j\to j} \in P_m\}|}$$

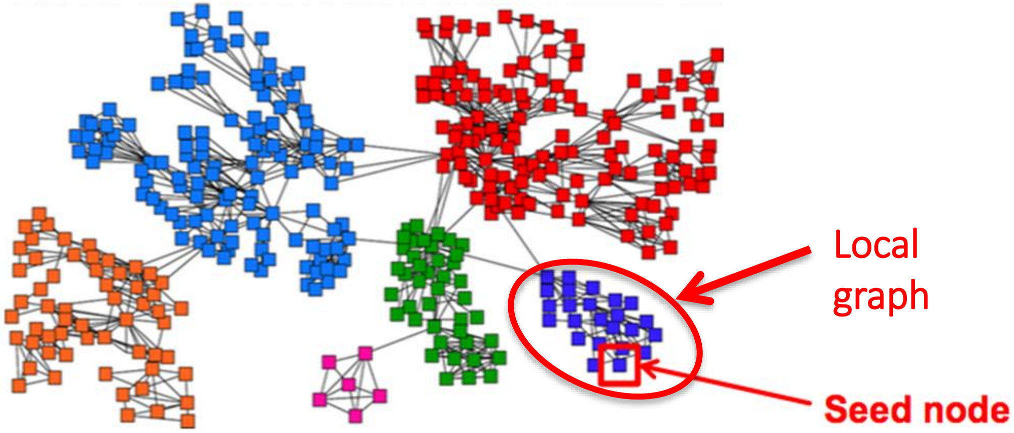
The weight/importance of each meta-path is different when the domain is different.

#1: How should we generate the large number of meta-paths at the same time?
Previous studies only focus on single meta-path, enumeration over the network is OK. In real world, what will happen when thousands of meta-paths are needed?

#2: How should we decide the weight of each meta-path?
Previous studies treat them equally. In real world, different meta-path should contribute differently in various domains.

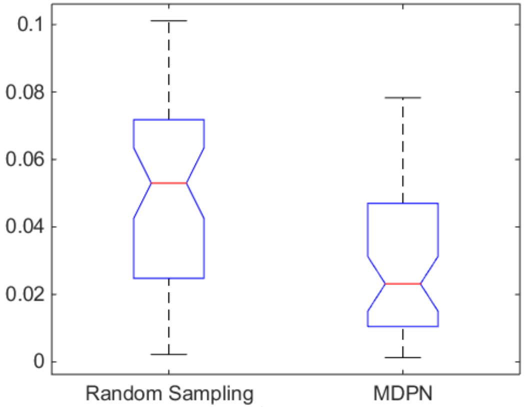
Meta-Path Dependent Random Walk

Intuition: Discovering compact sub-graph based on seed document nodes.



- Compute **Personalized PageRank (PPR)** around seed nodes.
- The random walk will get trapped inside the blue sub-graph.

- Algorithm outline
  - Run PPR (approximate connectivity to seed nodes) with teleport set = {S}
  - Sort the nodes by the decreasing PPR score
  - **Sweep** over the nodes and find compact **sub-graph**.
  - Use the sub-graph instead of the whole graph to compute # of meta-paths between nodes.



Frobenius norm of approximation of commuting matrices on 20NG dataset

Meta-Path Ranking

# of meta-paths: **20NG (325)** and **GCAT (1,682)**

- Maximal Spanning Tree based Selection [Sahami, 1998]

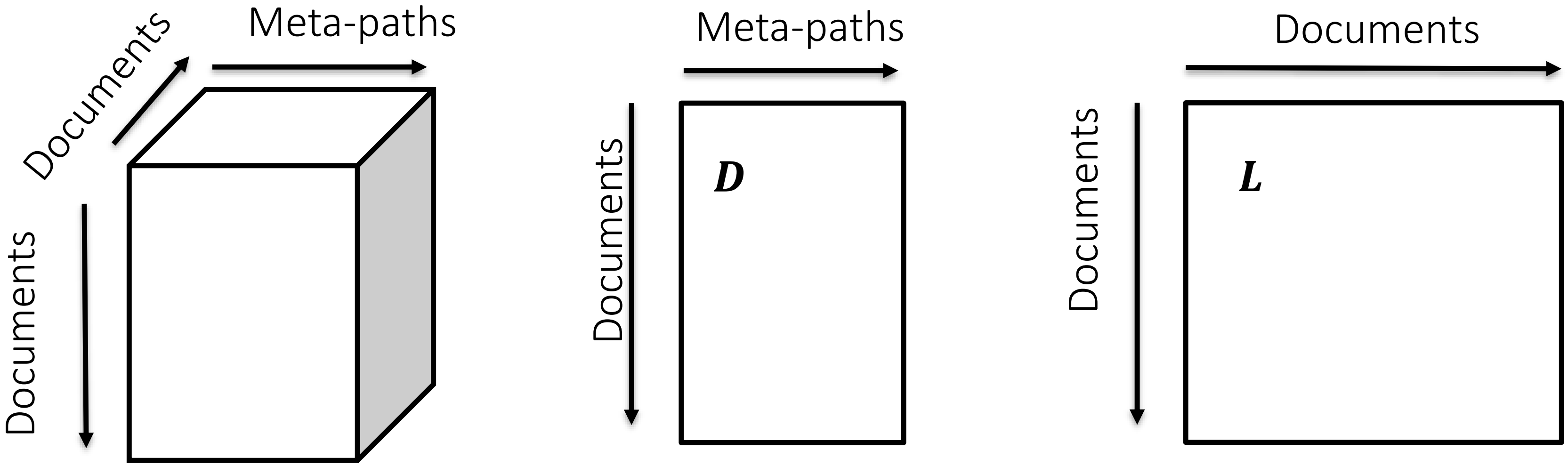
$$\frac{\sum_{j\neq i}^{M}\cos(\boldsymbol{D}_{.,j_1},\boldsymbol{D}_{.,j_2})}{M-1}$$

Select meta-paths with the largest dependencies with others

- Laplacian Score based Selection [He, 2006]

$$L_j = \frac{\widetilde{\boldsymbol{D}_{.,j}}^{T}\boldsymbol{L}\boldsymbol{D}_{.,j}}{\widetilde{\boldsymbol{D}_{.,j}}^{T}\Lambda\boldsymbol{D}_{.,j}}$$

Select a meta-path in **discriminating documents** from different clusters



Experiments

Document datasets			
Name	#(Categories)	#(Leaf Categories)	#(Documents)
20Newsgroups (20NG)	6	20	20,000
MCAT (Markets)	9	7	44,033
CCAT (Corporate/Industrial)	31	26	47,494
ECAT (Economics)	23	18	19,813

MCAT, CCAT, ECAT are top categories in RCV1 dataset containing manually labeled newswire stories from Reuter Ltd.

World knowledge bases				
Name	#(Entity Types)	#(Entity Instances)	#(Relation Types)	#(Relation Instances)
Freebase	1,500	40 millions	35,000	2 billions
publicly available knowledge base with entities and relations collaboratively collected by its community members.				
YAGO2	350,000	10 millions	100	120 millions
a semantic knowledge base, derived from Wikipedia, WordNet and GeoNames.				

The number is reported in [X. Dong et al. KDD'14], In our downloaded dump of Freebase, we found 79 domains, 2,232 types, and 6,635 properties.

Text Similarity Results

- Evaluation: correlation with document similarity
  - In the same category: 1
  - In different categories: 0

Datasets	Similarity Measures	BOW	BOW+ TOPIC	BOW+TOPIC+ ENTITY
20NG	Cosine	0.2400	0.2713	0.2768
	Jaccard	0.2352	0.2632	0.2650
	Dice	0.2400	0.2712	0.2767
GCAT	Cosine	0.3490	0.3639	0.3128
	Jaccard	0.3313	0.3460	0.2991
	Dice	0.3490	0.3638	0.3156

	KnowSim+UNIFORM	KnowSim+MST	KnowSim+LAP
20NG	0.2860	0.2891	0.2913 (+5.2%)
GCAT	0.3815	0.3833	0.4086 (+12.3%)

Outline

- Motivation
  - Two Challenges
    - Representation
    - Labels
- Text Categorization via HIN
  - HIN construction from texts
  - From HIN similarity to clustering and classification
  - World knowledge indirect supervision
- Conclusions and future work

Spectral Clustering with KnowSim

- Non-linear clustering (Ng et al., NIPS'01)
  - Construct k-NN graph based on pair-wise similarities
  - Perform k-means over Eigen vectors of the graph Laplacian

Datasets	Similarity Measures	BOW	BOW+TOPIC	BOW+TOPIC+ENTITY
20NG	Cosine	0.3440	0.3461	0.4247
	Jaccard	0.3547	0.3517	0.4292
	Dice	0.3440	0.3457	0.4248
GCAT	Cosine	0.3932	0.4352	0.4106
	Jaccard	0.3887	0.4292	0.4159
	Dice	0.3932	0.4355	0.4112

	KnowSim+UNIFORM	KnowSim+MST	KnowSim+LAP
20NG	0.4304	0.4304	**0.4461 (+3.9%)**
GCAT	0.4463	0.4653	**0.4736(+8.8%)**

Wang et al., KnowSim: A Document Similarity Measure on Structured Heterogeneous Information Networks. ICDM'15.

SVM with Indefinite HIN-Kernel

- SVM needs a positive semi-definite(PSD) kernel matrix
- KnowSim matrix is non-PSD
- Feed the non-PSD KnowSim kernel matrix to SVM [Luss and d'Aspremont 2008']
  - Learn a proxy of non-PSD KnowSim matrix
  - Simultaneously learn a SVM classifier.

Objective function:

Penalty factor

$$\min_{\kappa} \max_{\alpha} \boxed{1^{T}\alpha - \frac{1}{2}\alpha^{T}Y^{T}KY\alpha} + \boxed{\rho} \,\|\, \boxed{K} - \boxed{K_0} \,\|_F^2$$

Proxy kernel

Indefinite kernel

s.t. $\boxed{y^{T}a = 0, 0 £ a <= C1,} \boxed{K ³ 0}$

Original SVM Objective function

PSD Proxy kernel

Wang et al., Text Classification with Heterogeneous Information Network Kernels. AAAI'16.

Classification Results

Average accuracy

Model	Discrete		Embedding
Settings	BOW	BOW+ENTITY	Word2vec
20NG-SIM	90.81%	91.11%	91.67%
20NG-DIF	96.66%	96.90%	98.27%
GCAG-SIM	94.15%	94.29	96.81%
GCAT-DIF	88.98%	90.18%	90.64%

Mikolov
2013.
Window: 5
Dim: 400

Average accuracy

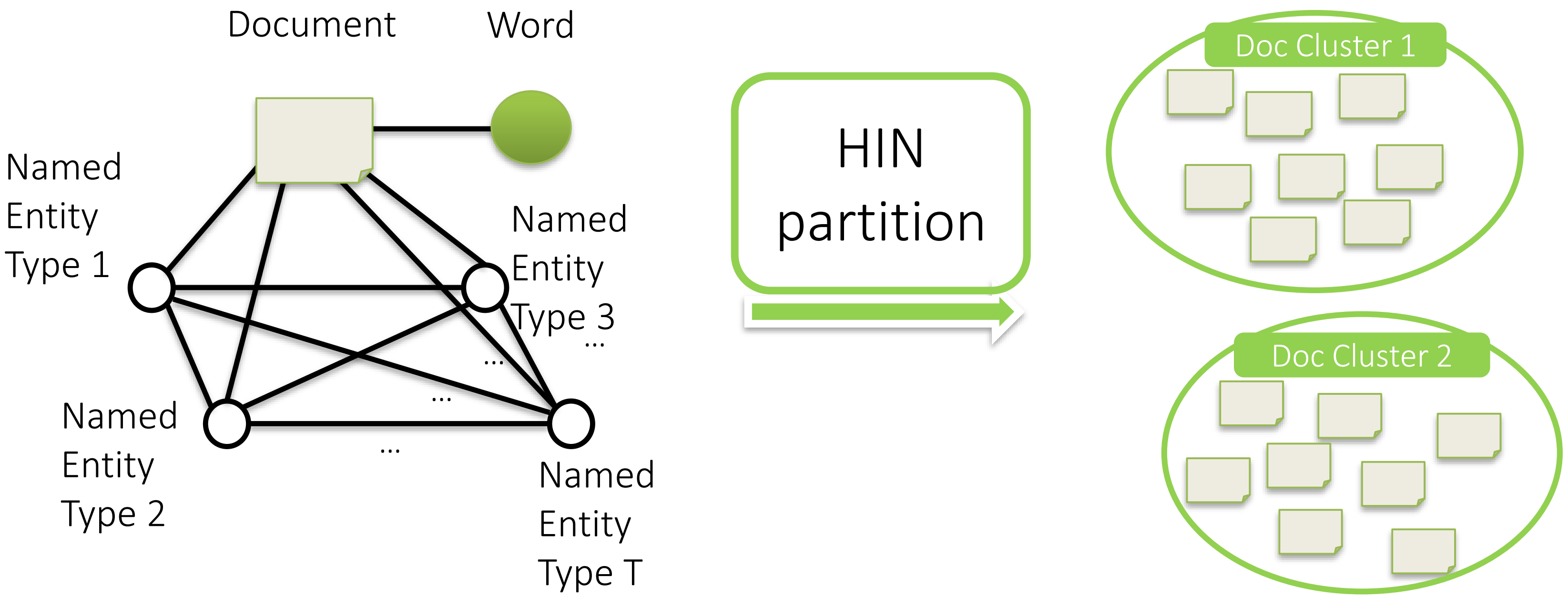
Model	SVM^{HIN}	SVM^{HIN}+KnowSim		IndefSVM^{HIN}+KnwoSim	
Settings		DWD	DWD+other MetaPaths	DWD	DWD+other MetaPaths
20NG-SIM	91.60%	92.32%	92.68%	92.65%	93.38%
20NG-DIF	97.20%	97.83%	98.01%	98.13%	98.45%
GCAG-SIM	94.82%	95.29%	96.04%	95.63%	98.10%
GCAT-DIF	91.19%	90.70%	91.88%	91.63%	93.51%

Collective classification: Lu and Gatoor 2003; Kong et al. 2012

Outline

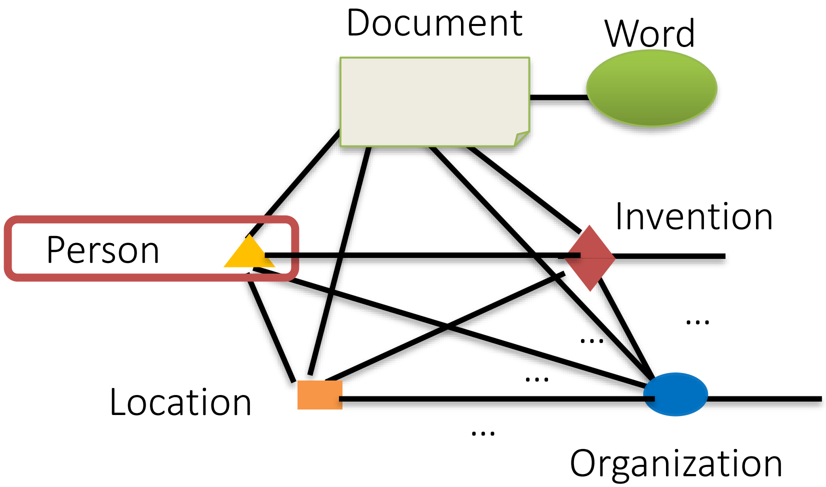
- Motivation
  - Two Challenges
    - Representation
    - Labels
- Text Categorization via HIN
  - HIN construction from texts
  - From HIN similarity to clustering and classification
  - World knowledge indirect supervision
- Conclusions and future work

HIN Constrained Clustering Modeling

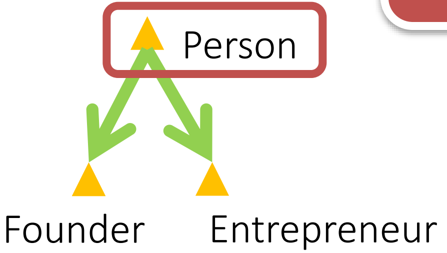


Wang et al., Incorporating World Knowledge to Document Clustering via Heterogeneous Information Networks. KDD'15.
Wang et al. World knowledge as indirect supervision for document clustering. TKDD'16.

HIN Constrained Clustering Modeling



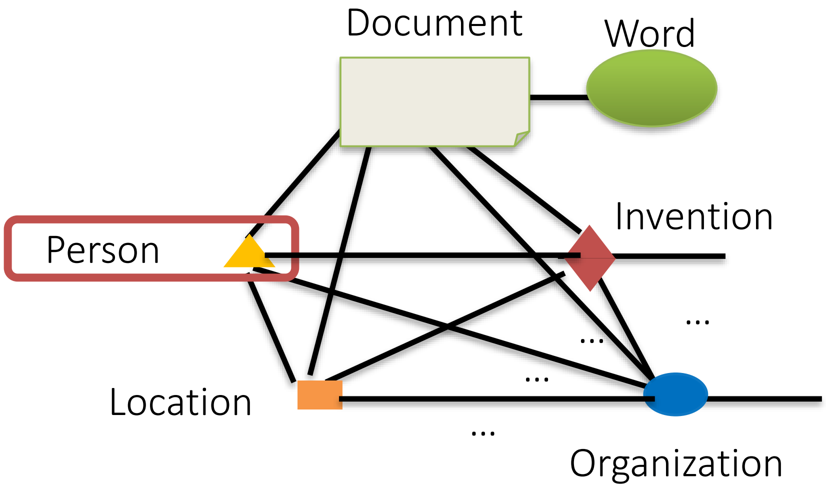
Named entity type hierarchy



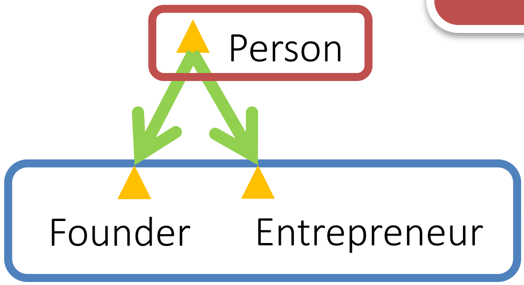
Larry Page

- Use the top level named entity types as the entity types in HIN.
  - have a relatively dense graph.

HIN Constrained Clustering Modeling

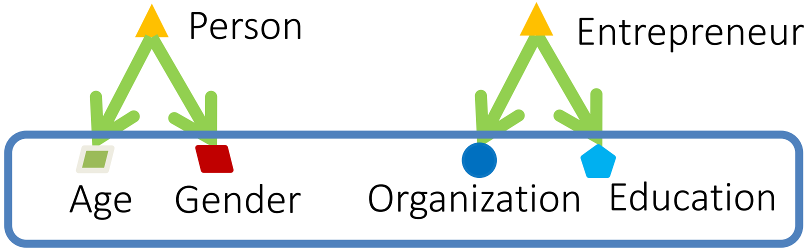


Named entity type hierarchy



Larry Page

Named entity sub-types

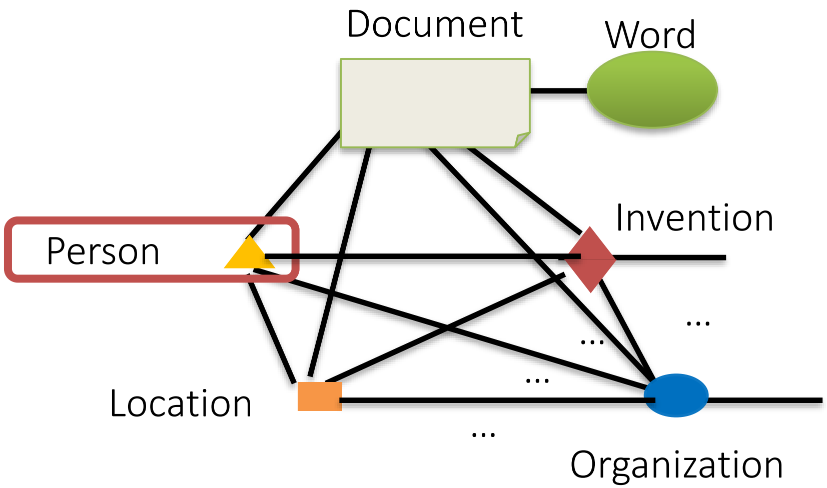


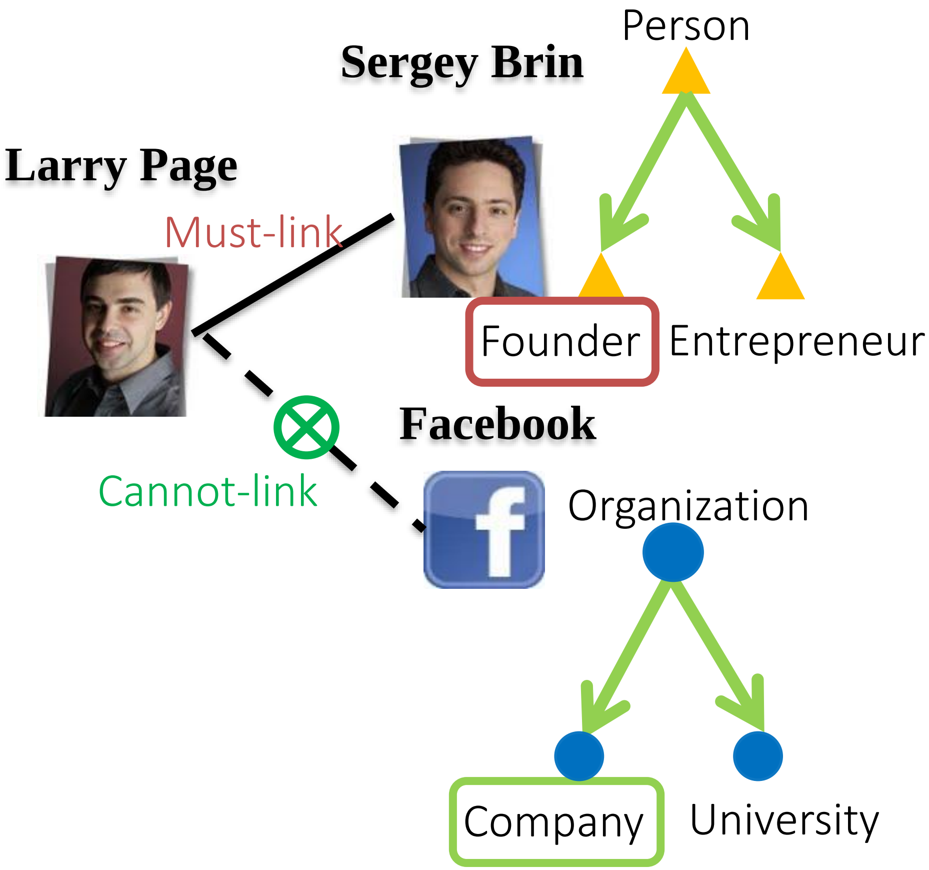
Attributes of named entity type

- Use the top level named entity types as the entity types in HIN.
  - have a relatively dense graph.
- Use named entity sub-types and attributes in HIN clustering model.
  - Useful to identify the topics or clusters of the documents.

HIN Constrained Clustering Modeling

Extend the framework of **information-theoretic co-clustering** (ITCC) [I. S. Dhillon et al. KDD'03] and constrained ITCC [Y. Song et al. TKDE'13].





- Use the top level named entity types as the entity types in HIN.
  - have a relatively dense graph.
- Use named entity sub-types and attributes in HIN clustering model.
  - Useful to identify the topics or clusters of the documents.

Song et al. Constrained Co-clustering with Unsupervised Constraints for Text Analysis. TKDE, 2013

HIN Constrained Clustering Modeling

For documents and words, factorize $q(d_m, w_i) = p(\boxed{\hat{d}_{k_d}}, \hat{w}_{k_w}) p(d_m | \hat{d}_{k_d}) p(w_i | \hat{w}_{\boxed{k_w}})$

Cluster indicators

Cluster indices

$$
\begin{aligned}
J_{CHINC} &= D_{KL}(p(D,W)||q(D,W)) \\
&+ \sum_{t=1}^{T} D_{KL}(p(D,E^t)||q(D,E^t)) \\
&+ \sum_{t=1}^{T} \sum_{s=1}^{T} D_{KL}(p(E^t,E^s)||q(E^t,E^s)) \\
&+ \sum_{t=1}^{T} \sum_{e_{i_1}^t=1}^{V_t} \sum_{e_{i_2}^t \in M_{e_{i_1}^t}} V_M\left(e_{i_1}^t, e_{i_2}^t \in M_{e_{i_1}^t}\right) \\
&+ \sum_{t=1}^{T} \sum_{e_{i_1}^t=1}^{V_t} \sum_{e_{i_2}^t \in C_{e_{i_1}^t}} V_C\left(e_{i_1}^t, e_{i_2}^t \in C_{e_{i_1}^t}\right)
\end{aligned}
$$

Minimizing KL means **approximation** q should be **similar to original** p.

Entity sub-type Must-links Cannot-links

$$\boxed{V_M\left(e_{i_1}^t, e_{i_2}^t \in M_{e_{i_1}^t}\right) = w_M D_{KL}\left(p(D|e_{i_1}^t)||p(D|e_{i_2}^t)\right) \cdot I_{l_{e_{i_1}^t} \neq l_{e_{i_2}^t}}}$$

$$\boxed{V_C\left(e_{i_1}^t, e_{i_2}^t \in C_{e_{i_1}^t}\right) = w_C(D_{max}^t - D_{KL}\left(p(D|e_{i_1}^t)||p(D|e_{i_2}^t)\right)) \cdot I_{l_{e_{i_1}^t} \neq l_{e_{i_2}^t}}}$$

Clustering Algorithm

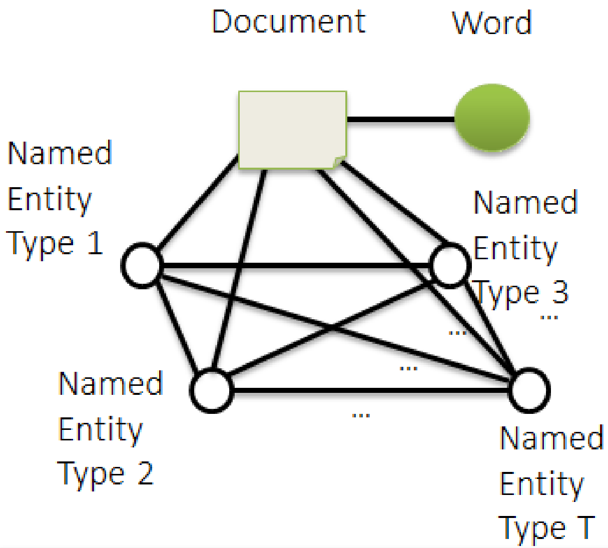
```
Algorithm: Alternating Optimization
Input: HIN defined on documents D, words W, entities $E^t, t = 1, \ldots, T$, Set maxIter and maxδ.
while iter < maxIter and δ > maxδ do
    D Label Update: minimize $J_{CHINC}$ w.r.t. $L_d$.
    D Model Update: update $q(d_m, w_i)$ and $q(d_m, e_i^t)$.

    for t = 1,…,T do                    Constrained by sub-types
        $E^t$ Label Update: minimize $J_{CHINC}$ w.r.t. $L_{e^t}$.
        $E^t$ Model Update: update $q(d_m, e_i^t)$ and $q(e_j^s, e_i^t)$.
    end for

    D Label Update: minimize $J_{CHINC}$ w.r.t. $L_d$.
    D Model Update: update $q(d_m, w_i)$ and $q(d_m, e_i^t)$.

    W Label Update: minimize $J_{CHINC}$ w.r.t. $L_w$.
    W Model Update: update$q(d_m, w_i)$.

    Compute cost change δ.
end while
```

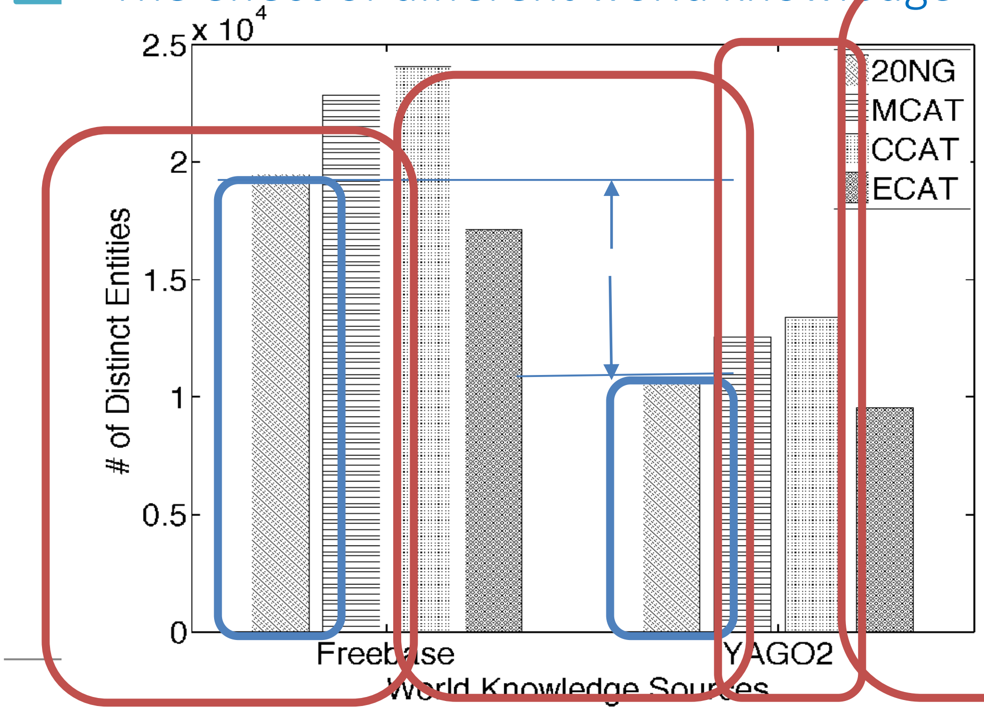


Knowledge indirect supervision: sub-types or attributes *cannot directly affect the document labels*. Constraints affect entity labels, entity labels will be transferred to affect the document labels.

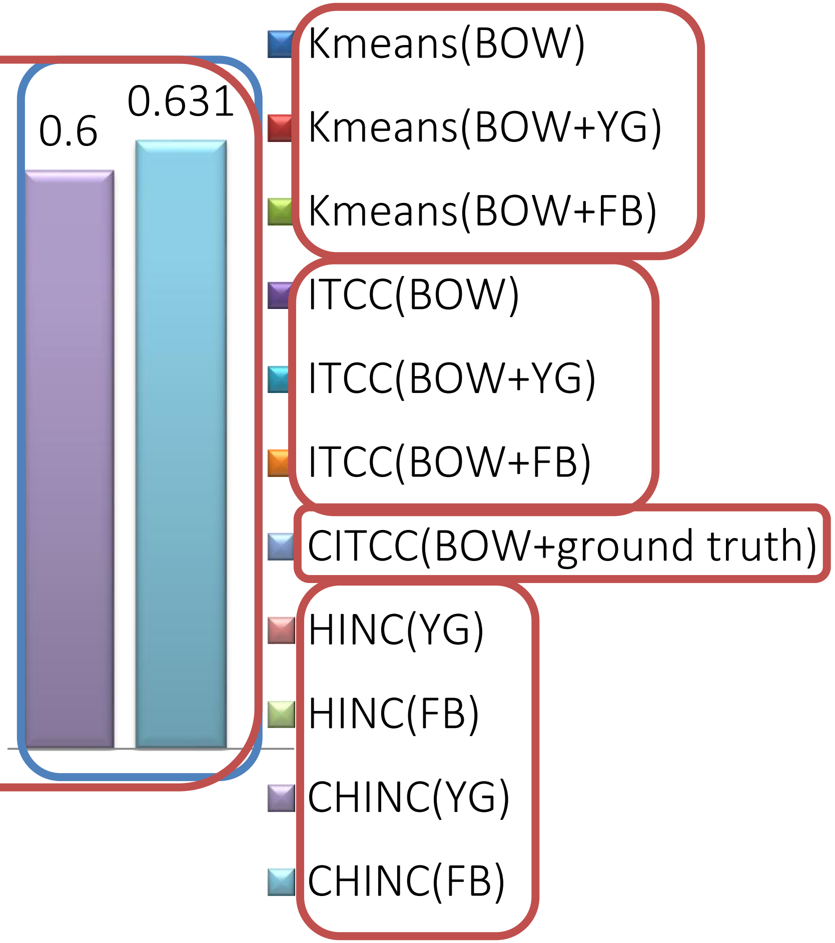
Clustering Results on 20 Newsgroups

Constrained information-theoretic co-clustering [Y. Song TKDE'13] with BOW + 250K ground-truth constraints.

The effect of different world knowledge



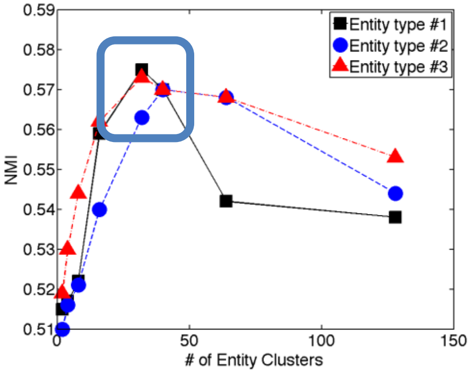
Freebase specifies more entities than YAGO2 does



Wang et al., Incorporating World Knowledge to Document Clustering via Heterogeneous Information Networks. KDD'15.
Wang et al. World knowledge as indirect supervision for document clustering. TKDD'16.

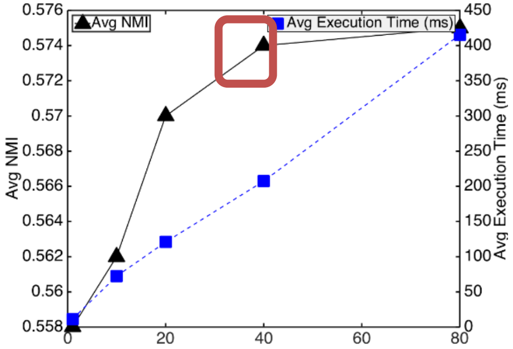
Parameter Study

Clustering with different numbers of entity clusters of each entity type



Finding #1: <u>certain values </u>of the number of entity clusters leading to the best clustering performance.

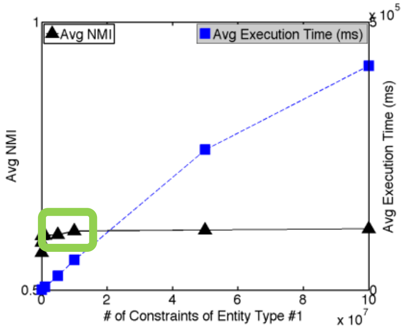
Optimization algorithm with different numbers of iterations

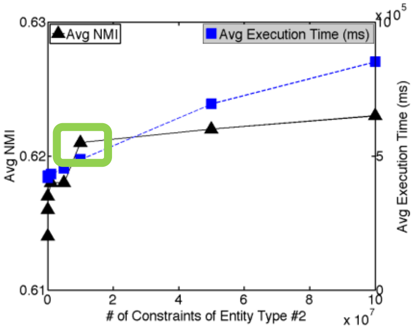


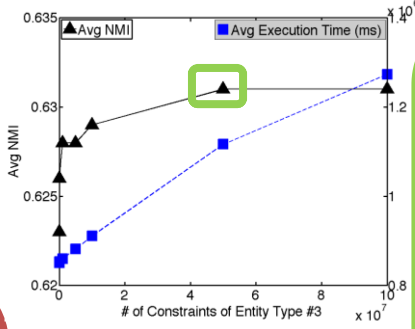
Finding #2: larger number of iterations, the clustering improves more, and <u>become stable.</u>

Because it comes to convergence.

Clustering with world knowledge constraints





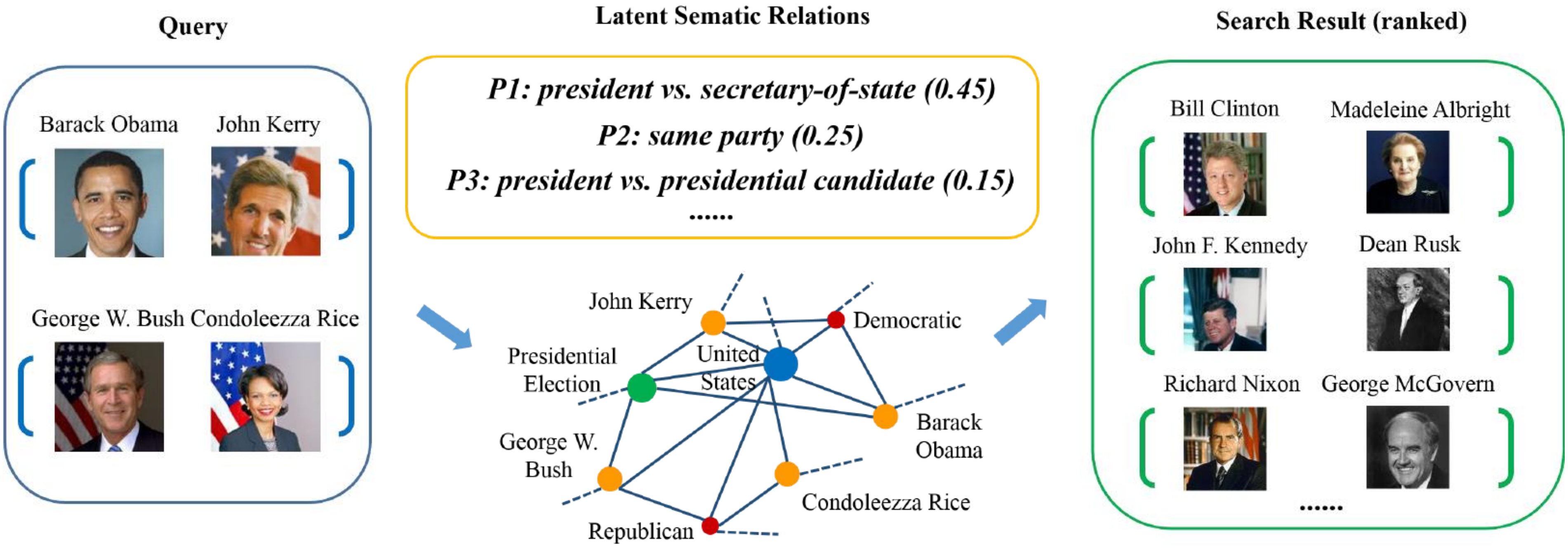


Finding #3: adding <u>more constraints</u> leading to <u>better performance</u>. Then become stable.

The entity sub-type information is transferred to the document side.

Other Research

- Relation search



Query

Latent Sematic Relations

Search Result (ranked)

P1: president vs. secretary-of-state (0.45)

P2: same party (0.25)

P3: president vs. presidential candidate (0.15)

......

Example query-based meta-paths on Rel-Full. We show the most important four query-based meta-paths of different queries.

Query: {⟨Google, Larry Page⟩, ⟨Microsoft, Bill Gates⟩, *etc.*}	ω
$\mathit{Organization} \xrightarrow{\text{is founded by}} \mathit{Founder}$	0.384
$\mathit{Organization} \xrightarrow{\text{run business in}} \mathit{Industry} \xrightarrow{\text{win award in}^{-1}} \mathit{Founder}$	0.274
$\mathit{Organization} \xrightarrow{\text{is founded by}} \mathit{Person} \xrightarrow{\text{is influence peer}^{-1}} \mathit{Founder}$	0.174
$\mathit{Organization} \xrightarrow{\text{'s leadership}} \mathit{Person} \xrightarrow{\text{mailing address}} \mathit{Location} \xrightarrow{\text{mailing address}^{-1}} \mathit{Founder}$	0.115
Query: {⟨Google, Larry Page⟩, ⟨Yahoo!, Marissa Mayer⟩, *etc.*}	ω
$\mathit{Organization} \xrightarrow{\text{run by}} \mathit{CEO} \xrightarrow{\text{job title}} \mathit{Founder}$	0.32
$\mathit{Organization} \xrightarrow{\text{founded date}} \mathit{Date} \xrightarrow{\text{graduation date}^{-1}} \mathit{Founder}$	0.229
$\mathit{Organization} \xrightarrow{\text{headquarter}} \mathit{Location} \xrightarrow{\text{education institute}} \mathit{Founder}$	0.207
$\mathit{Organization} \xrightarrow{\text{run business in}} \mathit{Industry} \xrightarrow{\text{win award in}^{-1}} \mathit{Founder}$	0.113

Wang et al. RelSim: Relation Similarity Search in Schema-Rich Heterogeneous Information Networks. SDM'16.

Future Work

Different domains

tweets, blogs, websites, medical, psychology

Which domain needs to consider more structured information?

+

World knowledge bases

With help of machine learning algorithms

What if there is no domain knowledge in the world knowledge base?

=

[Document similarity in ICDM’15]
[Document clustering in KDD’15]
[Document classification in AAAI’16]
[Item recommendation, ongoing]

More general and effective machine learning/ data mining

[Relation clustering in IJCAI’15]
[Similarity search in SDM’16]
[Paraphrasing in ACL’13]
[Data type refinement, ongoing]

Knowledge Networked learning
+
Deep learning
=

Conclusion

Problem	Text Representation and Annotation Efforts
Framework	World knowledge specification and representation; Text as HIN based learning and modeling
System	We are working on making analyzing text as network open source [Data and Code]

Thank You! ☺

Dataset

- 4 sub-datasets are constructed

20NewsGroup

Document datasets					
Sub-datasets	#(Document)	#(word)	#(Entity)	#(Total)	#(Types)
20NG-SIM	3000	22686	5549	31235	1514
20NG-DIF	3000	25910	6344	35254	1601
GCAG-SIM	3596	22577	8118	34227	1678
GCAT-DIF	2700	33345	12707	48752	1523
Each sub-datasets consists of three similar or distinct topics.					

RCV1-GCAT

More entities in GCAT