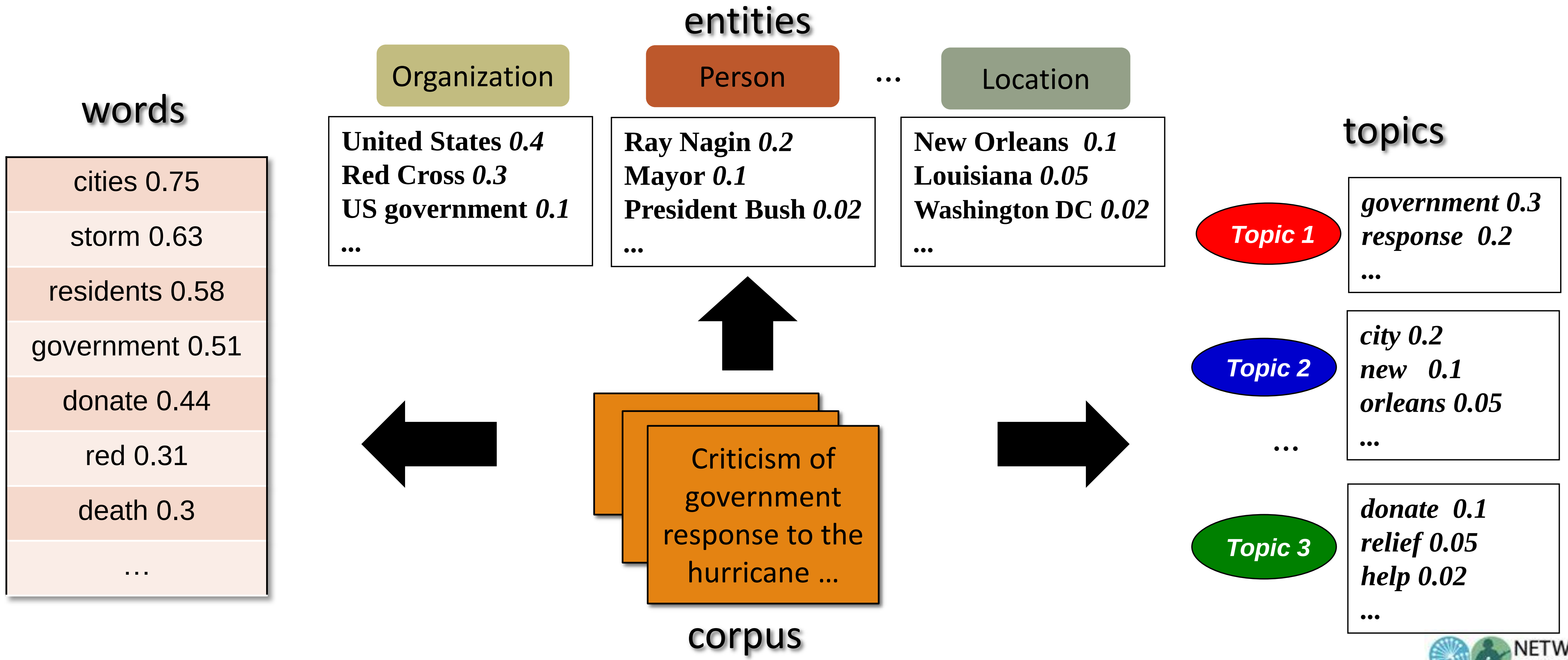
ClusType: Effective Entity Recognition and Typing by Relation Phrase-Based Clustering

Xiang Ren (UIUC), Ahmed El-Kishky (UIUC), Chi Wang (MSR), Fangbo Tao (UIUC), Clare R. Voss (ARL), Heng Ji (RPI), Jiawei Han (UIUC)

Motivation of Entity Recognition and Typing

- Making sense of large text corpus



Example: Linking Entities to Knowledge Base

The criticism consisted primarily of condemnations of mismanagement in response to Hurricane Katrina. Specifically, there was a delayed response to the flooding of New Orleans, Louisiana. New Orleans Mayor Ray Nagin was also criticized for failing to implement his evacuation plan.

Bush was criticized for not returning to Washington, D.C. from his vacation in Texas until after Wednesday afternoon. On the morning of August 28, the president telephoned Mayor Nagin to "plead" for a mandatory evacuation of New Orleans, and Nagin and Gov. Blanco decided to evacuate the city in response to that request

Criticism of government response to the hurricane …

WIKIPEDIA
The Free Encyclopedia

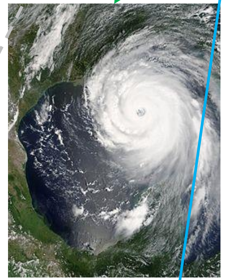
Link entity mentions to knowledge base entries for in-depth entity information

NETWORK SCIENCE CTA

Example: Linking Entities to Knowledge Base

The criticism consisted primarily of condemnations of mismanagement in response to Hurricane Katrina. Specifically, there was a delayed response to the flooding of New Orleans, Louisiana. New Orleans Mayor Ray Nagin was also criticized for failing to implement his evacuation plan.

Bush was criticized for not returning to Washington, D.C. from his vacation in Texas until after Wednesday afternoon. On the morning of August 28, the president telephoned Mayor Nagin to "plead" for a mandatory evacuation of New Orleans, and Nagin and Gov. Blanco decided to evacuate the city in response to that request



WIKIPEDIA
The Free Encyclopedia

"Entities" are what a large part of our knowledge is about

NETWORK SCIENCE CTA

What Power Can We Gain if More Structures are available?

- Structured database queries
- Structures facilitate heterogeneous network analysis, …

26 Yizhou Sun, Jiawei Han, Charu C. Aggarwal, Nitesh V. Chawla: When will it happen?: relationship prediction in heterogeneous information networks. WSDM 2012: 663-672

Knowledge hidden in DBLP Network	Mining Functions
Who are the **leading** researchers on Web search?	Ranking
Who are the **peer** researchers of Jure Leskovec?	Similarity Search
Whom **will** Christos Faloutsos **collaborate with**?	Relationship Prediction
Which types of **relationships** are most **influential** for an author to decide her topics?	Relation Strength Learning
How was the field of Data Mining **emerged** or **evolving**?	Network Evolution
Which authors are **rather different** from his/her peers in IR?	Outlier/anomaly detection

What Is Entity Recognition and Typing (ER)

- Identify token spans of entity mentions in text, and classify them into predefined set of types of interest

[Barack Obama] arrived this afternoon in [Washington, D.C.]. [President Obama]'s wife [Michelle] accompanied him

[TNF alpha] is produced chiefly by activated [macrophages]

What Is Entity Recognition and Typing (ER)

- **Identify token spans of entity mentions in text, and classify them into predefined set of types of interest**

[Barack Obama] arrived this afternoon in [Washington, D.C.]. [President Obama]'s wife [Michelle] accompanied him

PERSON
LOCATION

[TNF alpha] is produced chiefly by activated [macrophages]

PROTEIN
CELL

NETWORK SCIENCE CTA

Can We Rely on Existing Named Entity Recognition Methods?

- **Traditional named entity recognition systems from NLP**
  - Domain adaptation
  - Human annotation
  - Slow model training

- **Entity linking techniques**
  - Low coverage & freshness
  - >50% unlinkable entity mentions in Web corpus
  - >90% in our experiment corpora: tweets, Yelp reviews, …

NETWORK SCIENCE CTA

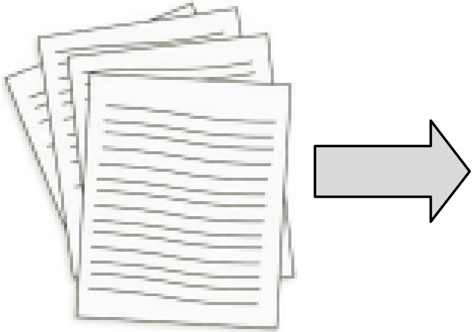
“Automated” Recognition of Typed Entities from Text

- **Goal**: recognizing entity mentions of target types with minimal/no human supervision and with no requirement that entities can be found in a KB.
- **Weak supervision**: relies on *manually selected seed entities*
  - pattern-based bootstrapping methods & label propagation methods
  - Assumptions on seeds: unambiguous and sufficiently frequent

    → requires careful seed selection by human
- **Distant supervision**: leverages entity information in KBs to reduce human supervision (cont.)

Typical Workflow of Distant Supervision

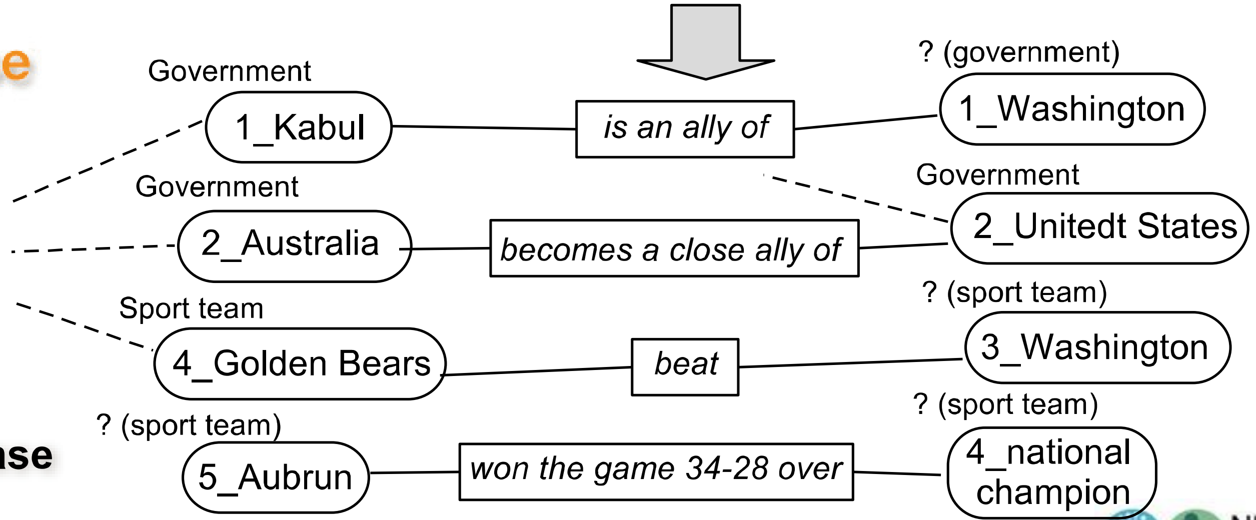
- Detect entity mentions from text
- Map candidate mentions to KB entities of target types
- Use confidently mapped {mention, type} to infer types of remaining candidate mentions



Text corpus

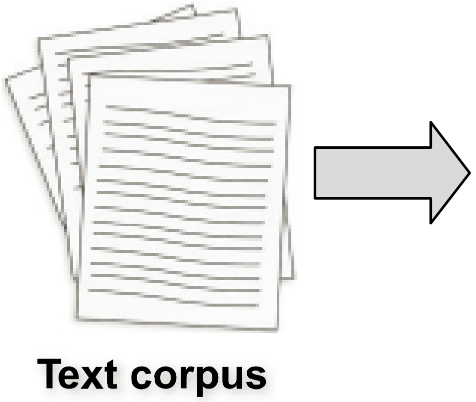
ID	Document Text
1	... has concerns whether **Kabul** *is an ally of* **Washington.**
2	... **Australia** *becomes a close ally of* the **United States**. ...
3	He *has offices in* **Washington, Boston** and **San Francisco**.
4	... The **Cardinal** *will share the title with* **California** if the **Golden Bears** *beat* **Washington** later Saturday. ...
5	... **Auburn** *won the game 34-28 over* the defending **national champions**. ...

Knowledge base

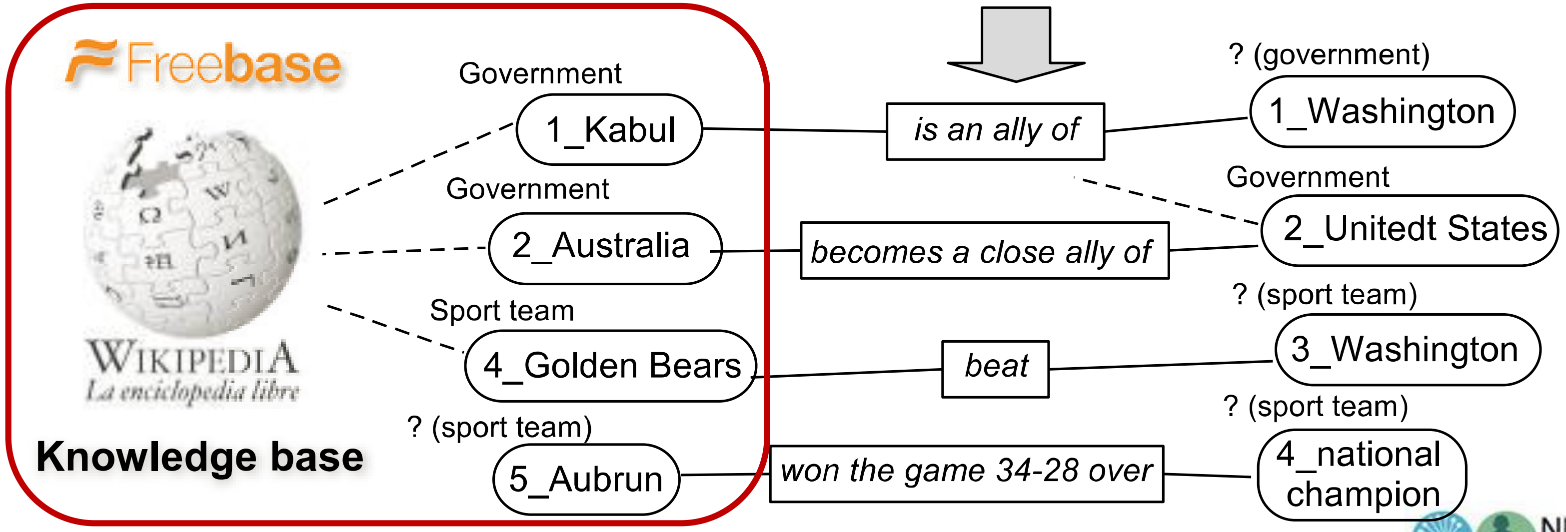


Typical Workflow of Distant Supervision

- ❑ Detect entity mentions from text
- ❑ Map candidate mentions to KB entities of target types
- ❑ Use confidently mapped {mention, type} to infer types of remaining candidate mentions

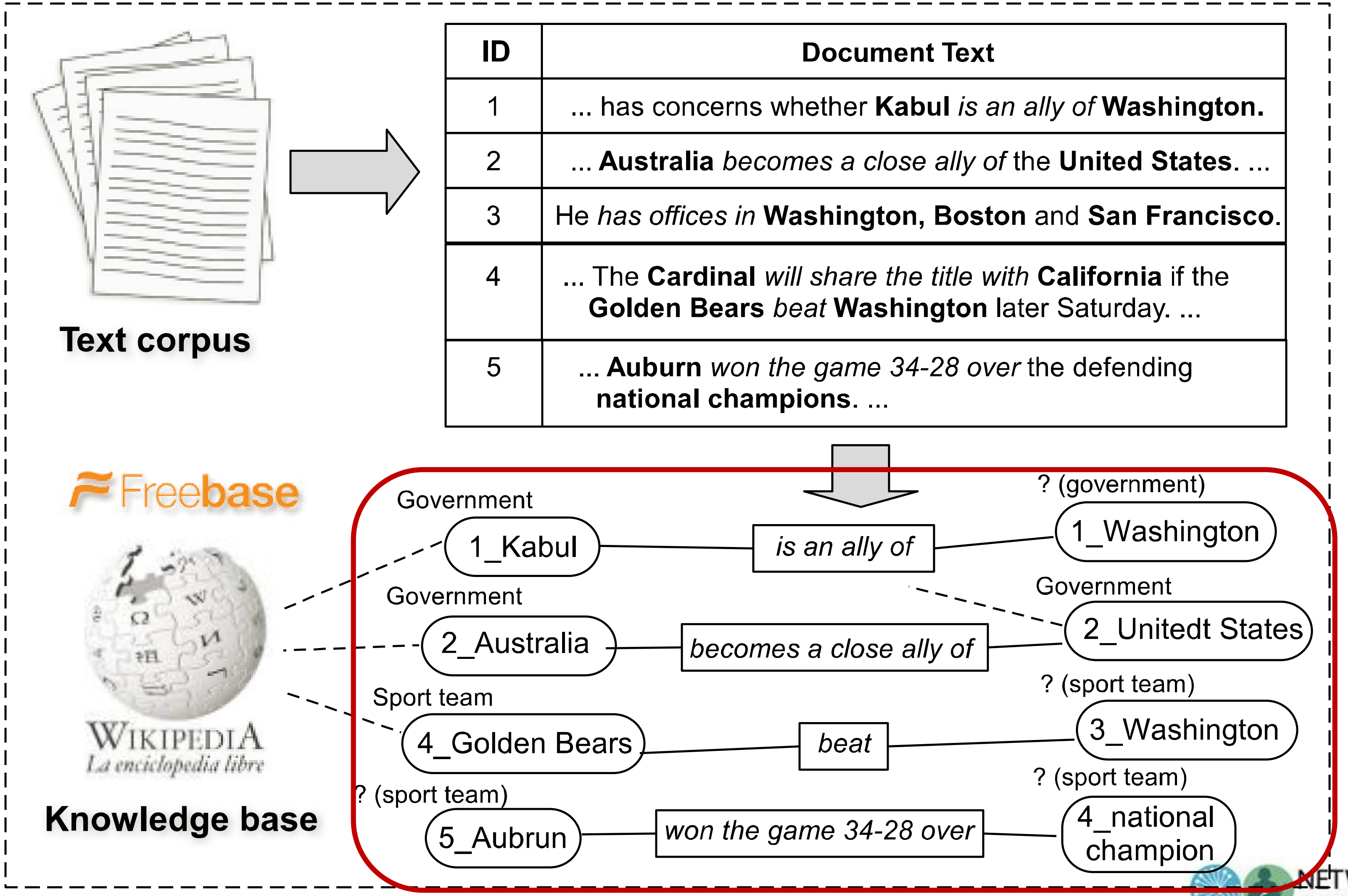


ID	Document Text
1	... has concerns whether **Kabul** *is an ally of* **Washington.**
2	... **Australia** *becomes a close ally of* the **United States**. ...
3	He *has offices in* **Washington, Boston** and **San Francisco**.
4	... The **Cardinal** *will share the title with* **California** if the **Golden Bears** *beat* **Washington** later Saturday. ...
5	... **Auburn** *won the game 34-28 over* the defending **national champions**. ...



Typical Workflow of Distant Supervision

- Detect entity mentions from text
- Map candidate mentions to KB entities of target types
- Use confidently mapped {mention, type} to infer types of remaining candidate mentions



Challenge I: Domain Restriction

- Most existing work assume entity mentions are already extracted by existing entity detection tools
  - Trained on general-domain corpora (clean, grammatical)
  - Depends on various linguistic features (dependency structures)
  - Specific, dynamic or emerging domains (*e.g.*, tweets, Yelp reviews)
    - E.g, "*in-and-out*" from Yelp review

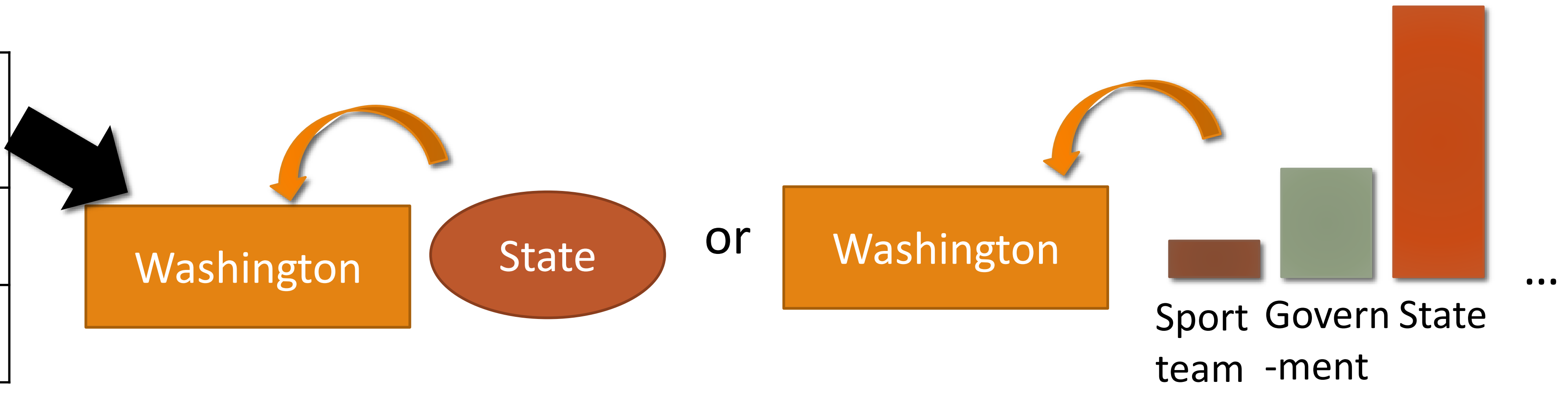
Challenge II: Name Ambiguity

- Multiple entities may share the same surface name

While Griffin is not the part of Washington's plan on Sunday's game, …	Sport team
…has concern that Kabul is an ally of Washington.	US government
He has office in Washington, Boston and San Francisco	US capital city

- Previous methods simply output a single type/type distribution for each surface name, instead of an exact type for each entity mention

While Griffin is not the part of Washington's plan on Sunday's game, …
… news from Washington indicates that the congress is going to…
It is one of the best state parks in Washington.



Challenge III: Context Sparsity

- A variety of contextual clues are leveraged to find sources of shared semantics across different entities
  - Keywords, Wiki concepts, linguistic patterns, textual relations, …

ID	Sentence	Freq
1	The magnitude 9.0 quake **caused widespread devastation in [Kesennuma city]**	12
2	… tsunami that **ravaged [northeastern Japan]** last Friday	31
3	The resulting tsunami **devastate [Japan]**’s northeast	244

- Previous methods have difficulties in handling entity mention with sparse (infrequent) context

NETWORK SCIENCE CTA

Our Solution

Domain-agnostic phrase mining algorithm

- Extracts candidate entity mentions & relation phrases with minimal linguistic/domain assumption → ~~domain restriction~~

Do not simply merge entity mentions with *identical surface names*

- Model each mention based on its surface name and context, in a scalable way → ~~name ambiguity~~

Mine synonymous *relation phrase* co-occurring with entity mentions

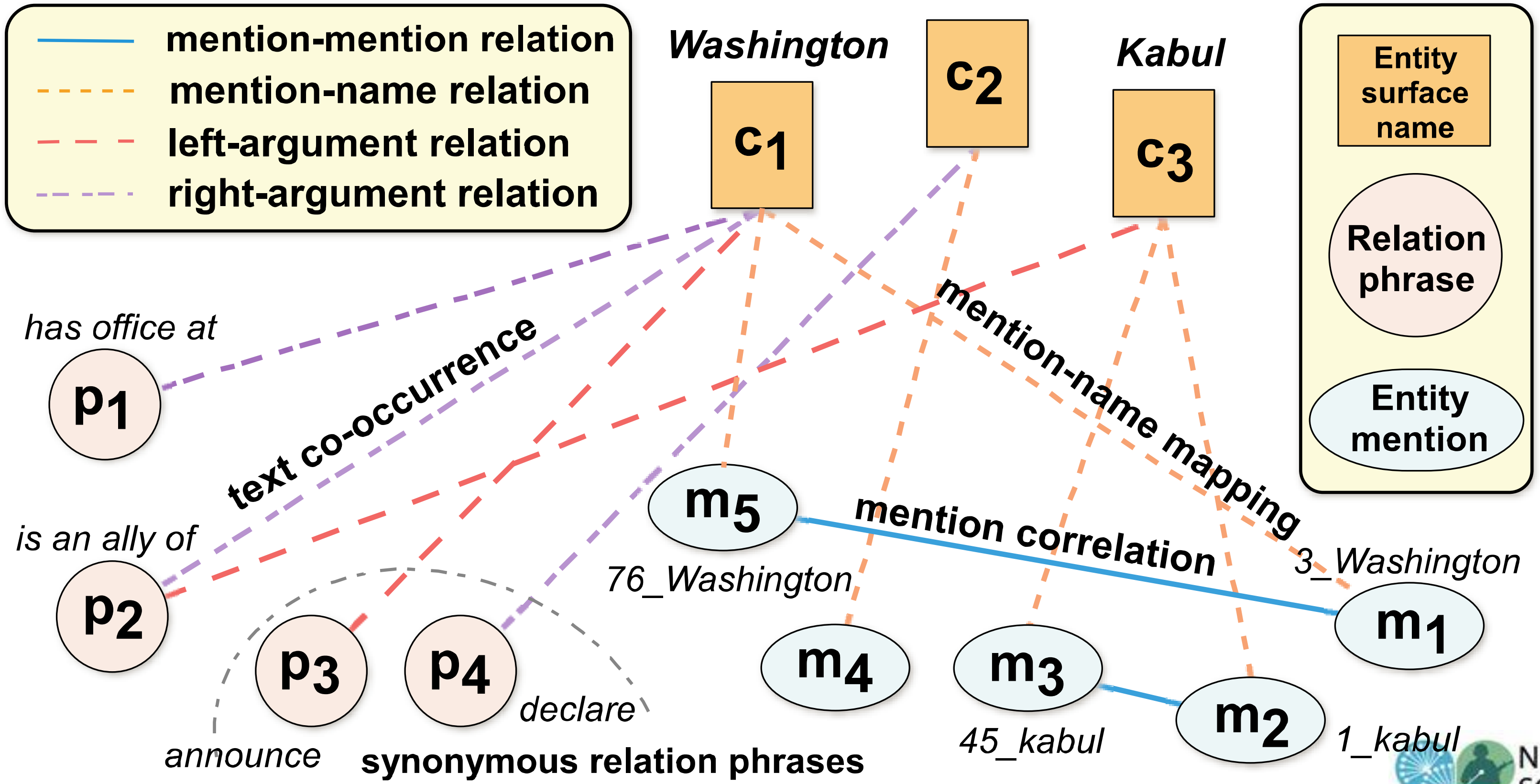
- Helps form connecting bridges among entities that do not share identical context, but share synonymous relation phrases → ~~context sparsity~~

NETWORK SCIENCE CTA

A Relation Phrase-Based Entity Recognition Framework

- POS-constrained phrase segmentation for mining candidate entity mentions and relation phrases, simultaneously
- Construct a heterogeneous graph to represent available information in a unified form

Entity mentions are kept as individual objects **to be disambiguated**



Why Jointly Clustering Relation Phrases?

- With the constructed graph, formulate a graph-based semi-supervised learning of two tasks jointly:

Type propagation on heterogeneous graph

Multi-view relation phrase clustering

Entity argument types serve as **good feature** for clustering relation phrases

Type propagation via **synonymous relation phrases**

Type signatures of frequent relation phrases can help infer the type signatures of infrequent (sparse) ones that have similar cluster memberships

Candidate Generation

- An efficient phrase mining algorithm incorporating both *corpus-level statistics* and *syntactic constraints*
  - **Global significance score**: Filter low-quality candidates;
  - **Generic POS tag patterns**: remove phrases with improper syntactic structure

Relation phrase: phrase that denotes a unary or binary relation in a sentence

Pattern	Example
V	disperse; hit; struck; knock;
P	in; at; of; from; to;
V P	locate in; come from; talk to;
$VW^*(P)$	caused major damage on; come lately

V-verb; P-prep; W-{adv | adj | noun | det | pron}
W^* denotes multiple W; (P) denotes optional.

Partitions corpus into segments which meet both significance threshold and POS patterns

Over:RP the weekend the system:EP dropped:RP nearly inches of snow in:RP western Oklahoma:EP and at:RP [Dallas Fort Worth International Airport]:EP sleet and ice caused:RP hundreds of [flight cancellations]:EP and delays. ...... It is forecast:RP to reach:RP [northern Georgia]:EP by:RP [Tuesday afternoon]:EP, Washington:EP and [New York]:EP by:RP [Wednesday afternoon]:EP, meteorologists:EP said:RP.

EP: entity mention candidate; RP: relation phrase

Candidate Generation

- An efficient phrase mining algorithm incorporating both *corpus-level statistics* and *syntactic constraints*
  - **Global significance score**: Filter low-quality candidates;
  - **Generic POS tag patterns**: remove phrases with improper syntactic structure

- Entity detection performance comparison with an NP chunker

Method	NYT		Yelp		Tweet	
	Prec	Recall	Prec	Recall	Prec	Recall
Our method	**0.469**	**0.956**	**0.306**	**0.849**	0.226	**0.751**
NP chunker	0.220	0.609	0.296	0.247	**0.287**	0.181

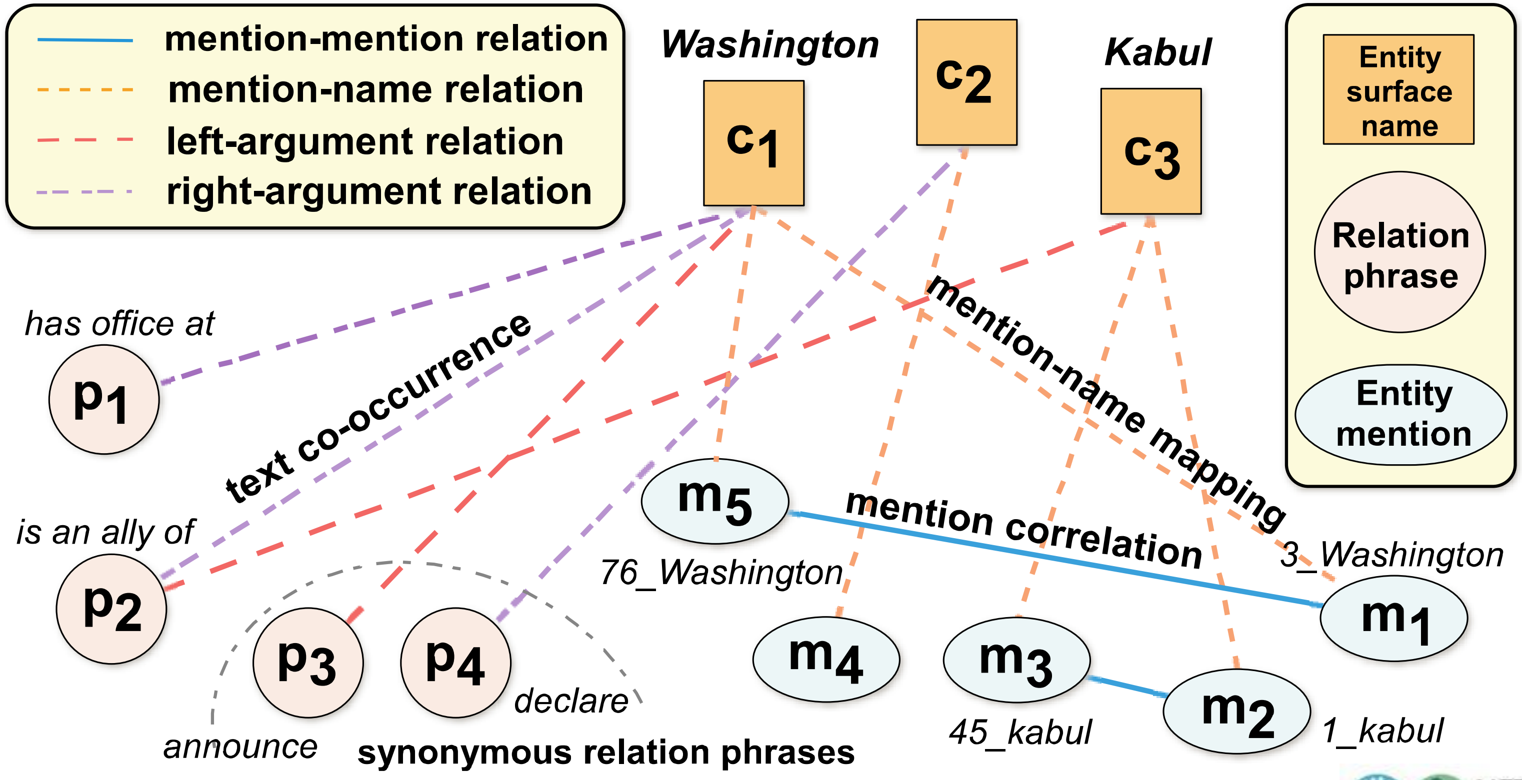
Recall is most critical for this step, since later we cannot detect the misses (i.e., false negatives)

Construction of Heterogeneous Graphs

- With three types of objects extracted from corpus: candidate entity mentions, entity surface names, and relation phrases
  - We can construct a heterogeneous graph to **enforce several hypotheses** (cont.)

Smoothness Assumption

the more two objects are likely to share the same label, the larger the weight will be associated with their connecting edge

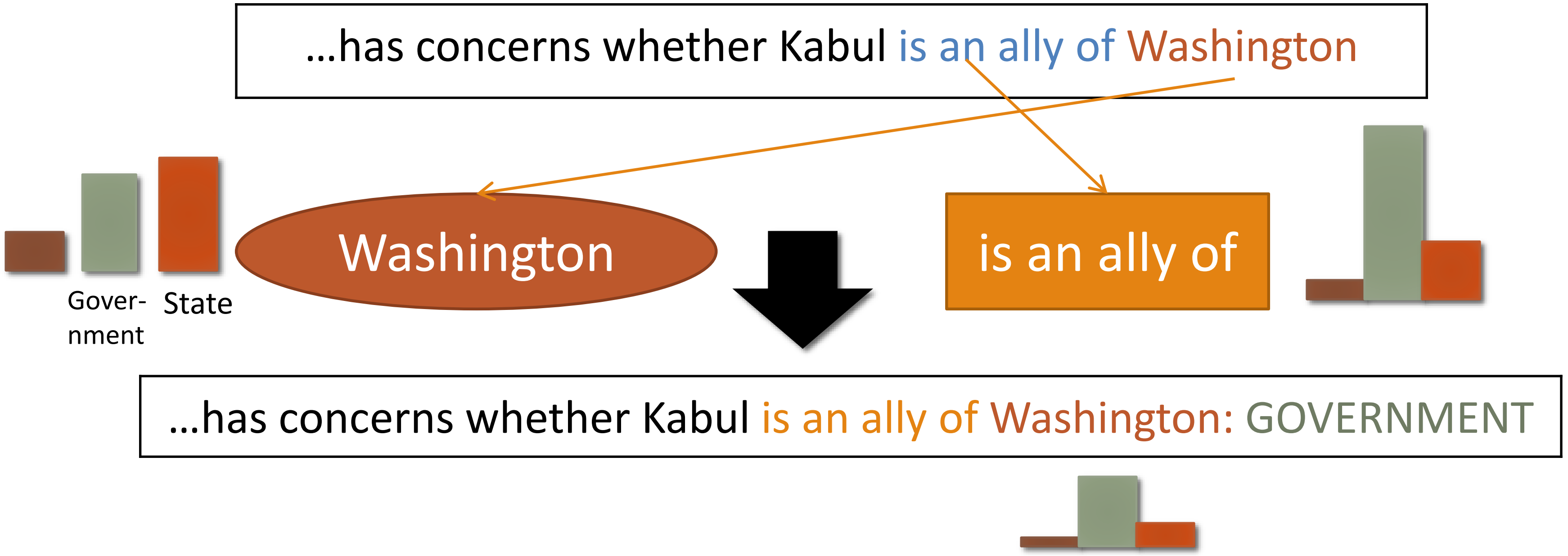


Construction of Heterogeneous Graphs

- Q1: How to model exact type for each entity mention?
- Q2: How to leverage relation phrases to propagate types?
- Q3: How to disambiguate entity mentions?

Modeling Type for Entity Mention

- Directly modeling type indicator of each entity mention in label propagation
  - → Intractable size of parameter space
- Type cues for entity mention:
  1. Types distribution of its surface name (entity mention-surface name subgraph)
  2. Type signature of its surrounding relation phrases



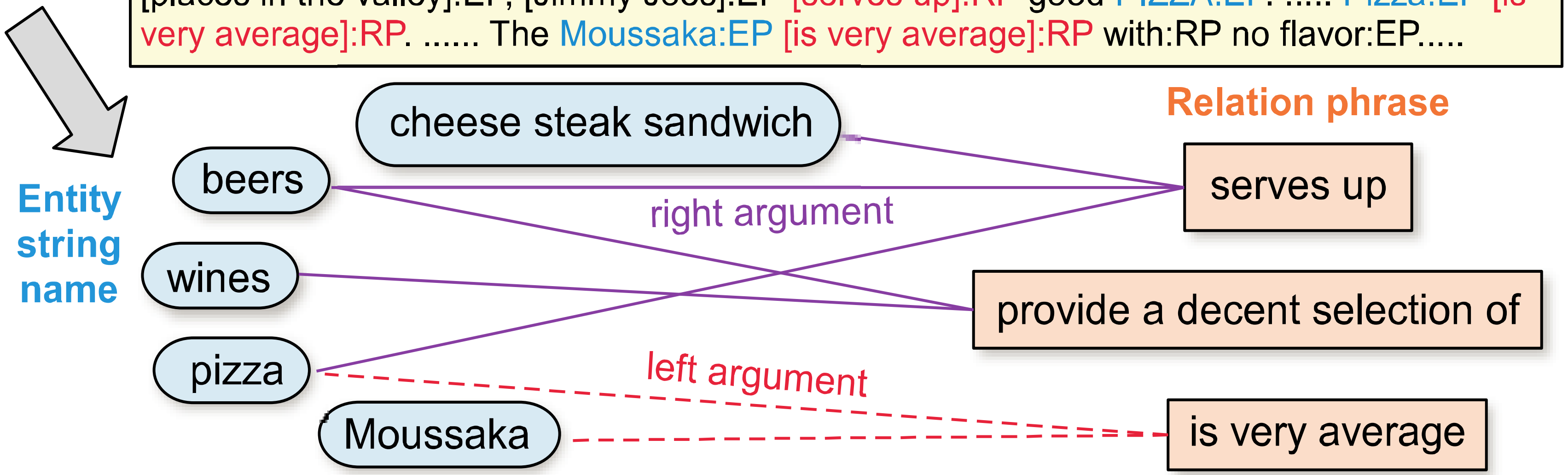
Entity Name-Relation Phrase Subgraph

- Weight importance of different relation phrases for different surface names

Text Corpus

This place:EP [serves up]:RP the best [cheese steak sandwich]:EP west of:RP the Mississippi:EP. ...... Four Peaks:EP [serves up]:RP some beers:EP and great eats:RP. ...... They [provide a decent selection of]:RP beers:EP and high-end wines:EP. ...... Tons of:RP [places in the valley]:EP, [Jimmy Joes]:EP [serves up]:RP good PIZZA:EP. ..... Pizza:EP [is very average]:RP. ...... The Moussaka:EP [is very average]:RP with:RP no flavor:EP.....

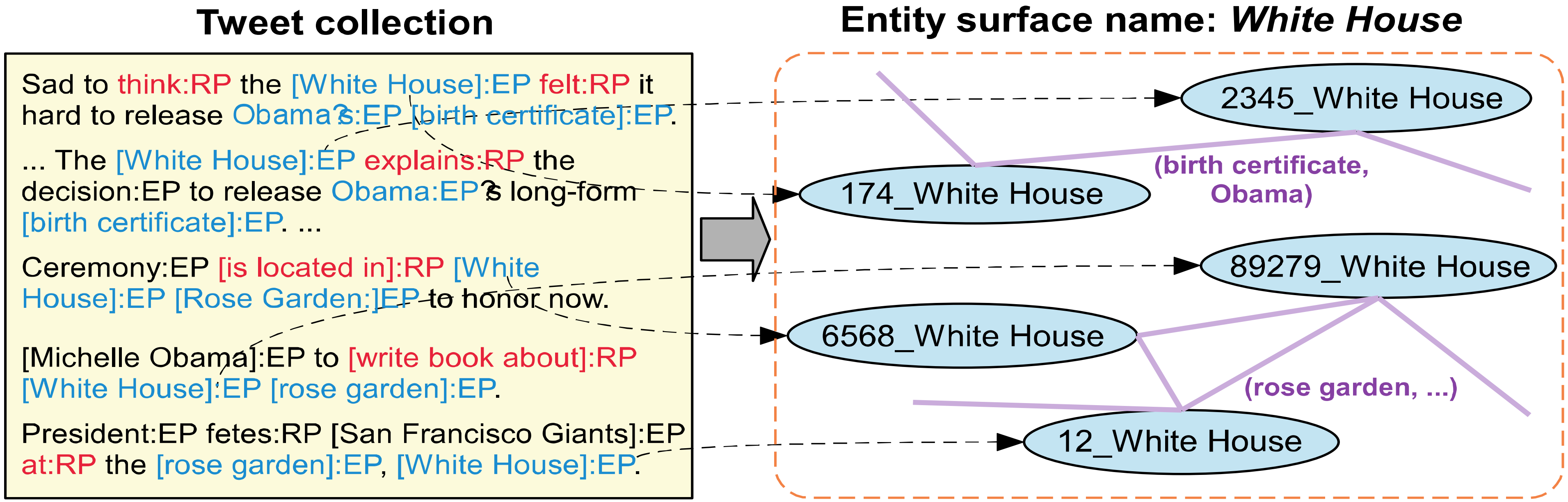
Aggregated co-occurrences between entity surface names and relation phrases across corpus



HYPOTHESIS 1 (ENTITY-RELATION CO-OCCURRENCES). *If surface name c often appears as the left (right) argument of relation phrase p, then c's type indicator tends to be similar to the corresponding type indicator in p's type signature.*

Mention Correlation Subgraph

- Name with ambiguous type: "***White House***" (GOV or LOC?)
- Other co-occurring mentions may provide good hints
  - E.g., "birth certificate" and "rose garden"



HYPOTHESIS 2 (MENTION CORRELATION). *If there exists a strong correlation (i.e., within sentence, common neighbor mentions) between two candidate mentions that share the same name, then their type indicators tend to be similar.*

How to Cluster Relation Phrases Jointly?

- **Signals we consider:**
  1. String similarity (e.g., edit distance)
  2. Context similarity (e.g., distributional & embedding features)
  3. Type signature consistency (type distributions of entity arguments)
  4. More to try…
- **Softly clustering synonymous relation phrases**
  - Multi-view NMF as part of the final objective function

HYPOTHESIS 3 (TYPE SIGNATURE CONSISTENCY). *If two relation phrases have similar cluster memberships, the type indicators of their left and right arguments (type signature) tend to be similar, respectively.*

HYPOTHESIS 4 (RELATION PHRASE SIMILARITY). *Two relation phrases tend to have similar cluster memberships, if (1) their strings are similar; (2) their context words are similar; and (3) the type indicators of their left and right arguments are similar, respectively.*

Type Inference: A Joint Optimization Problem

$$\mathcal{O}_{\alpha,\gamma,\mu} = \mathcal{F}(\mathbf{C},\mathbf{P}_L,\mathbf{P}_R) + \mathcal{L}_\alpha\big(\mathbf{P}_L,\mathbf{P}_R,\{\mathbf{U}^{(v)},\mathbf{V}^{(v)}\},\mathbf{U}^*\big) + \Omega_{\gamma,\mu}(\mathbf{Y},\mathbf{C},\mathbf{P}_L,\mathbf{P}_R). \quad (2)$$

$$\mathcal{F}(\mathbf{C},\mathbf{P}_L,\mathbf{P}_R) = \sum_{i=1}^{n}\sum_{j=1}^{l} W_{L,ij}\left\|\frac{\mathbf{C}_i}{\sqrt{D^{(\mathcal{C})}_{L,ii}}} - \frac{\mathbf{P}_{L,j}}{\sqrt{D^{(\mathcal{P})}_{L,jj}}}\right\|_2^2 + \sum_{i=1}^{n}\sum_{j=1}^{l} W_{R,ij}\left\|\frac{\mathbf{C}_i}{\sqrt{D^{(\mathcal{C})}_{R,ii}}} - \frac{\mathbf{P}_{R,j}}{\sqrt{D^{(\mathcal{P})}_{R,jj}}}\right\|_2^2$$

Type propagation between entity surface names and relation phrases (H.1)

Mention modeling & mention correlation (H.2)

$$\Omega_{\gamma,\mu}(\mathbf{Y},\mathbf{C},\mathbf{P}_L,\mathbf{P}_R) = \left\|\mathbf{Y} - f(\Pi_{\mathcal{C}}\mathbf{C},\Pi_L\mathbf{P}_L,\Pi_R\mathbf{P}_R)\right\|_F^2 + \frac{\gamma}{2}\sum_{c\in\mathcal{C}}\sum_{i,j=1}^{M_c} W^{(c)}_{ij}\left\|\frac{\mathbf{Y}_i}{\sqrt{D^{(c)}_{ii}}} - \frac{\mathbf{Y}_j}{\sqrt{D^{(c)}_{jj}}}\right\|_2^2 + \mu\|\mathbf{Y}-\mathbf{Y}_0\|_F^2$$

$$\mathcal{L}_\alpha\big(\mathbf{P}_L,\mathbf{P}_R,\{\mathbf{U}^{(v)},\mathbf{V}^{(v)}\},\mathbf{U}^*\big) \quad (3)$$
$$= \sum_{v=0}^{d}\beta^{(v)}\big(\|\mathbf{F}^{(v)} - \mathbf{U}^{(v)}\mathbf{V}^{(v)T}\|_F^2 + \alpha\|\mathbf{U}^{(v)}\mathbf{Q}^{(v)} - \mathbf{U}^*\|_F^2\big).$$

Multi-view relation phrases clustering (H.3 & 4)

The ClusType Algorithm

$$\min_{\substack{\mathbf{Y},\mathbf{C},\mathbf{P},\mathbf{P}_R,\mathbf{U}^* \\ \{\mathbf{U}^{(v)},\mathbf{V}^{(v)},\beta^{(v)}\}}} \mathcal{O}_{\alpha,\gamma,\mu,\lambda_L,\lambda_\Omega}$$

$$\text{s.t.}\quad \mathbf{Y} \in \{0,1\}^{M\times T},\quad \mathbf{Y}\mathbf{1} = \mathbf{1};$$

$$\mathbf{U}^* \geq 0,\ \mathbf{U}^{(v)} \geq 0,\ \mathbf{V}^{(v)} \geq 0;$$

$$\sum_{v=0}^{d} \exp(-\beta^{(v)}) = 1,\ \forall 0 \leq v \leq d.$$

- Efficiently solved by alternate minimization based on block coordinate descent algorithm
- Algorithm complexity is linear to #entity mentions, #relation phrases, #cluster, #clustering features and #target types

The ClusType algorithm:

Update type indicators and type signatures

$$\mathbf{Y}^{(c)} = \big[(1+\gamma+\mu)\mathbf{I}_c - \gamma\mathbf{S}_{\mathcal{M}}^{(c)}\big]^{-1}\big(\mathbf{\Theta}^{(c)} + \mu\mathbf{Y}_0^{(c)}\big),\quad \forall c \in \mathcal{C},\quad (7)$$

$$\mathbf{C} = \frac{1}{2}\big[\mathbf{S}_L\mathbf{P}_L + \mathbf{S}_R\mathbf{P}_R + \Pi_{\mathcal{C}}^T(\mathbf{Y} - \Pi_L\mathbf{P}_L - \Pi_R\mathbf{P}_R)\big];\quad (8)$$

$$\mathbf{P}_L = \mathbf{X}_0^{-1}\big[\mathbf{S}_L^T\mathbf{C} + \Pi_L^T(\mathbf{Y} - \Pi_{\mathcal{C}}\mathbf{C} - \Pi_R\mathbf{P}_R) + \beta^{(0)}\mathbf{U}^{(0)}\mathbf{V}^{(0)T}\big];$$

$$\mathbf{P}_R = \mathbf{X}_1^{-1}\big[\mathbf{S}_R^T\mathbf{C} + \Pi_R^T(\mathbf{Y} - \Pi_{\mathcal{C}}\mathbf{C} - \Pi_L\mathbf{P}_L) + \beta^{(1)}\mathbf{U}^{(1)}\mathbf{V}^{(1)T}\big];$$

For each view, performs single-view NMF until converges

$$V_{jk}^{(v)} = V_{jk}^{(v)}\frac{[\mathbf{F}^{(v)T}\mathbf{U}^{(v)}]_{jk} + \alpha\sum_{i=1}^{l}U_{ik}^{*}U_{ik}^{(v)}}{\mathbf{\Delta}_{jk}^{(v)} + \alpha\big(\sum_{i=1}^{l}U_{ik}^{(v)2}\big)\big(\sum_{i=1}^{T}V_{ik}^{(v)}\big)},\quad (9)$$

$$U_{ik}^{(v)} = U_{ik}^{(v)}\frac{[\mathbf{F}^{(v)+}\mathbf{V}^{(v)} + \alpha\mathbf{U}^*]_{ik}}{[\mathbf{U}^{(v)}\mathbf{V}^{(v)T}\mathbf{V}^{(v)} + \mathbf{F}^{(v)-}\mathbf{V}^{(v)} + \alpha\mathbf{U}^{(v)}]_{ik}}.\quad (10)$$

Update consensus matrix and relative weights of different views

$$\mathbf{U}^* = \frac{\sum_{v=0}^{d}\beta^{(v)}\mathbf{U}^{(v)}\mathbf{Q}^{(v)}}{\sum_{v=0}^{d}\beta^{(v)}};\quad \beta^{(v)} = -\log\Big(\frac{\delta^{(v)}}{\sum_{i=0}^{d}\delta^{(i)}}\Big).\quad (12)$$

Until the objective converges

NETWORK SCIENCE CTA

Comparing ClusType with Other Methods and Its Variants

Performance comparison on three datasets

Data sets	NYT			Yelp			Tweet		
Method	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Pattern [9]	0.4576	0.2247	0.3014	0.3790	0.1354	0.1996	0.2107	0.2368	0.2230
FIGER [16]	0.8668	0.8964	0.8814	0.5010	0.1237	0.1983	**0.7354**	0.1951	0.3084
SemTagger [12]	0.8667	0.2658	0.4069	0.3769	0.2440	0.2963	0.4225	0.1632	0.2355
APOLLO [29]	0.9257	0.6972	0.7954	0.3534	0.2366	0.2834	0.1471	0.2635	0.1883
NNPLB [15]	0.7487	0.5538	0.6367	0.4248	0.6397	0.5106	0.3327	0.1951	0.2459
ClusType-NoClus	0.9130	0.8685	0.8902	0.7629	0.7581	0.7605	0.3466	0.4920	0.4067
ClusType-NoWm	0.9244	0.9015	0.9128	0.7812	0.7634	0.7722	0.3539	**0.5434**	0.4286
ClusType-TwoStep	0.9257	0.9033	0.9143	0.8025	0.7629	0.7821	0.3748	0.5230	0.4367
ClusType	**0.9550**	**0.9243**	**0.9394**	**0.8333**	**0.7849**	**0.8084**	0.3956	0.5230	**0.4505**

46.08% and 48.94% improvement in F1 score compared to the best baseline on the Tweet and the Yelp datasets

- **vs. FIGER**: effectiveness of our candidate generation and proposed hypotheses on type propagation
- **vs. NNPLB** and **APOLLO**: ClusType not only utilizes semantic-rich relation phrase as type cues, but only cluster synonymous relation phrases to tackle context sparsity
- **vs. variants**: (i) models mention correlation for name disambiguation; (ii) integrates clustering in a mutually enhancing way

NETWORK SCIENCE CTA

Comparing on Trained NER System

- Compare with Stanford NER, which is trained on general-domain corpora including ACE corpus and MUC corpus, on three types: PER, LOC, ORG

Method	**NYT**	**Yelp**	**Tweet**
Stanford NER [6]	0.6819	0.2403	0.4383
ClusType-NoClus	0.9031	0.4522	0.4167
ClusType	**0.9419**	**0.5943**	**0.4717**

- ClusType and its variants outperform Stanford NER on both dynamic corpus (NYT) and domain-specific corpus (Yelp)
- ClusType has lower precision but higher Recall and F1 score on Tweet → Superior recall of ClusType mainly come from domain-independent candidate generation

Example Output and Relation Phrase Clusters

Table 7: Example output of ClusType and the compared methods on the Yelp dataset.

ClusType	SemTagger	NNPLB
The best **BBQ:Food** I've tasted in **Phoenix:LOC** ! I had the **[pulled pork sandwich]:Food** with **coleslaw:Food** and **[baked beans]:Food** for lunch. ...	The best BBQ I've tasted in **Phoenix:LOC** ! I had the pulled **[pork sandwich]:LOC** with **coleslaw:Food** and **[baked beans]:LOC** for lunch. ...	The best **BBQ:Loc** I've tasted in **Phoenix:LOC** ! I had the pulled pork **sandwich:Food** with coleslaw and baked **beans:Food** for **lunch:Food**. ...
I only go to **ihop:LOC** for **pancakes:Food** because I don't really like anything else on the menu. Ordered **[chocolate chip pancakes]:Food** and a **[hot chocolate]:Food**.	I only go to ihop for pancakes because I don't really like anything else on the menu. Ordered **[chocolate chip pancakes]:LOC** and a **[hot chocolate]:LOC**.	I only go to ihop for pancakes because I don't really like anything else on the menu. Ordered chocolate chip pancakes and a hot chocolate.

- Extracts more mentions and predicts types with higher accuracy

Table 8: Example relation phrase clusters and their corpus frequency from the NYT dataset.

ID	Relation phrase
1	recruited by (5.1k); employed by (3.4k); want hire by (264)
2	go against (2.4k); struggling so much against (54); run for re-election against (112); campaigned against (1.3k)
3	looking at ways around (105); pitched around (1.9k); echo around (844); present at (5.5k);

- Not only synonymous relation phrases, but also both sparse and frequent relation phrase can be clustered together
- → boosts sparse relation phrases with type information of frequent relation phrases

NETWORK SCIENCE CTA

Conclusions and Future Work

- A principled way to leverage distant supervision for entity recognition and typing
- Domain-independent, language independent!
- General framework that can be applied to other NLP tasks!

Ongoing:

- Extend to role discovery for scientific concepts → paper profiling
- Study of relation phrase clustering, such as
  - joint entity/relation clustering
  - synonymous relation phrase canonicalization
- Study of joint entity and relation phrase extraction with phrase mining

Software

- Entity Typing
  - ClusType: http://shanzhenren.github.io/ClusType
- Phrase Mining
  - SegPhrase: https://github.com/shangjingbo1226/SegPhrase
  - TopMine: http://web.engr.illinois.edu/~elkishk2/code/ToPMine.zip
- Checking our research package dissemination portal
  - IlliMine http://illimine.cs.uiuc.edu/