Effort-Light StructMine: Turning Massive Corpora into Structures

Xiang Ren

Department of Computer Science

University of Illinois at Urbana-Champaign

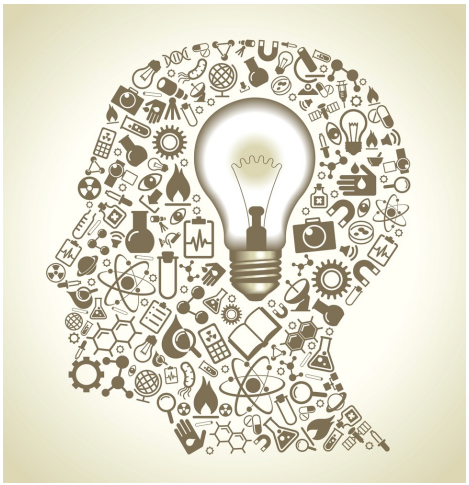
Feb. 9, 2017

Turning Unstructured Text Data into Structures

Unstructured Text Data

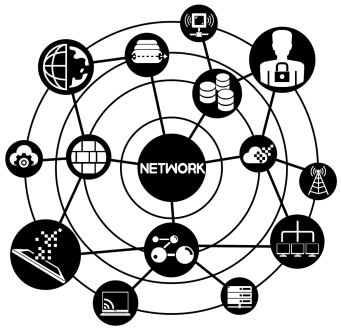
(~80% of the data collected)

?



Knowledge & Insights

Structures



STORE

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

PRODUCT

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

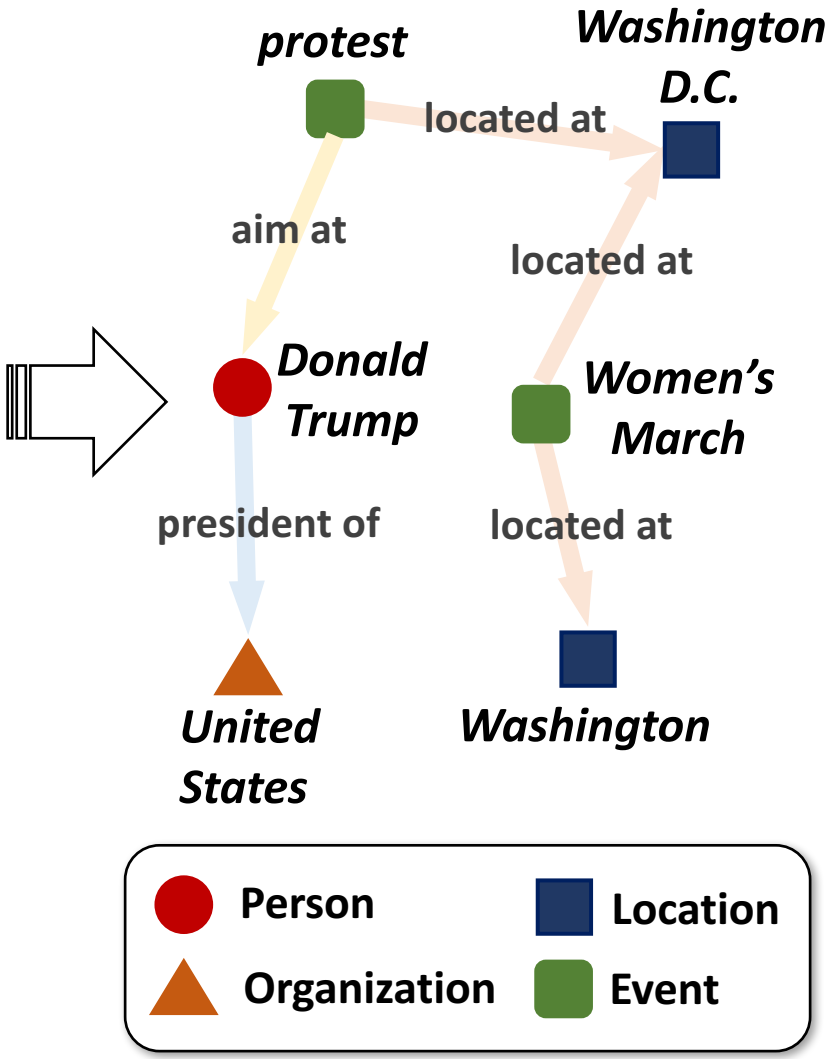
SALES_FACT

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

Reading the News: Text to Structures

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

-- *CNN*

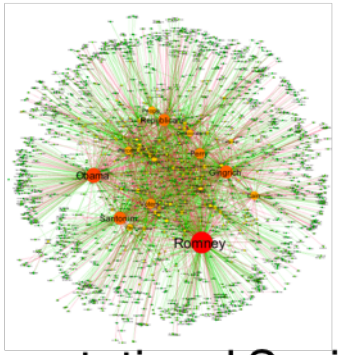


Text to Structures: Applications

Medical records
Scientific papers
Clinical reports
…

Healthcare

Social media posts
Web blogs
News articles
…



Computational Sociology

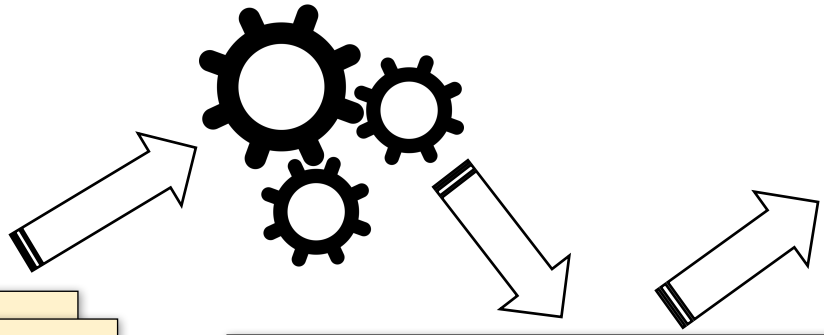
Corporate reports
News streams
Customer reviews
…

Business Intelligence

Prior Work: Mining Factual Structures with Human Efforts

Human labeling

Extraction Rules
Features / Models



Factual Structures

Women's March

United States

Donald Trump

… ***The June 2013 Egyptian protest*** were mass ***protest*** event that occurred in ***Egypt*** on 30 June 2013, …

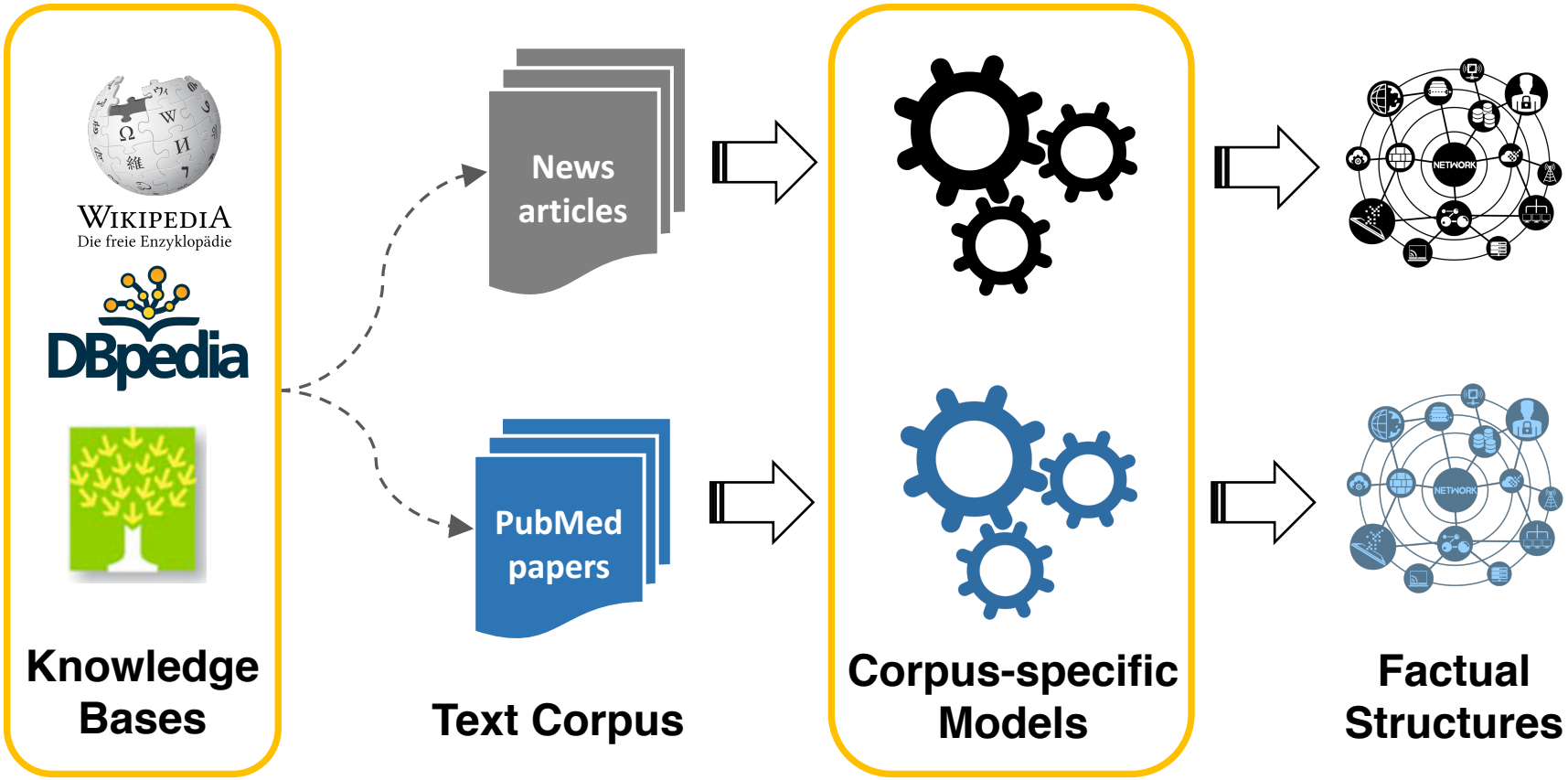
Training data

… The ***Women's March*** was a worldwide ***protest*** on January 21, 2017. The ***protest*** was aimed at ***Donald Trump***, the recently inaugurated President of the ***United States***…

New data

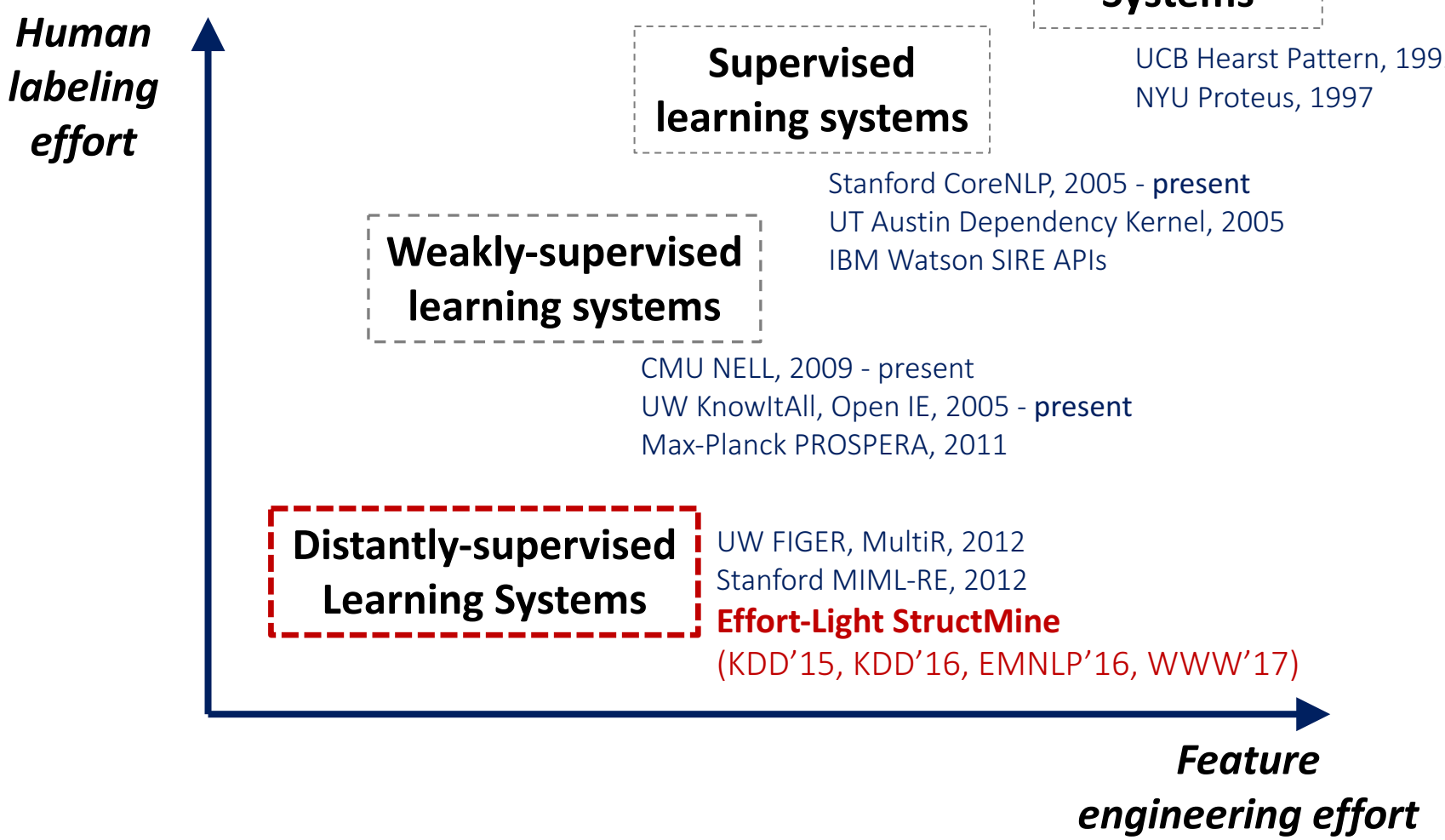
Stanford CoreNLP
CMU NELL
UW KnowItAll
IBM Alchemy APIs
…

My Work: Effort-Light StructMine



Enables quick development of applications over various corpora

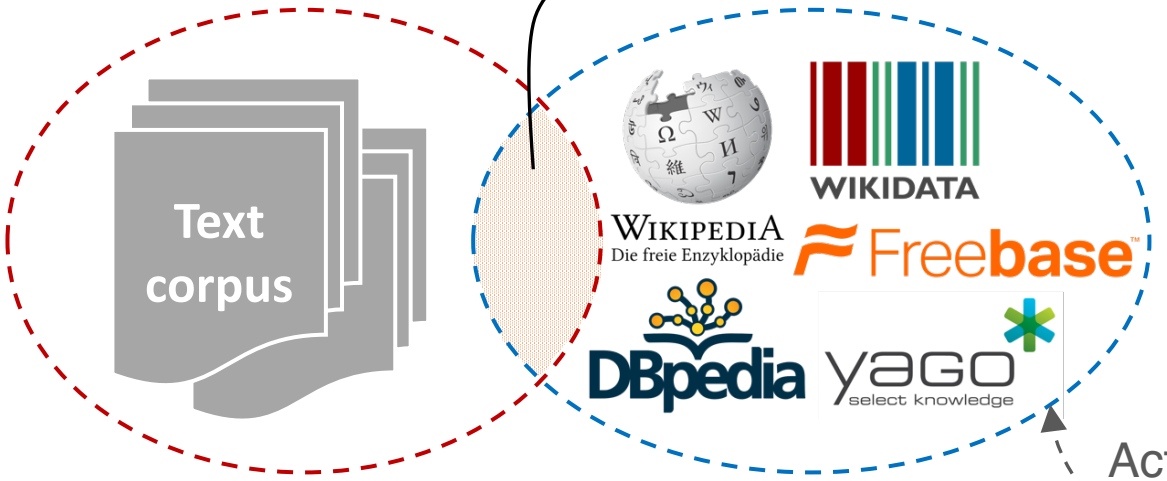
Effort–Light StructMine: **Where Are We?**



My General Approach: Distant Supervision

1% of 10M sentences
→ 100K labeled sentences

Overlapping factual information: entity names, entity types, relationships ...

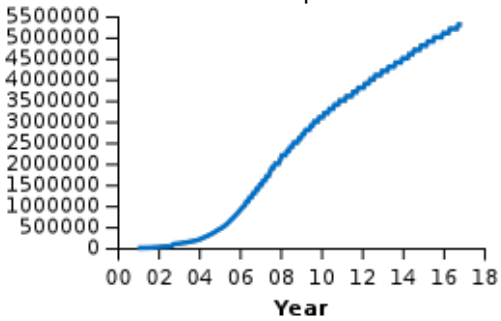


Knowledge Bases as background knowledge

Publicly available in many domains:

- Common knowledge
- Biomedical sciences
- Arts
- …

Number of Wikipedia articles



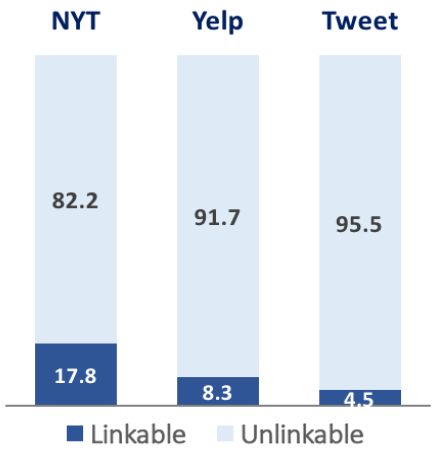
Actively enriched by human crowds

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

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

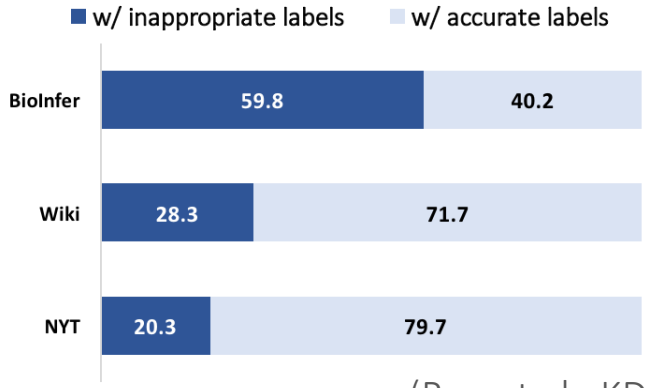
Distant Supervision: **Challenges**

1. Data sparsity of KBs
   - Entity/fact coverage in KBs
   - Confidence of mapping to KBs



(Ren et al., KDD'15)

2. Context-agnostic label assignment
   - Are ***all*** the assigned type labels appropriate for the instance's context?



(Ren et al., KDD'16)

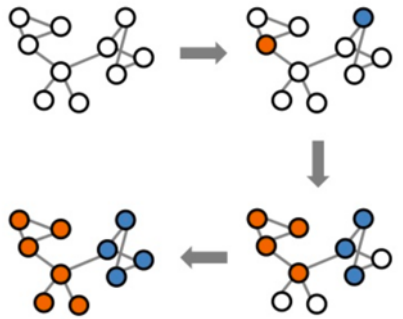
Lin et al. *No noun phrase left behind: detecting and typing unlinkable entities*. EMNLP, 2012.

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

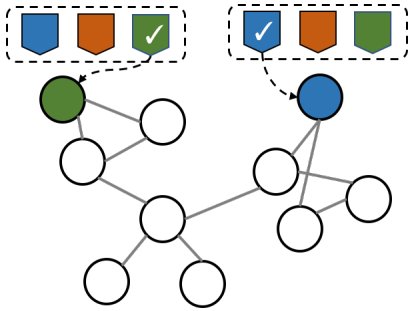
Min et al. *Distant Supervision for Relation Extraction with an Incomplete Knowledge Base*. NAACL, 2013.

Effort-Light StructMine: **Key Ideas**

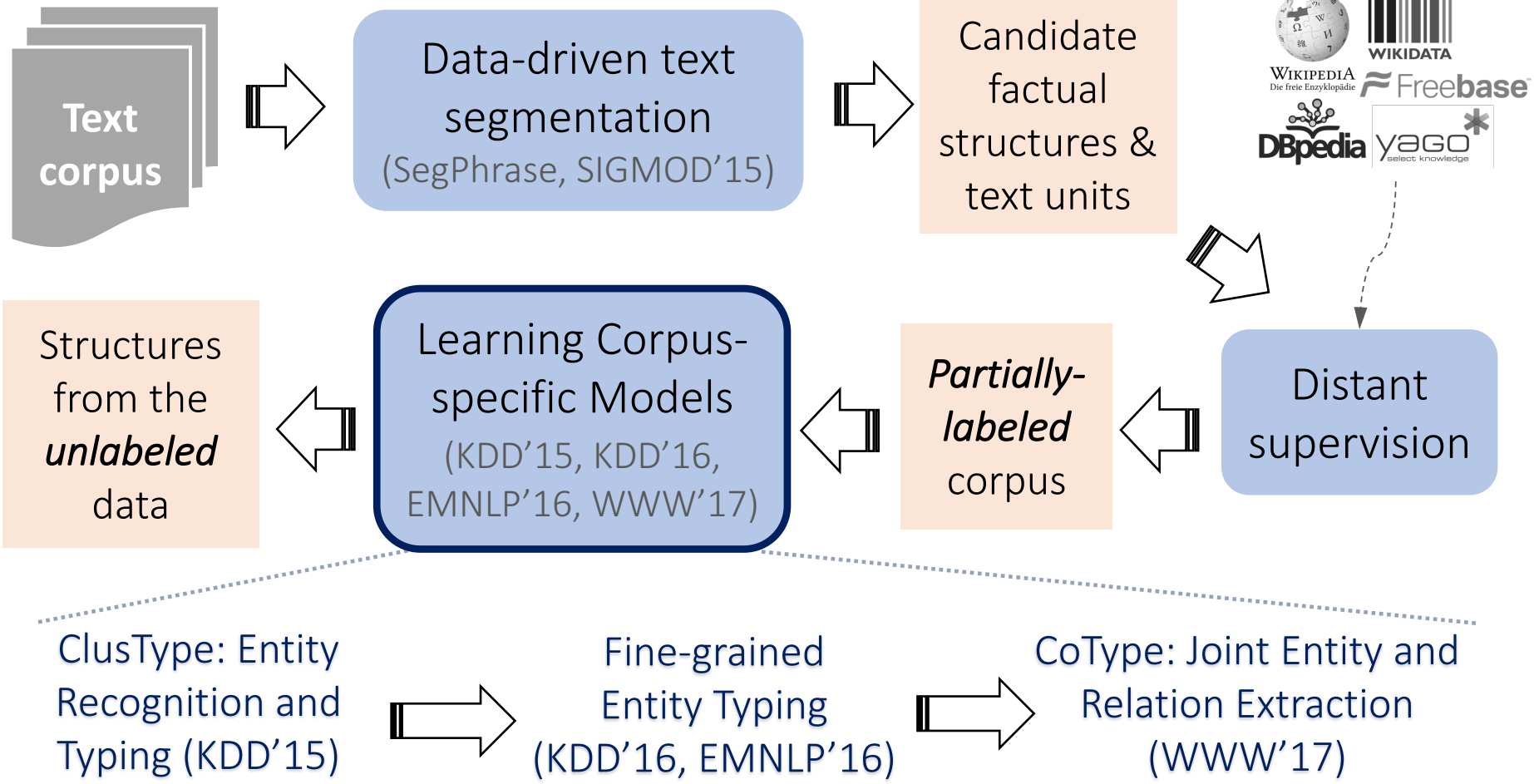
Challenge	Key Idea
Data Sparsity	**Propagate** type information via "textual bridges" + **consolidate** "similar" bridges
Context-Agnostic Label Assignment	**Select** "best" label for context with specialized optimization objectives



A Joint Optimization Framework



Effort-Light StructMine: **Methodology**



Corpus to Structured Network: The Roadmap

Outline

- Introduction
- **Entity Recognition and Typing** [KDD'15, KDD'16]
- Joint Entity and Relation Extraction [WWW'17]
- Summary and Future Directions

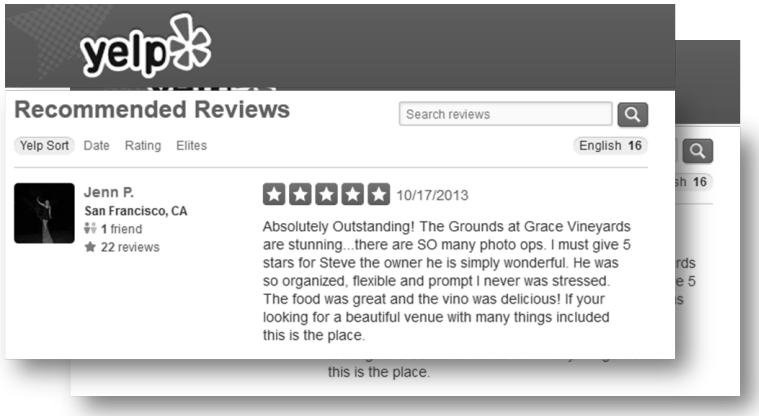
Outline

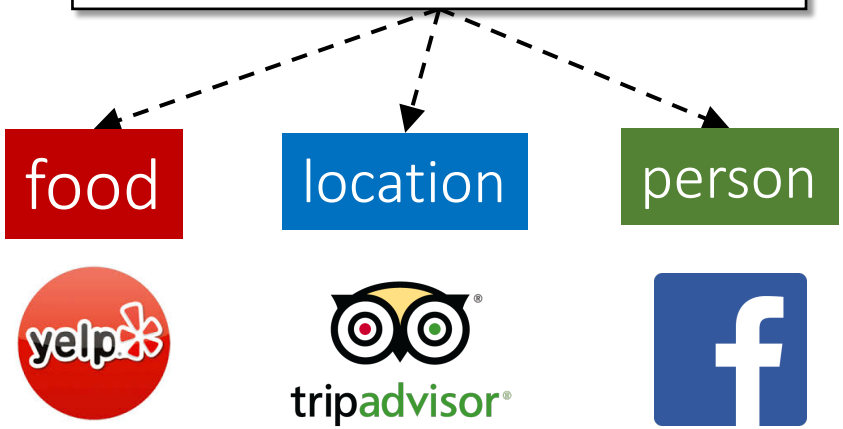
- Introduction
- Entity Recognition and Typing [KDD'15, KDD'16]
- Joint Entity and Relation Extraction [WWW'17]
- Summary and Future Directions

Recognizing Entities of Target Types from Text

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

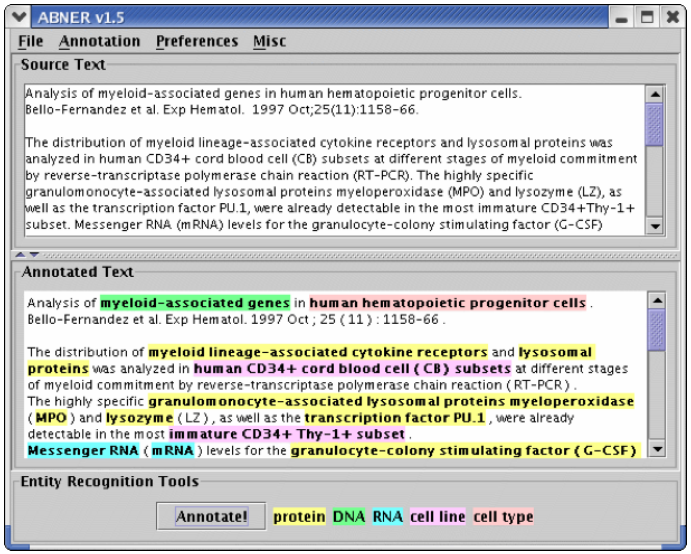
The best *BBQ* I've tasted in *Phoenix* ! I had the *pulled pork sandwich* with *coleslaw* and *baked beans* for lunch. The owner is very nice. …



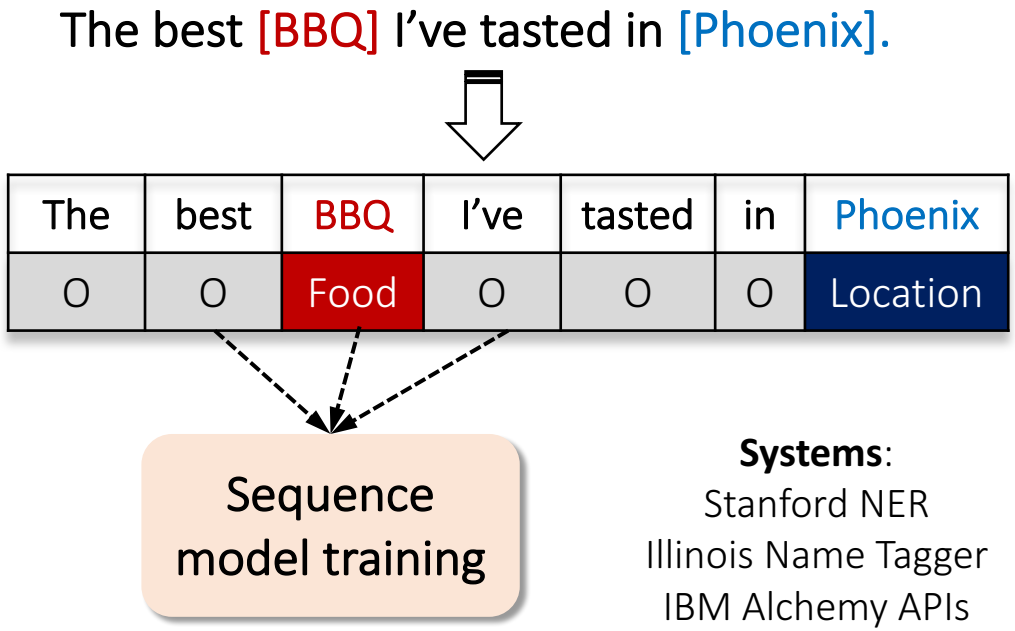


Traditional Named Entity Recognition (NER) Systems

- Heavy reliance on human annotated data
- Training sequence models is slow



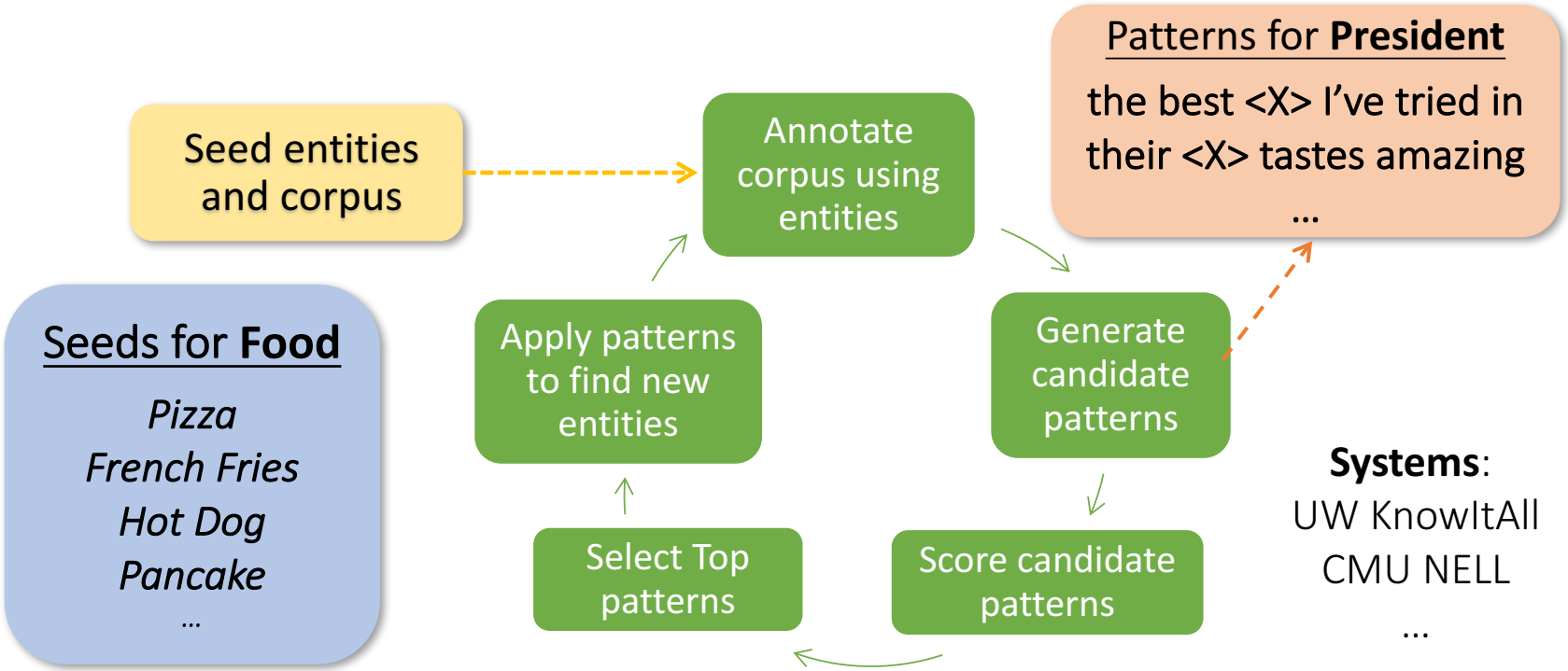
A manual annotation interface



Finkel et al., *Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling*, ACL 2005

Weak Supervision Systems: Pattern-Based Bootstrapping

- Requires manual seed selection & mid-point checking
  - Sufficiently frequent & No ambiguity

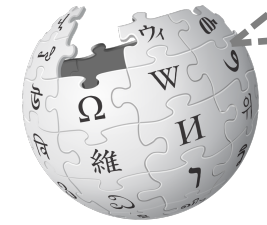


Etzioni et al., *Unsupervised named-entity extraction from the web: An experimental study*, Artificial Intelligence 2005.
Mitchell et al. Never-ending Learning, AAAI, 2015

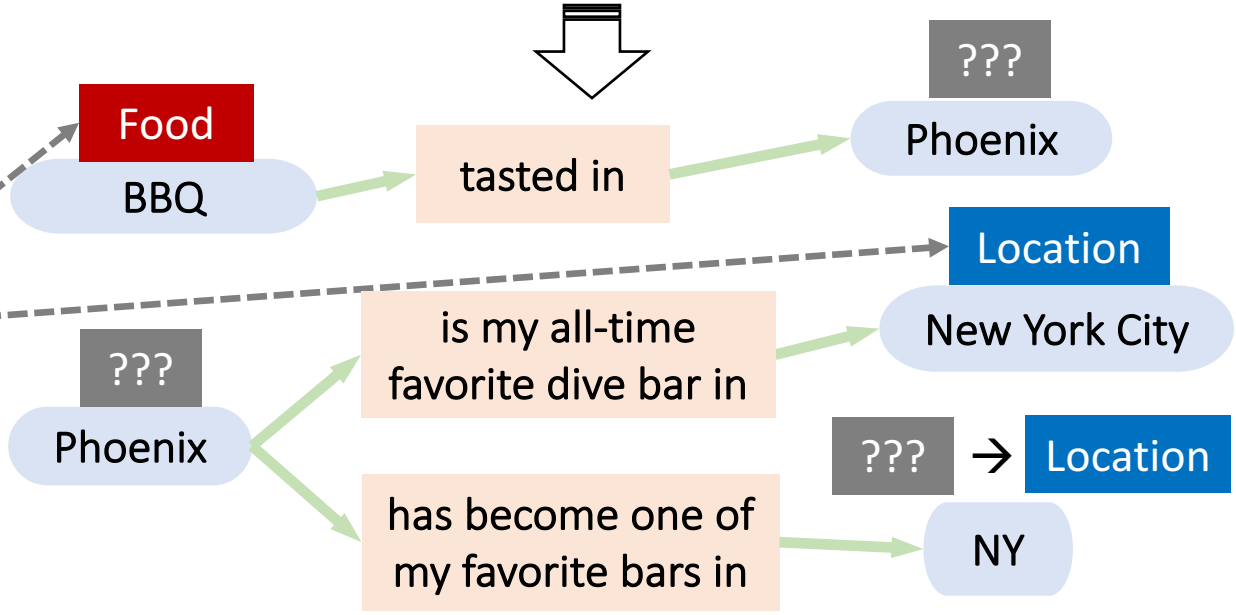
Entity Typing with Distant Supervision

1. Detect entity names from text
2. Link entity names to KB entities
3. Propagate type information to the unlinkable entities

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



WIKIPEDIA
Die freie Enzyklopädie



Lin et al. *No noun phrase left behind: detecting and typing unlinkable entities*. EMNLP, 2012.

Previous Methods: Limitation 1

1. Context-agnostic type prediction
   - Predict types for entity mentions regardless of contexts
2. Textual bridge sparsity

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

Previous Methods: Limitation 2

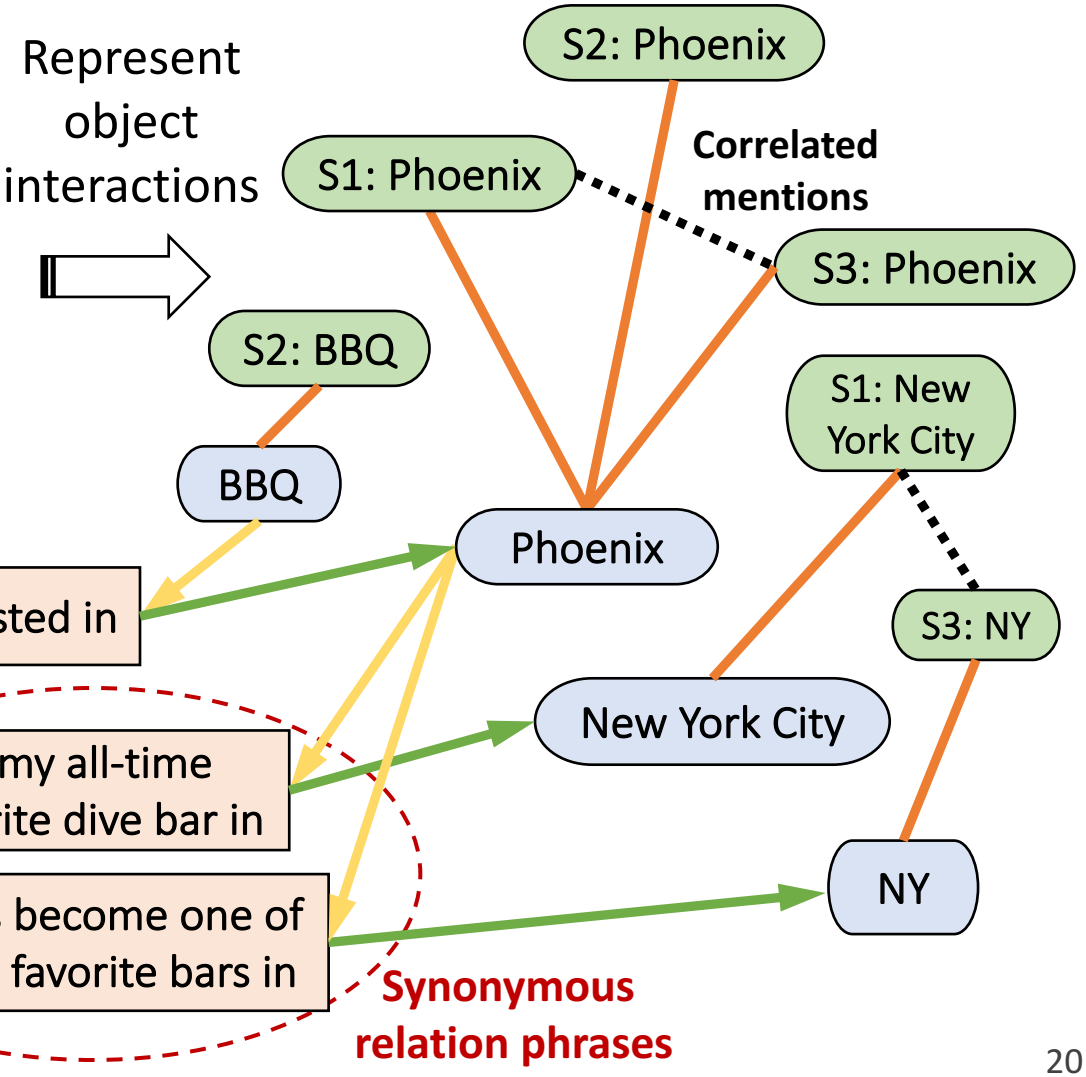
1. Context-agnostic type prediction
2. Sparsity of textual bridges
   - Some relational phrases are **infrequent** in the corpus
     → hard to propagate information

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

My Solution: **ClusType** (KDD'15)

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

Represent object interactions



Jointly optimize two sub-tasks on the graph:

1. Type label propagation
2. Relation phrase clustering

ClusType: Data-Driven Entity Mention Detection

- **Significance** of a merging between two sub-phrases

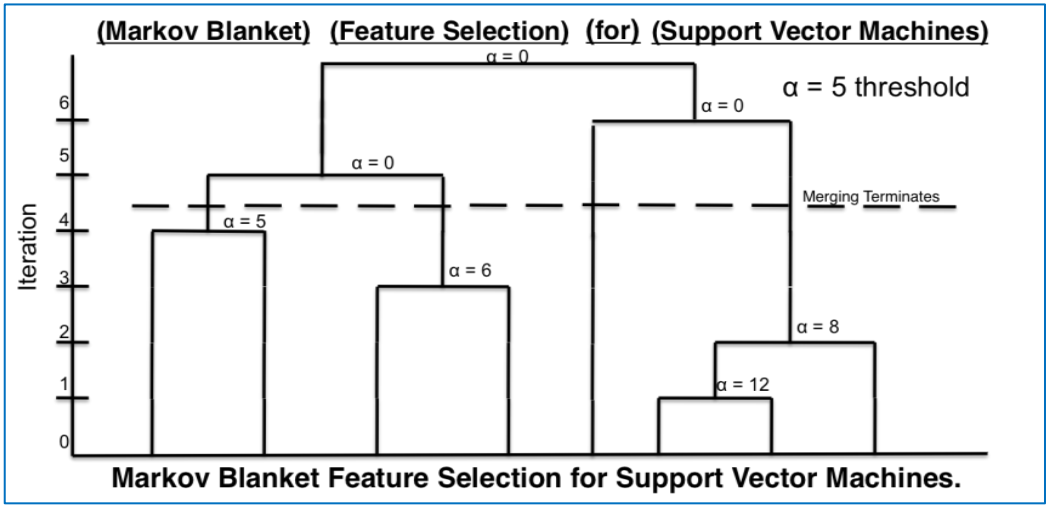
Quality of merging

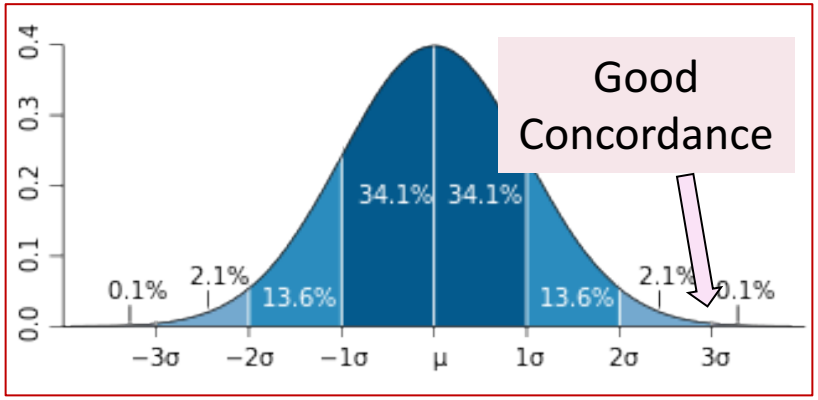
$$\rho_X(S_1, S_2) = \frac{\upsilon(S_1 \oplus S_2) - N\frac{\upsilon(S_1)}{N}\frac{\upsilon(S_2)}{N}}{\sqrt{\upsilon(S_1 \oplus S_2)}} \cdot I_X(S_1 \oplus S_2)$$

Corpus-level Concordance

Syntactic quality

Pattern	Example
(J*)N*	support vector machine
VP	tasted in, damage on
VW*(P)	train a classifier with





ClusType: Data-Driven Entity Mention Detection

- **Significance** of a merging between two sub-phrases

Quality of merging

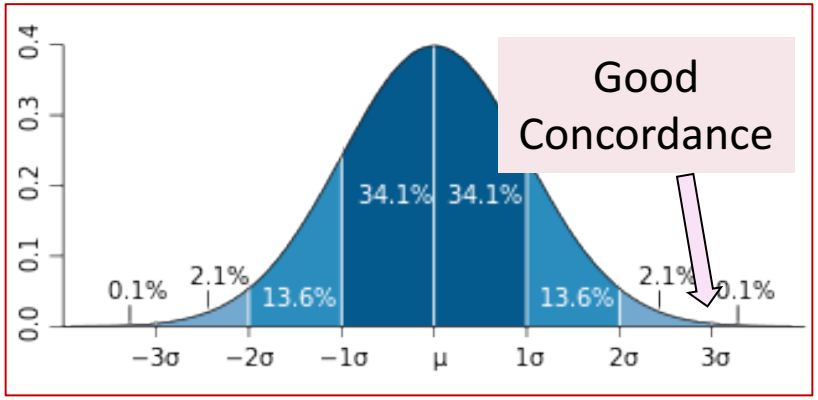
$$\rho_X(S_1, S_2) = \frac{v(S_1 \oplus S_2) - N\frac{v(S_1)}{N}\frac{v(S_2)}{N}}{\sqrt{v(S_1 \oplus S_2)}} \cdot I_X(S_1 \oplus S_2)$$

Corpus-level Concordance

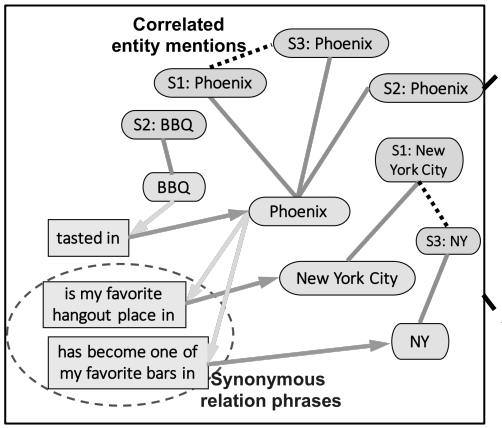
Syntactic quality

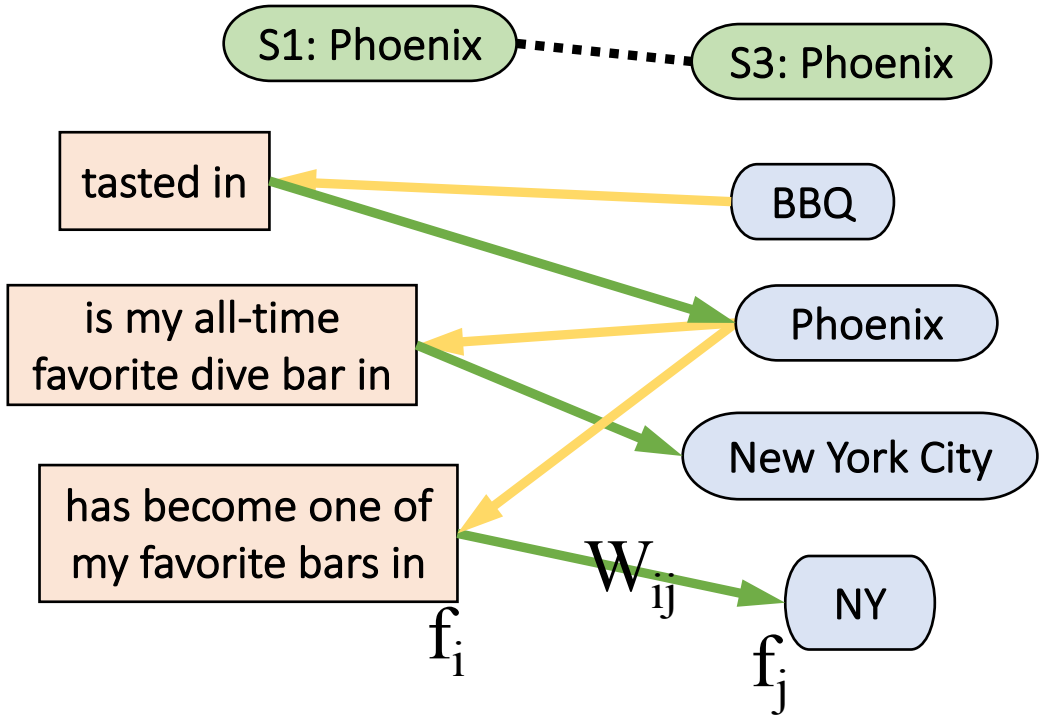
Pattern	Example
(J*)N*	support vector machine
VP	tasted in, damage on
VW*(P)	train a classifier with

The best *BBQ* I've tasted in *Phoenix* ! I had the *pulled pork sandwich* with *coleslaw* and *baked beans* for lunch. … This *place* serves up the best *cheese steak sandwich* in west of *Mississippi*.



Type Propagation in **ClusType**





Smoothness Assumption

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

Edge weight / object similarity

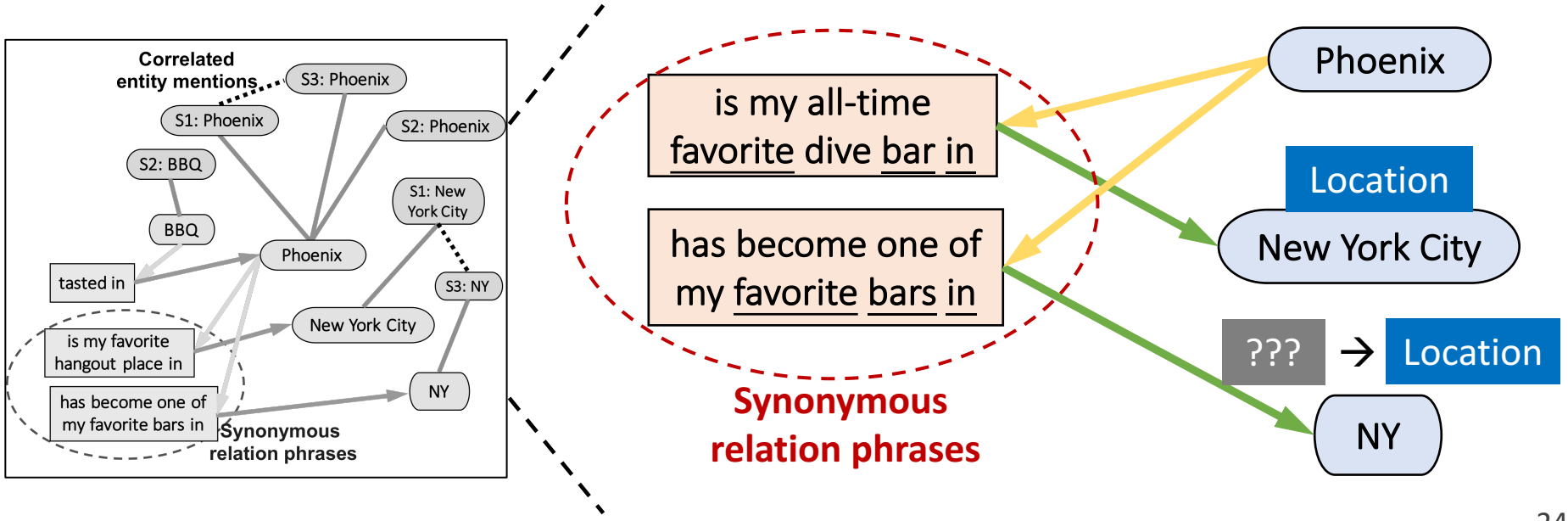
Measure of Non-Smoothness

Vector of scores for single label on nodes

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

Relation Phrase Clustering in **ClusType**

- Two relation phrases should be grouped together if:
  - Share similar string
  - Share similar context
  - Entity arguments' types are similar



Putting All Together: A Joint Optimization Framework

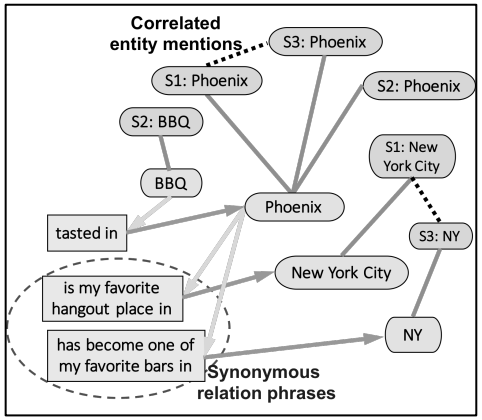
$$\mathcal{O}_{\alpha,\gamma,\mu} = \mathcal{F}(\mathbf{C}, \mathbf{P}_L, \mathbf{P}_R) + \mathcal{L}_\alpha\big(\mathbf{P}_L, \mathbf{P}_R, \{\mathbf{U}^{(v)}, \mathbf{V}^{(v)}\}, \mathbf{U}^*\big) + \Omega_{\gamma,\mu}(\mathbf{Y}, \mathbf{C}, \mathbf{P}_L, \mathbf{P}_R).$$

Type propagation between entity names and relation phrases

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

Mention correlation & mention type modeling

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



$$\sum_{v=0}^{d} \beta^{(v)} \big(\|\mathbf{F}^{(v)} - \mathbf{U}^{(v)}\mathbf{V}^{(v)T}\|_F^2 + \alpha\|\mathbf{U}^{(v)}\mathbf{Q}^{(v)} - \mathbf{U}^*\|_F^2\big)$$

Multi-view relation phrases clustering

ClusType: Comparing with State-of-the-Art Systems

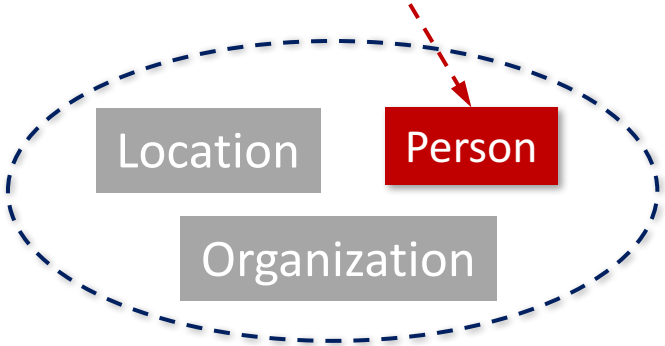
Methods	NYT (118k 2013 news articles)	Yelp (230k business reviews)	Tweet (302k tweets)
Pattern (Stanford, CONLL'14)	0.301	0.199	0.223
SemTagger (U Utah, ACL'10)	0.407	0.296	0.236
NNPLB (UW, EMNLP'12)	0.637	0.511	0.246
APOLLO (THU, CIKM'12)	0.795	0.283	0.188
FIGER (UW, AAAI'12)	0.881	0.198	0.308
ClusType (KDD'15)	0.939	0.808	0.451

- **Pattern** (Stanford, CONLL'14): explicit textual pattern; semantic drift
- **NNPLB** (UW, EMNLP'12)**:** type propagation on surface name level (name ambiguity)
- **APOLLO** (THU, CIKM'12): **context sparsity** in type propagation
- **FIGER** (UW, AAAI'12): reliance on complex linguistic features (domain restriction)

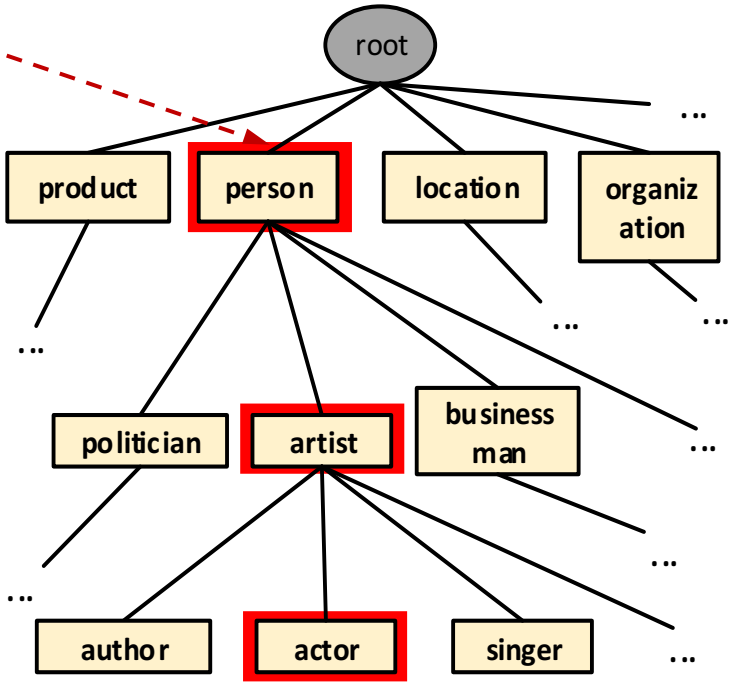
Precision (P) = $\frac{\#Correctly-typed\ mentions}{\#System-recognized\ mentions}$, Recall ($R$) = $\frac{\#Correctly-typed\ mentions}{\#ground-truth\ mentions}$, F1 score = $\frac{2(P\times R)}{(P+R)}$

Fine-Grained Entity Typing

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



A few common types

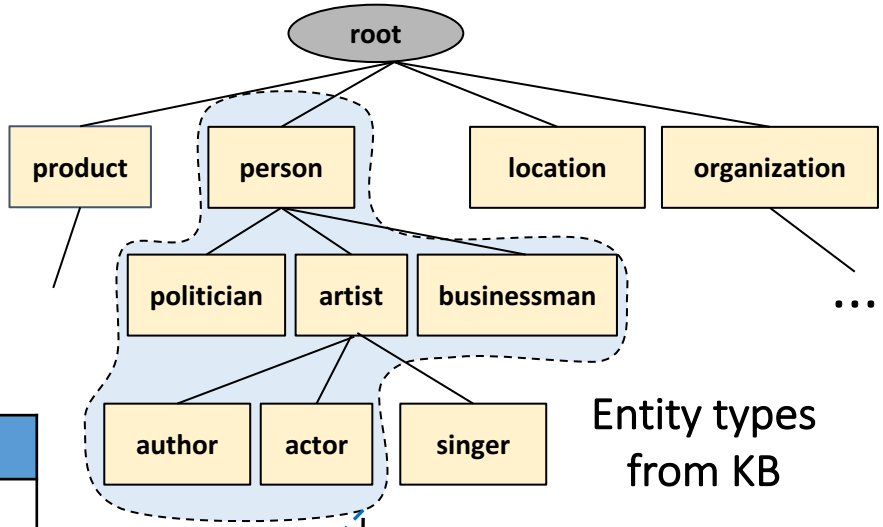


A type hierarchy with 100+ types

- **Features for deeper NLP tasks**
  - Relation extraction (Ling & Weld, 2012)
- **Assists downstream applications**
  - Question answering

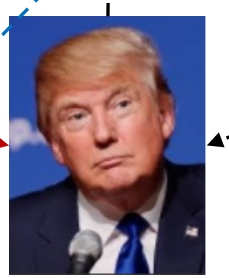
Current Distant Supervision: Context-Agnostic Labeling

- "Context-agnostic" type assignment in **training data**
- **Prior work**: all labels are "perfect" training labels



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

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



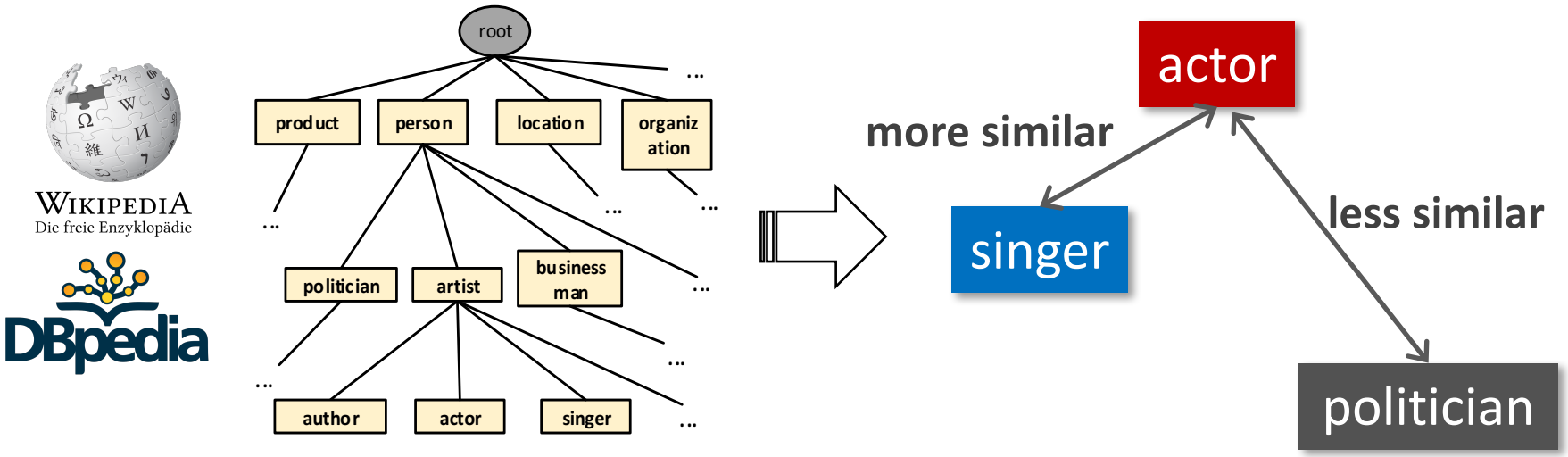
Entity: ***Donald Trump***

WIKIPEDIA
Die freie Enzyklopädie

DBpedia

Current Distant Supervision: "Type Independence" Assumption

- Entity types are not independent → **correlated**



- Existing studies ignore such correlation information

How to deal with infrequent (fine-grained) entity types?

My Solution: **P**artial **L**abel **E**mbedding (KDD'16)

Extract Text Features

Label Noise Reduction

Train Classifiers on De-noised Data

Prediction on New Data

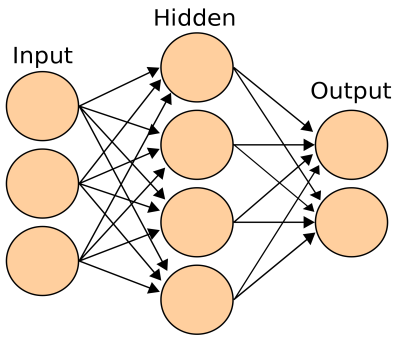
ID	Sentence
S1	*Donald Trump* spent 14 television seasons presiding over a business-themed game show, NBC's The Apprentice

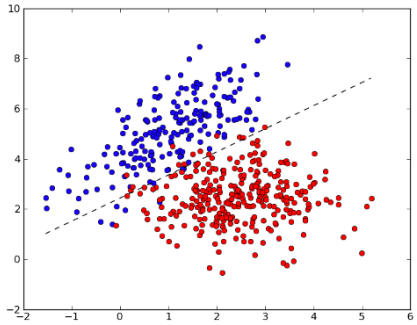
Text features: HEAD_Donald, CXT_A: television, CXT_A: season, POS: NN, TKN_trump, SHAPE: AA

S1: *Donald Trump*

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

De-noised labeled data





"Robust" classifier

Ren et al. *Label Noise Reduction in Entity Typing by Heterogeneous Partial-Label Embedding*. KDD, 2016.

PLE: Modeling Clean and Noisy Mentions Separately

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

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

S1: *Donald Trump*

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

Types ranked

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

"Best" candidate type

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

Measured based on currently estimated embedding space

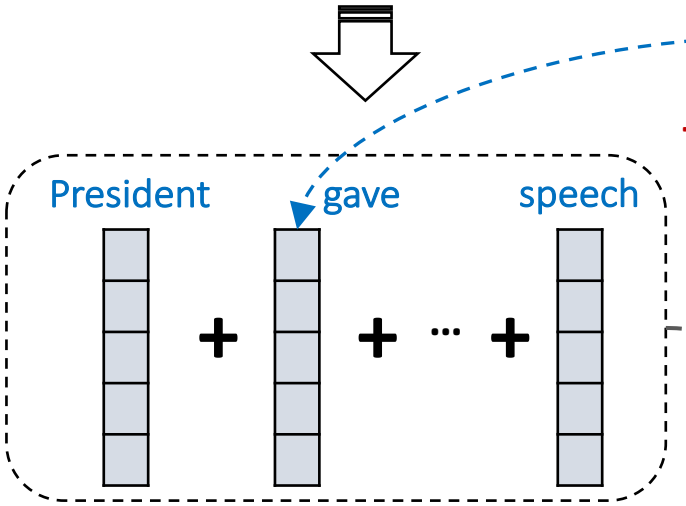
For a noisy mention, its "*best candidate type*" should be **ranked higher** than all its "*non-candidate types*"

Ren et al. *AFET: Automatic Fine-Grained Entity Typing by Hierarchical Partial-Label*. EMNLP, 2016.

Type Inference in PLE

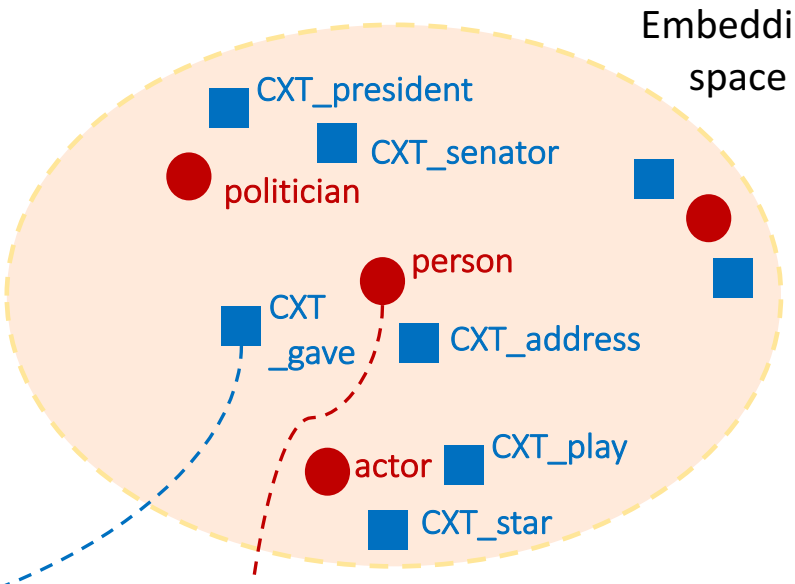
- Top-down nearest neighbor search in the given type hierarchy

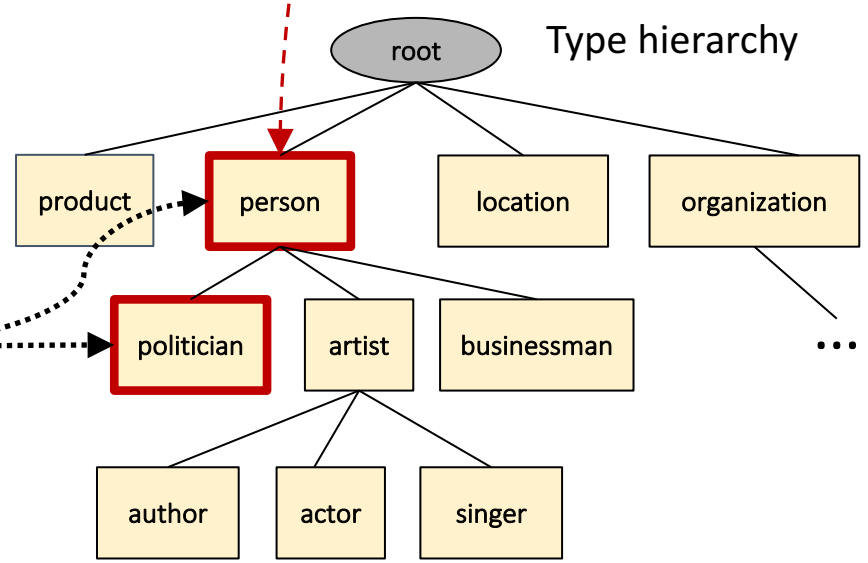
ID	Sentence
S_i	President ***Trump*** gave an all-hands address to troops at the U.S. Central Command headquarters



Embedding vectors for text features

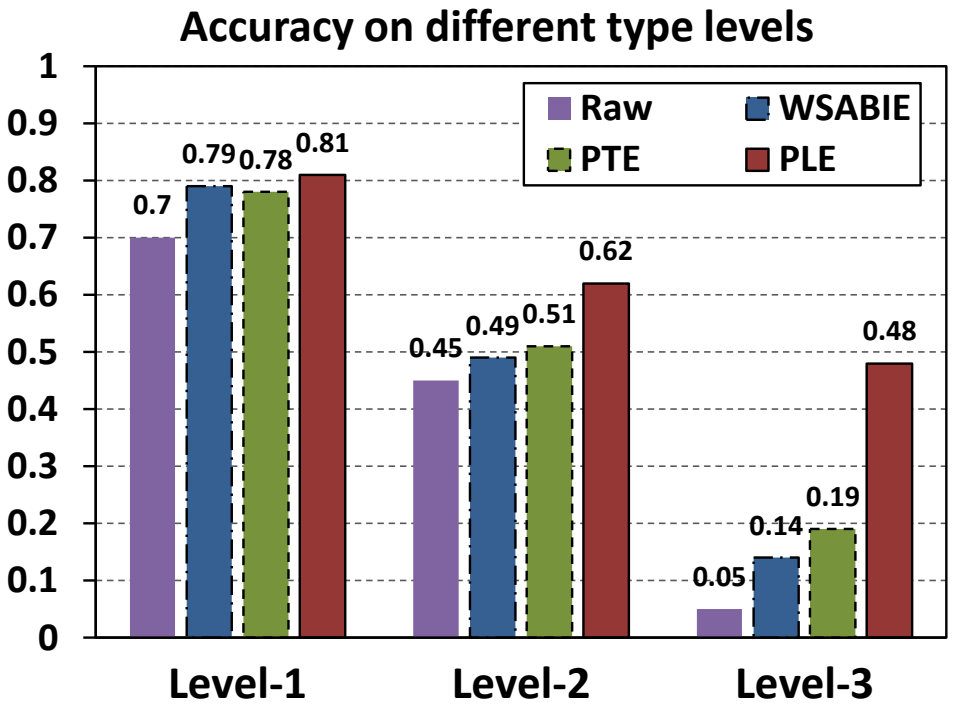
Test mention: S_i_*Trump*





PLE: Performance of Fine-Grained Entity Typing

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



- **Raw**: candidate types from distant supervision
- **WASBIE** (Google, ACL'14): joint feature and type embedding
- **PTE** (MSR, WWW'15): joint mention, feature and type embedding
  - Both WASBIE and PTE suffer from context-agnostic labels
- **PLE** (KDD'16): partial-label loss + type correlation modeling

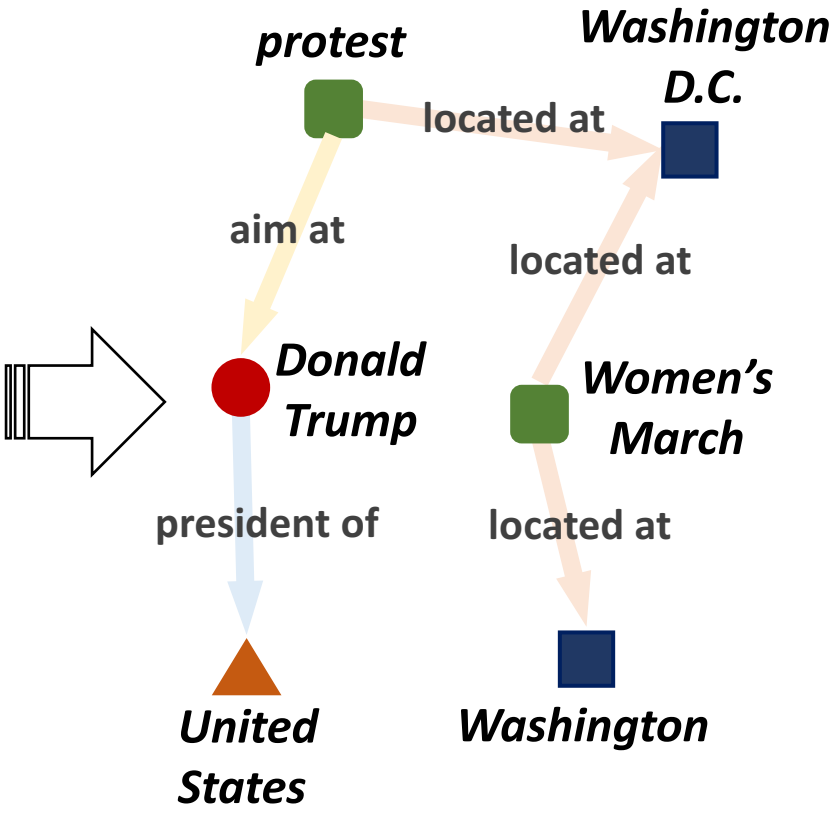
OntoNotes dataset (Weischedel et al. 2011, Gillick et al., 2014): 13,109 news articles, 77 annotated documents, 89 types
https://catalog.ldc.upenn.edu/LDC2013T19

Outline

- Introduction
- Entity Recognition and Typing [KDD'15, KDD'16]
- **Joint Entity and Relation Extraction** [WWW'17]
- Summary and Future Directions

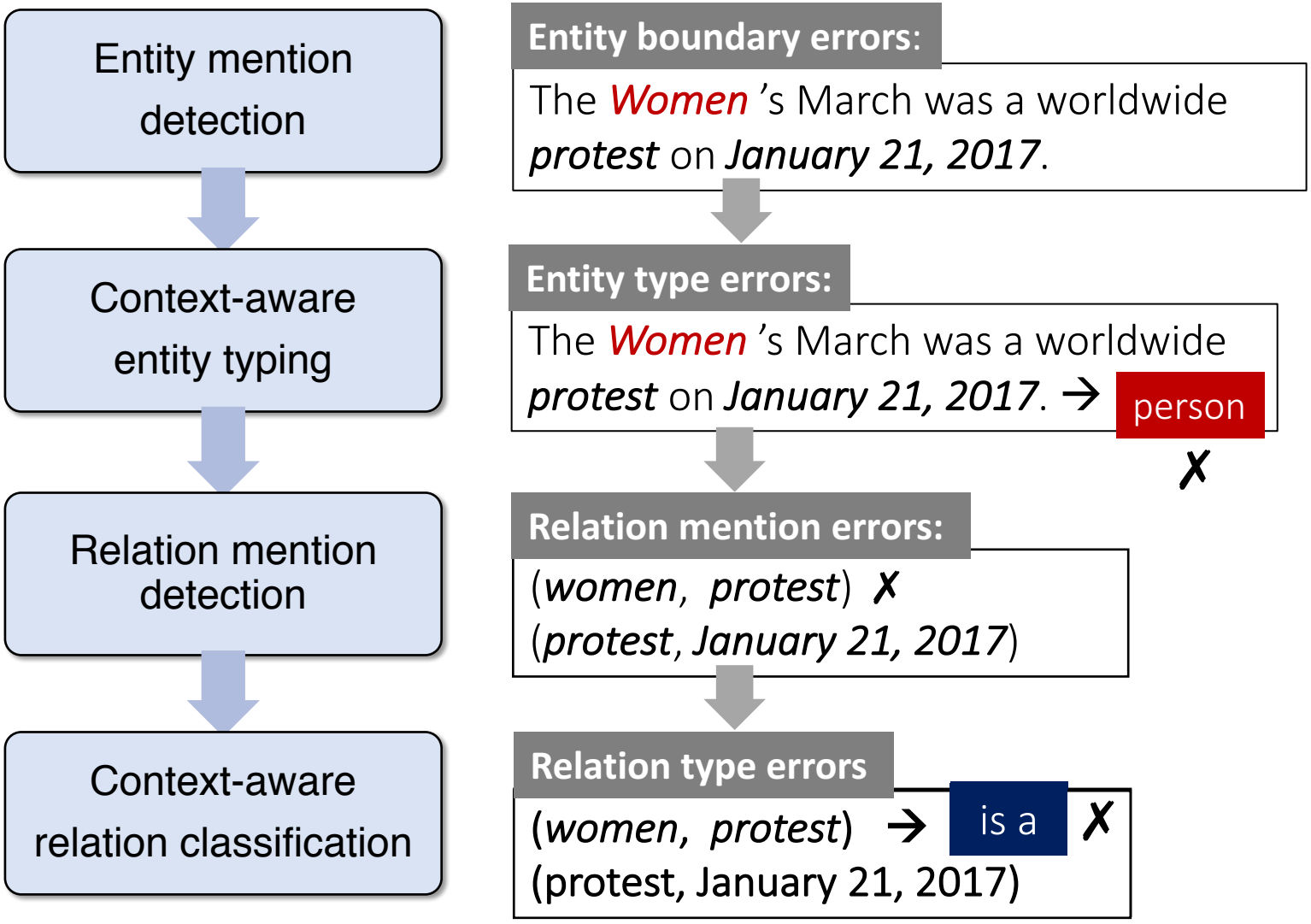
Extraction of Typed Entities and Relations

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

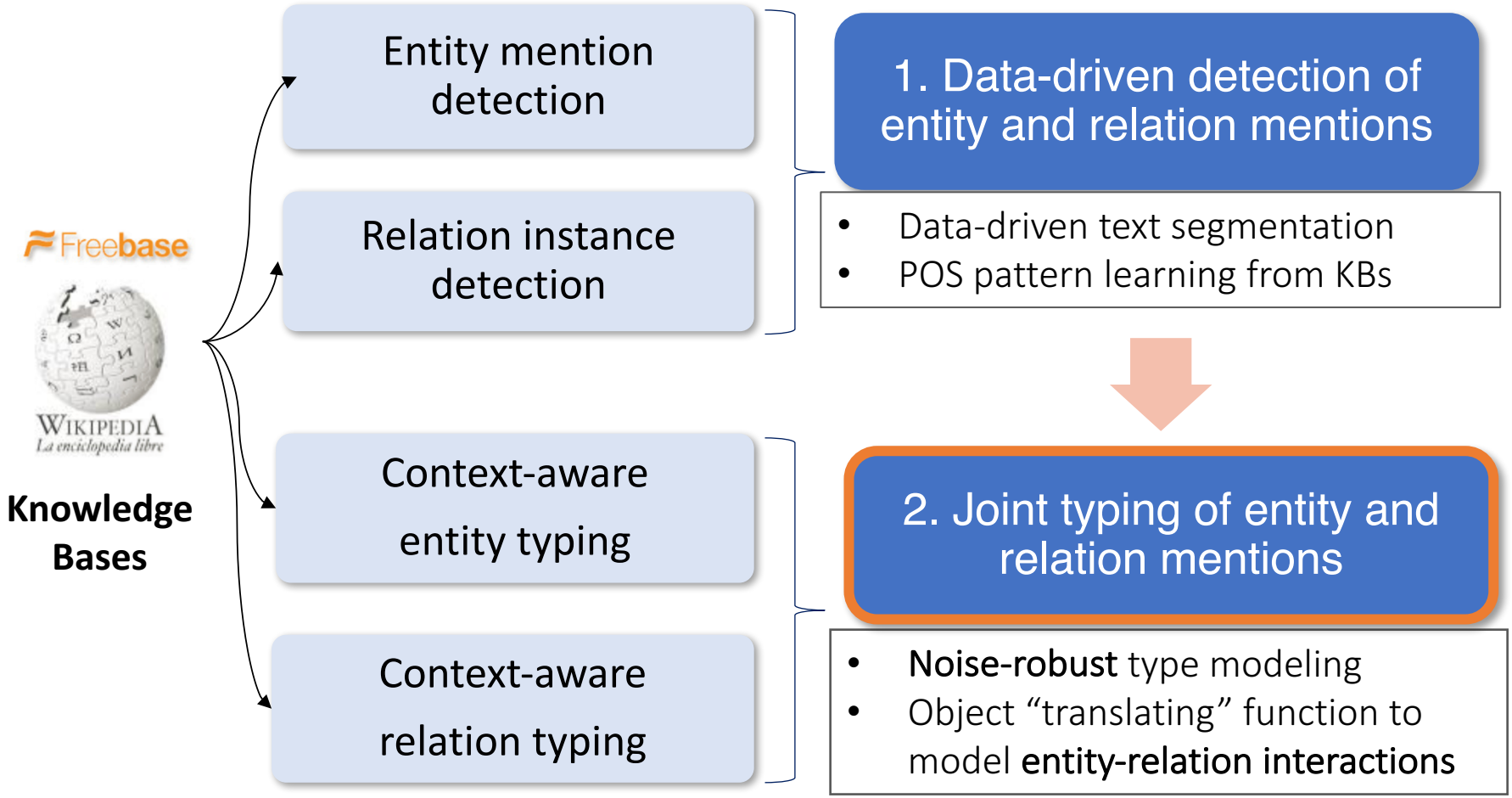


Prior Work: An "Incremental" System Pipeline

Error propagation cascading down the pipeline

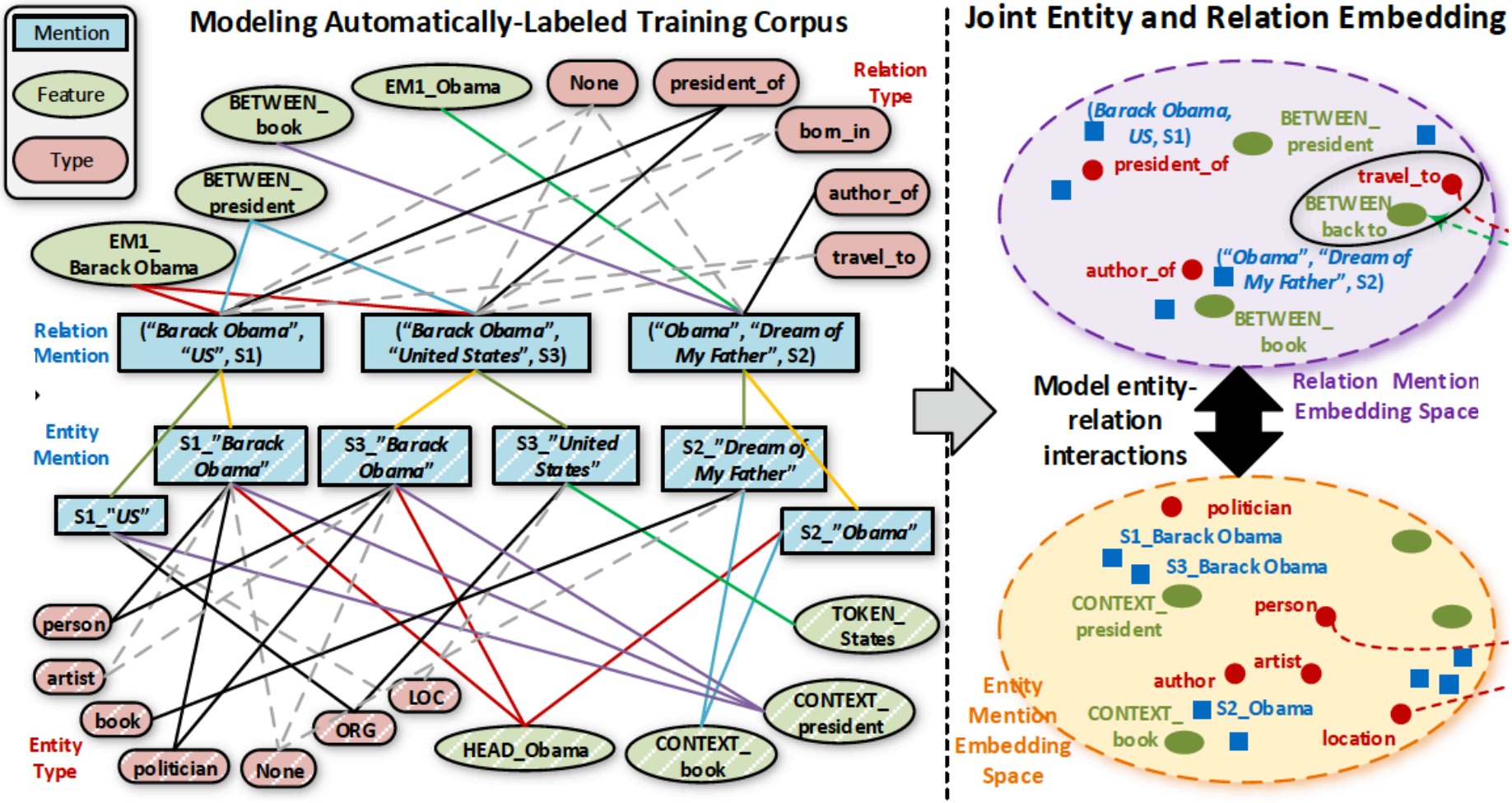


My Solution: **CoType** (WWW'17)



Ren et al. *CoType: Joint Extraction of Typed Entities and Relations with Knowledge Bases*. WWW, 2017.

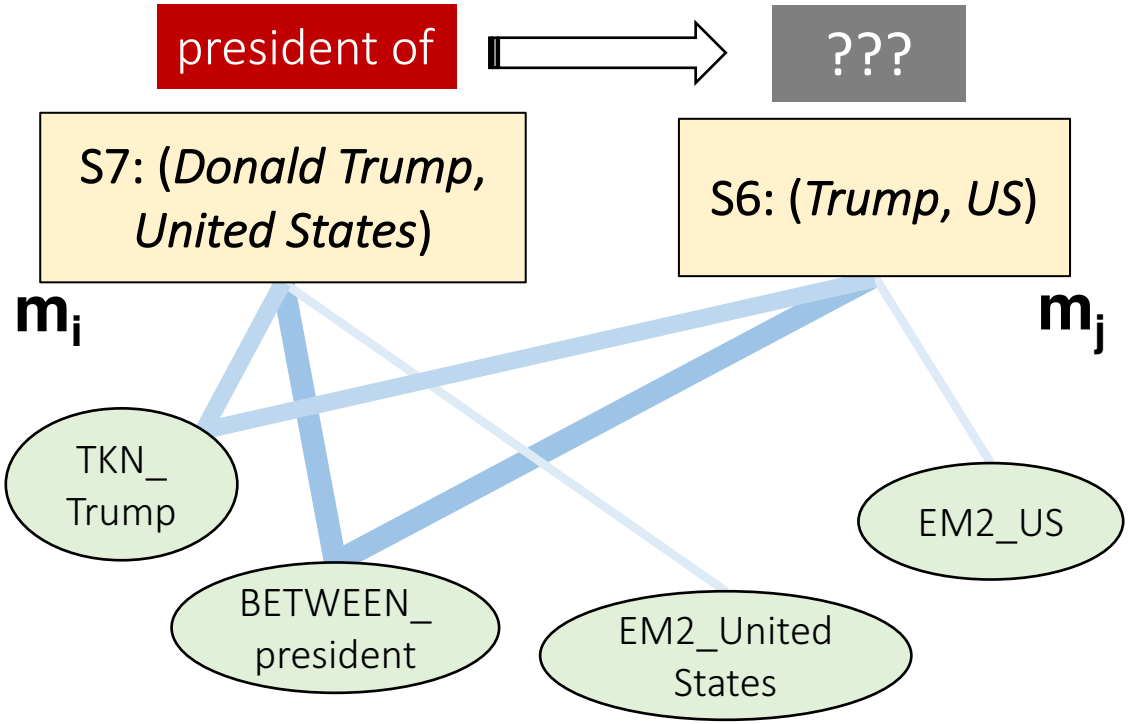
CoType: Co-Embedding for Typing Entities and Relations



Modeling Mention-Feature Co-Occurrences

- Second-order Proximity
  - Mentions with similar distributions over text features should have similar types (*i.e.*, close to each other in the latent space)

Vertex $\mathbf{m_i}$ and $\mathbf{m_j}$ have a large second-order proximity

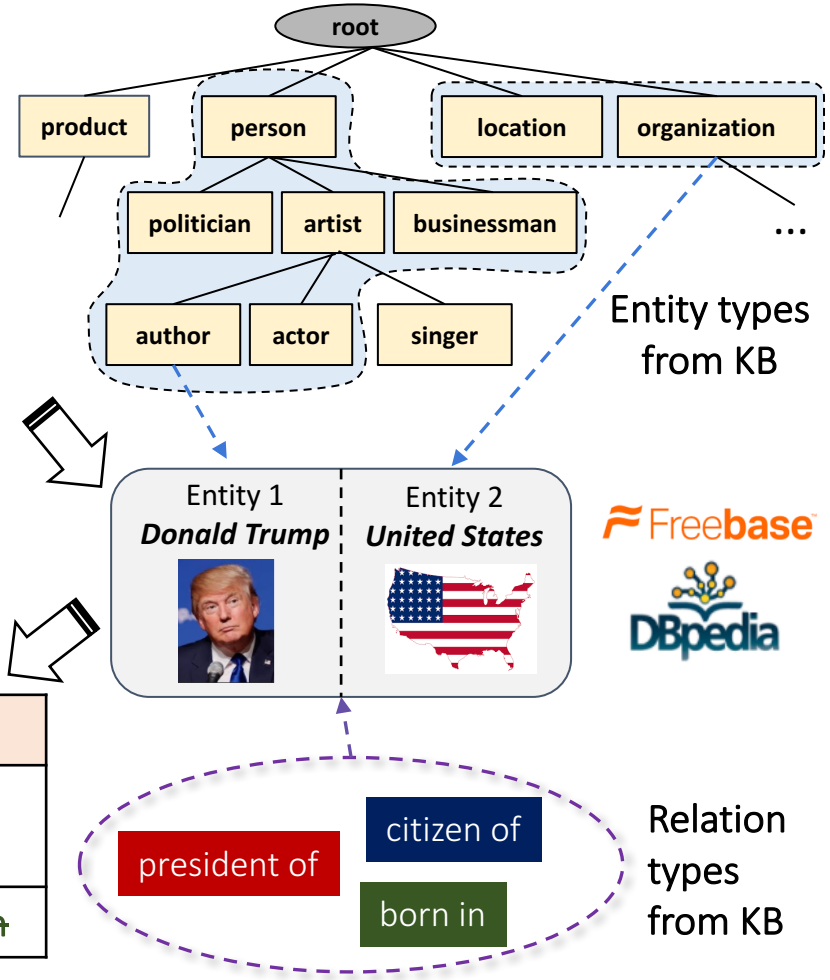


Current Distant Supervision: Context-Agnostic Labeling

ID	Sentence
S1	*Donald Trump* was born in Queens, New York, *USA* on June 14,1946.
S2	The protest was aimed at *Donald Trump*, the recently inaugurated president of the *United States*.
S3	There is a method to *Donald Trump*'s madness and he laid it all out in his book, "*The Art of the Deal*."

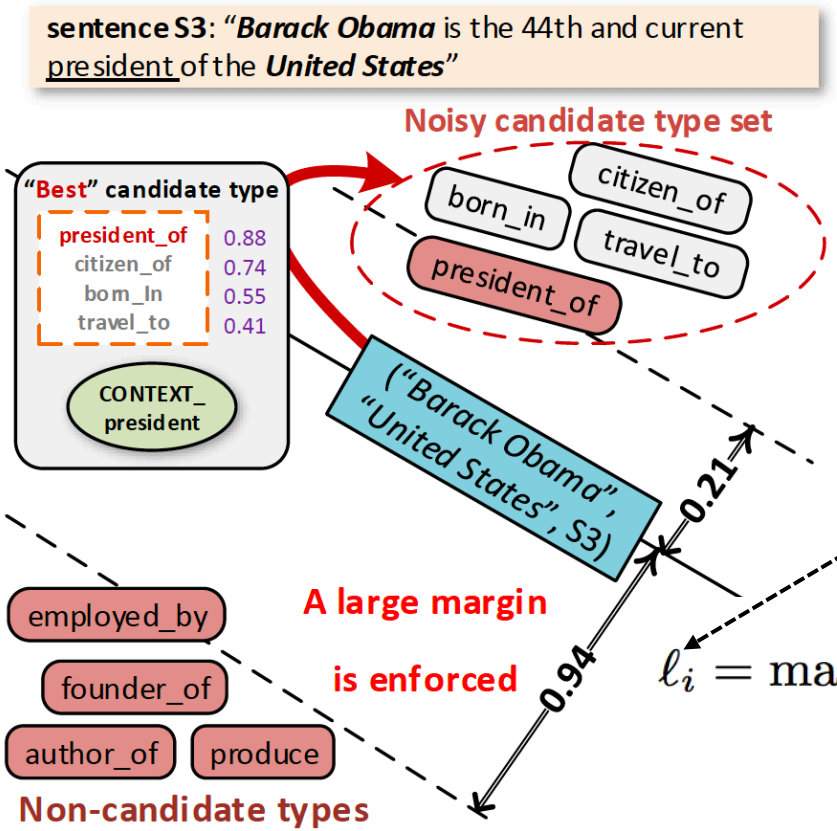
Type labels for relation mention in S2:

E1: *Donald J. Trump*	**E2**: *United States*
E1 Types: person, politician, ~~businessman~~, ~~author~~, ~~actor~~	**E2 Types**: location, ~~organization~~
Relations between E1, E2: president of, ~~citizen of~~, ~~born in~~	



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

Context-Aware Type Modeling



Partial-label Loss Function

- A relation mention should be **more similar** to its "**most relevant**" **candidate type**, than to any other **non-candidate type**

Enforce: mention is more similar to its **"most relevant" candidate type**

$$\ell_i = \max\left\{0, 1 - \left[\max_{y \in \mathcal{Y}_i} s(m_i, y) - \max_{y' \in \overline{\mathcal{Y}}_i} s(m_i, y')\right]\right\}$$

Score for "most relevant" type

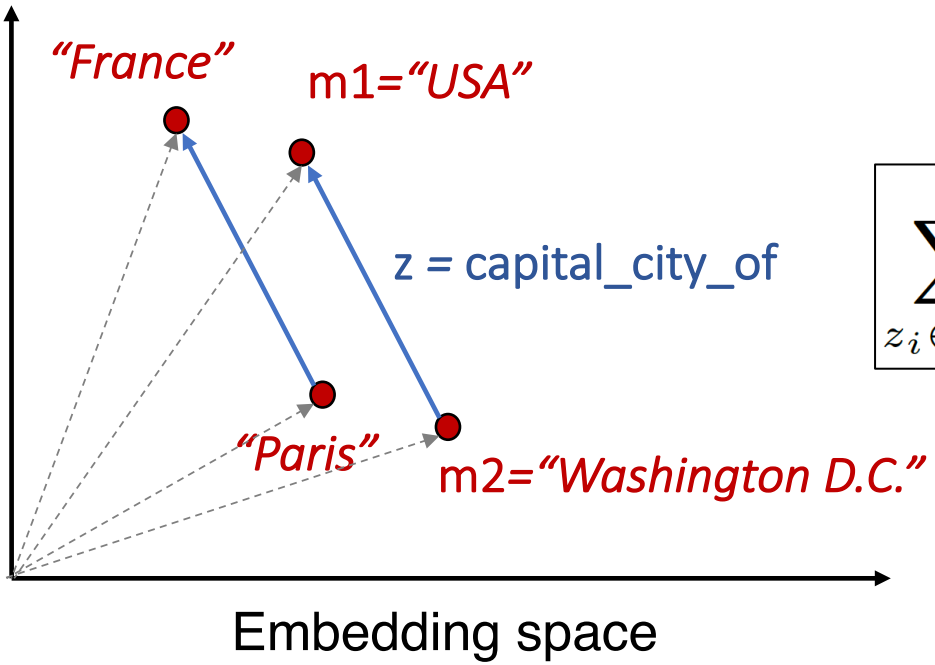
Maximal score for non-candidate types

Modeling Entity-Relation Interactions

Object "Translating" Assumption

For a relation mention **z** of entity mentions **m1** and **m2**,

vec(m1) ≈ **vec**(m2) + **vec**(z)



- Error on an entity-relation triple (z, m1, m2):

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

Enforce: error on a positive triple should be smaller than error on a negative triple

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

positive relation triple

negative relation triple

Reducing Error Propagation: A Joint Optimization Framework

Modeling **Entity-relation interactions**

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

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

Modeling **relation mentions**

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

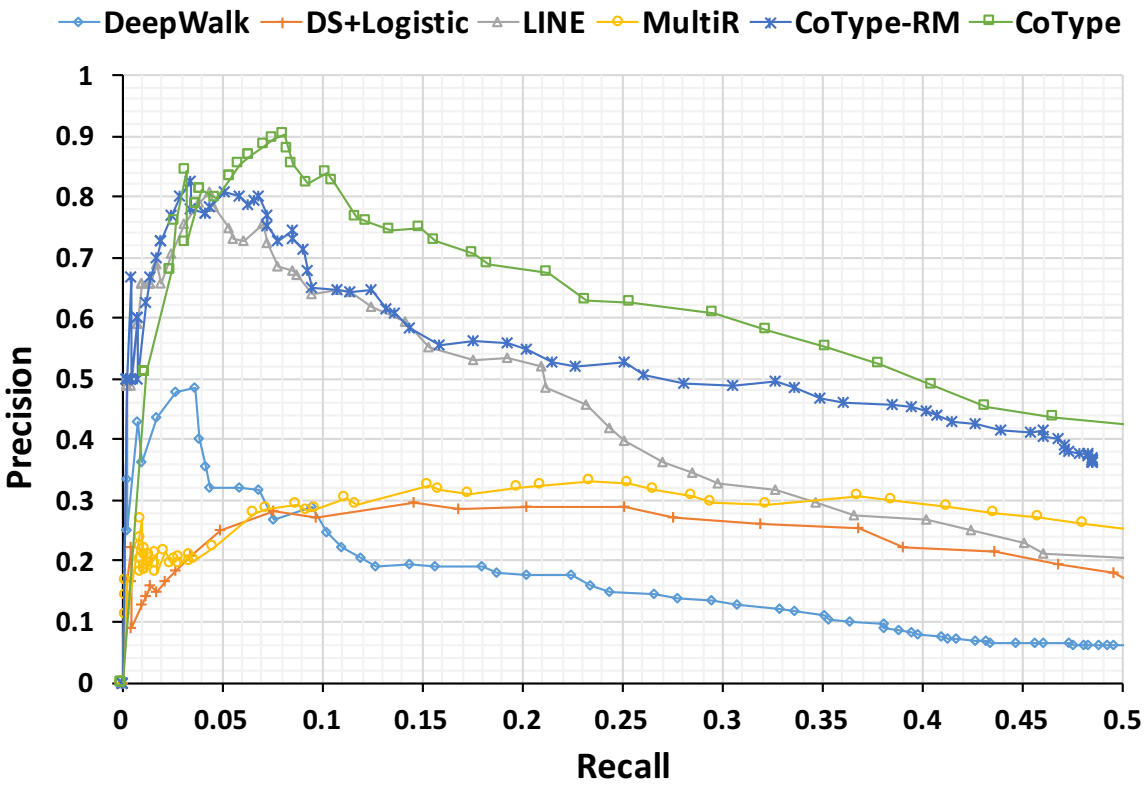
Modeling **entity mentions**

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

Details of the formulas can be found in:
Ren et al. *CoType: Joint Extraction of Typed Entities and Relations with Knowledge Bases*. WWW, 2017.

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

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



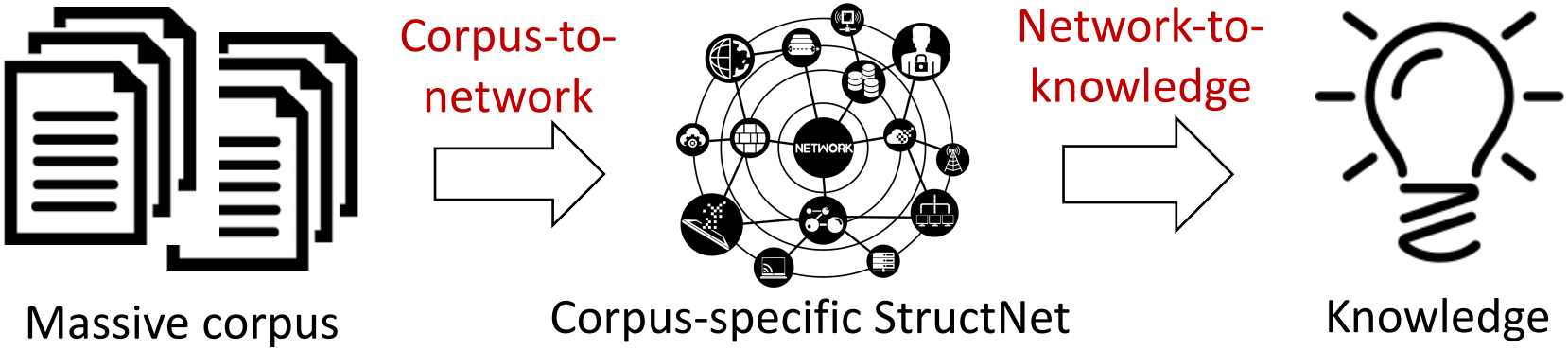
- DeepWalk (StonyBrook, KDD'14): homogeneous graph embedding
- DS+Logistic (Stanford, ACL'09): trains logistic classifier on DS
- LINE (MSR, WWW'15): joint feature and type embedding
- MultiR (UW, ACL'11): distantly-supervised, models noisy labels
- CoType-RM (WWW'17): only models relation mentions
- CoType (WWW'17): models entity-relation interactions

Outline

- Introduction
- Entity Recognition and Typing [KDD'15, KDD'16]
- Joint Entity and Relation Extraction [WWW'17]
- **Summary and Future Directions**

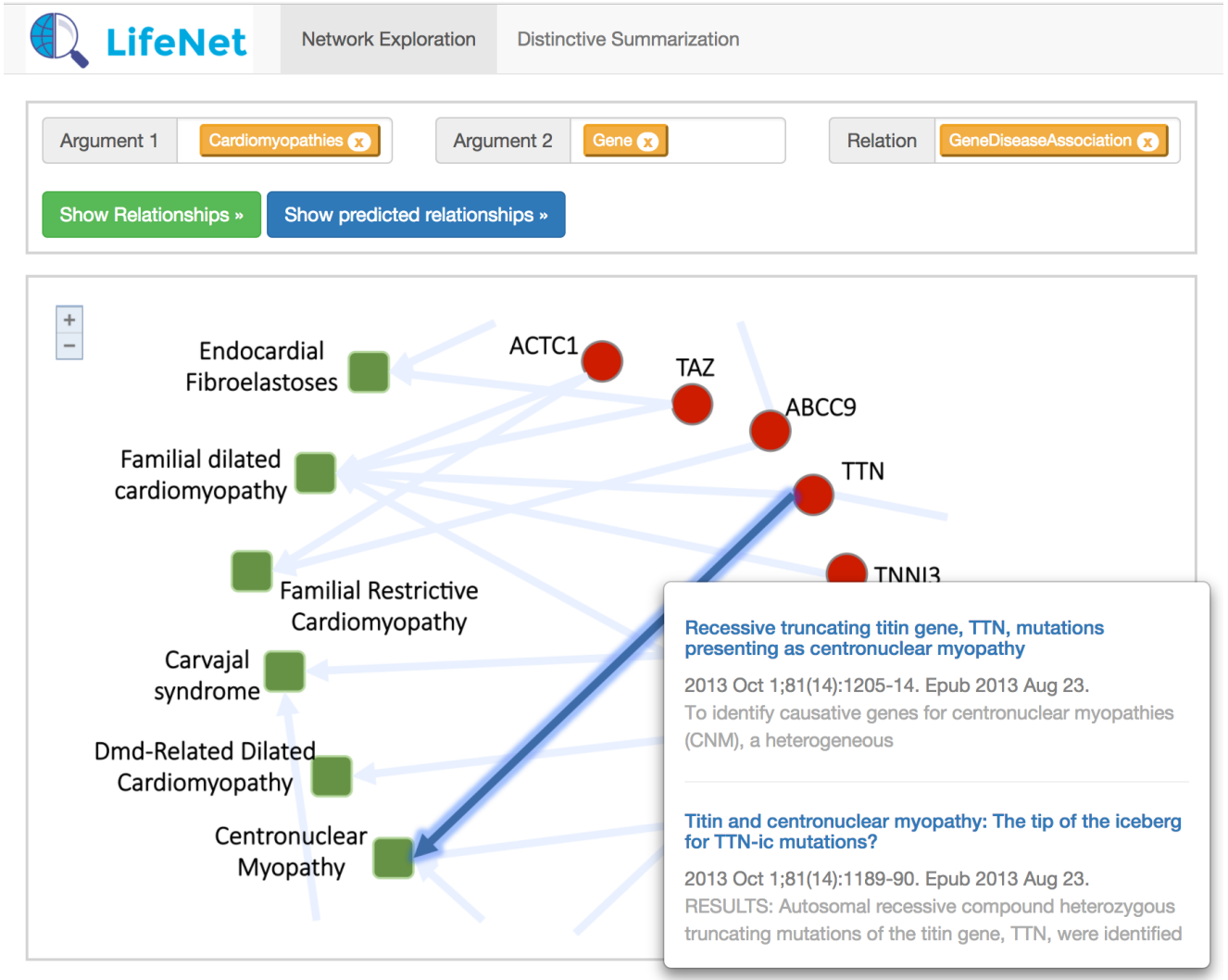
Overall Contributions

- Study the "Corpus-specific StructNet Construction" problem
- Create a novel framework: "Effort-Light StructMine"
- Apply the framework to solve three subtasks to progressively construct StructNet
- A principled approach to explore and analyze "Big Text Data"



Ongoing Application of Effort-Less StructMine

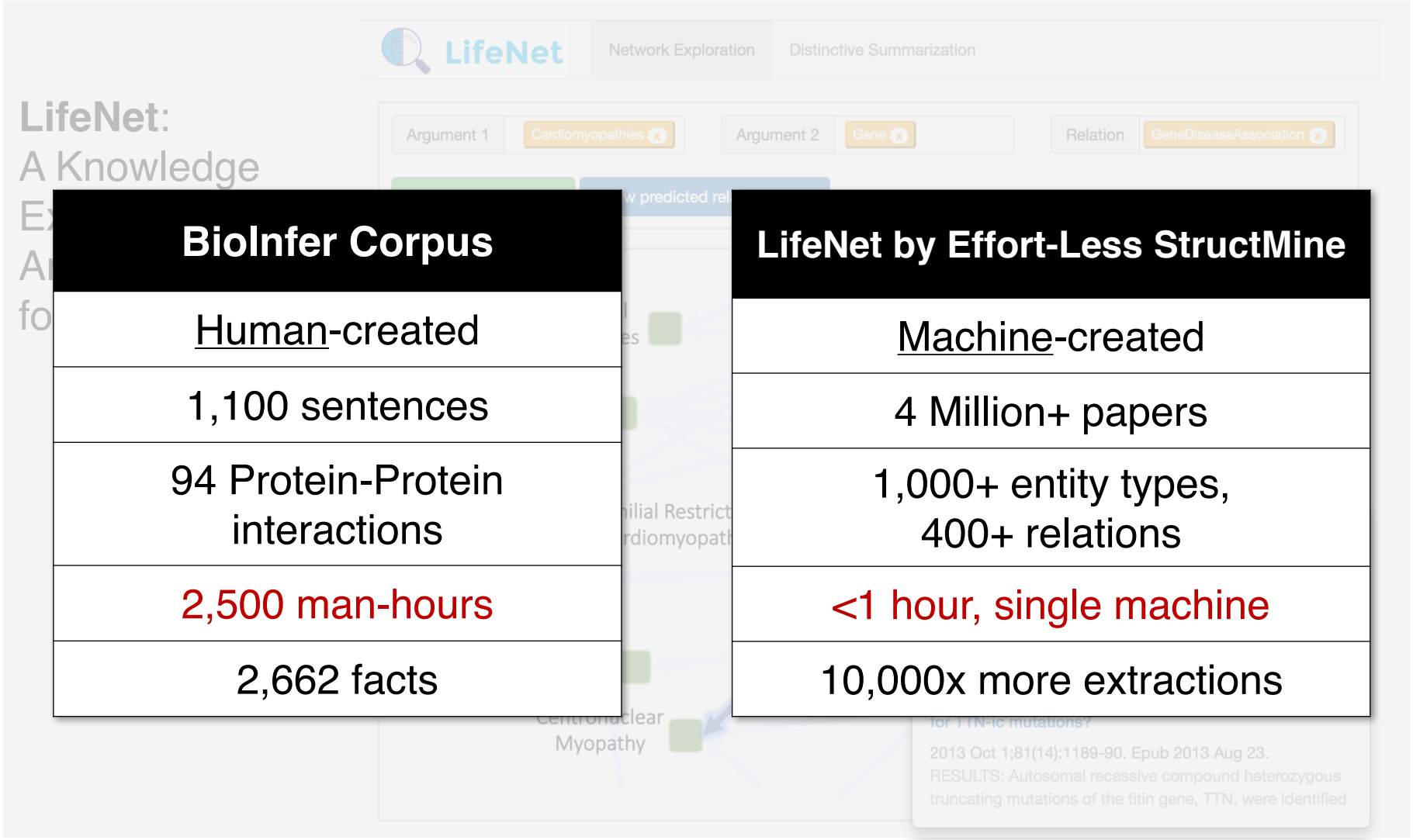
LifeNet:
A Knowledge Exploration and Analytics System for Life Sciences



BioInfer: a corpus for information extraction in the biomedical domain, BMC Bioinformatics, 2007
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1808065/

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

Looking Forward: Applications on Life Sciences



BioInfer Corpus
Human-created
1,100 sentences
94 Protein-Protein interactions
2,500 man-hours
2,662 facts

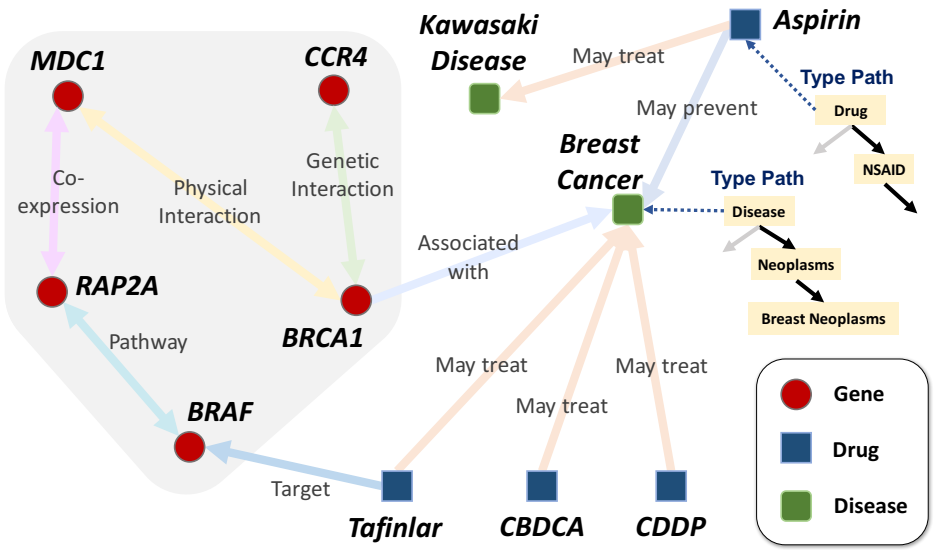
LifeNet by Effort-Less StructMine
Machine-created
4 Million+ papers
1,000+ entity types, 400+ relations
<1 hour, single machine
10,000x more extractions

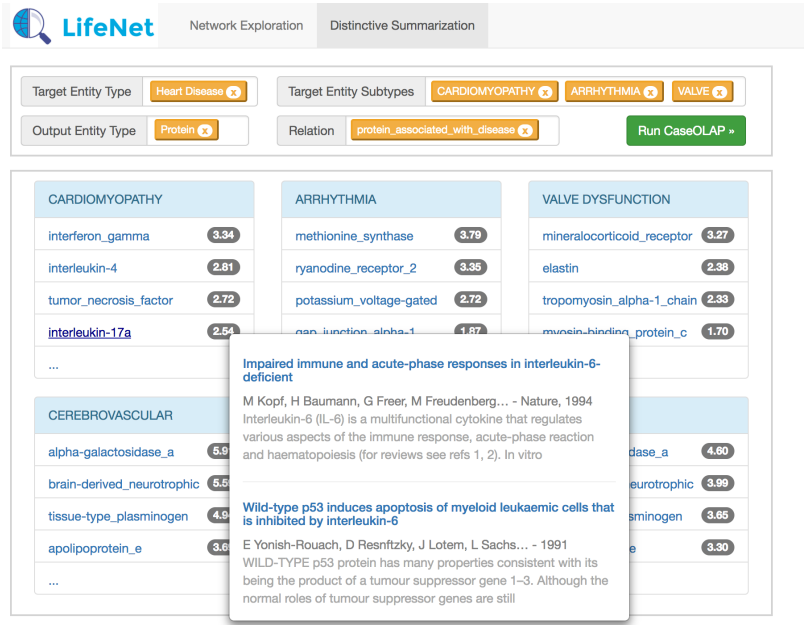
BioInfer: a corpus for information extraction in the biomedical domain, BMC Bioinformatics, 2007
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1808065/

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

Looking Forward: Mining StructNets for Scientific Research

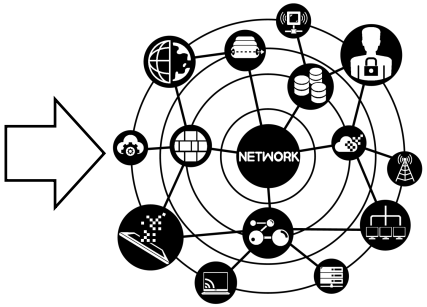
- Literature → StructNet → Knowledge Exploration and Analytics
- Collaborate with **life scientists, physicits, computer scientists.**
- ClusCite (KDD'14), Comparative Document Analysis (WSDM'17), FacetGist (CIKM'16)



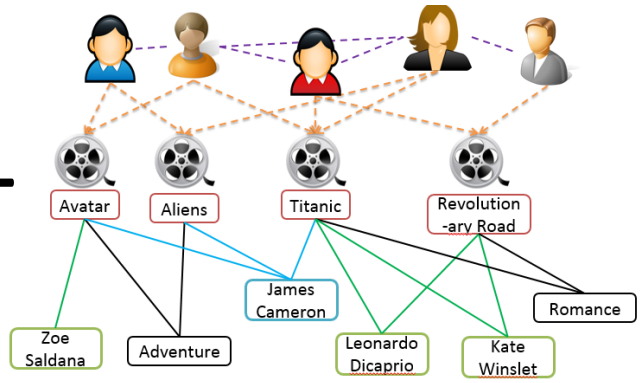


Looking Forward: Engaging with Human Behaviors

- StructNet + User Behavioral Data → Intelligent Systems
  - Social networks, transaction records
- User-generated Content to StructNet: Human Behavior Modeling?
  - Social media posts, customer reviews, fictions, etc.
- Collaborate with **social / political scientists, HCI researchers**
- Personalized entity recommendation (WSDM'14a, RecSys'13)

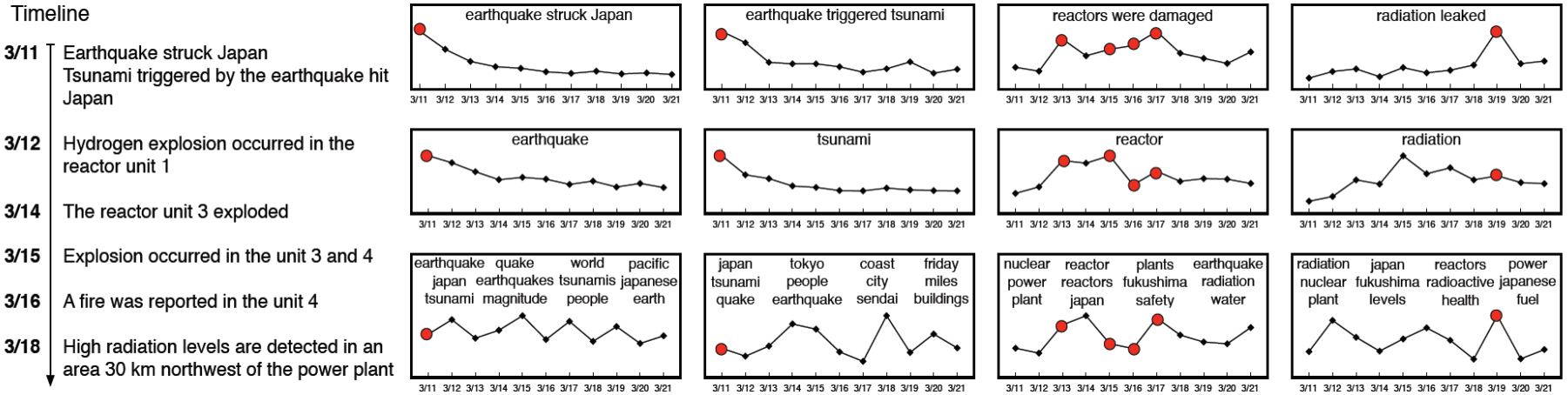


+



Looking Forward: Integrating with Physical World

- Text signals (e.g., social media posts) + sensor signals (geo-sensors in phones) → better smart city operating systems
- Collaborate with **networking / system researchers**, **environmental scientist**
- Event-centric summarization of massive corpora (ICDM'13)



Acknowledgement

- Academic Collaborators

- Industry Collaborators

- Funding

Q&A

Work done at UIUC	My Research Publications
Phrase mining	SDM'14, SIGMOD'15
Entity recognition and typing	KDD'15, KDD'16, EMNLP'16
Relation Extraction	WWW'17
Entity synonym mining	WWW'15
Facet discovery	CIKM'16
Automatic summarization	ICDM'13, WWW'16, WSDM'17
Entity recommendation in text-rich environment	RecSys'13, WSDM'14a, WSDM'14b

Work mentioned in this talk

Reference I

- **Xiang Ren**, Zeqiu Wu, Wenqi He, Meng Qu, Clare R. Voss, Heng Ji, Tarek F. Abdelzaher, Jiawei Han. CoType: Joint Extraction of Typed Entities and Relations with Knowledge Bases. WWW, 2017.
- **Xiang Ren**, Ahmed El-Kishky, Heng Ji, and Jiawei Han. Automatic Entity Recognition and Typing in Massive Text Data (Conference Tutorial). SIGMOD, 2016.
- **Xiang Ren***, Wenqi He*, Meng Qu, Lifu Huang, Heng Ji, Jiawei Han. AFET: Automatic Fine-Grained Entity Typing by Hierarchical Partial-Label Embedding. EMNLP, 2016.
- **Xiang Ren***, Wenqi He*, Meng Qu, Heng Ji, Clare R. Voss, Jiawei Han. Label Noise Reduction in Entity Typing by Heterogeneous Partial-Label Embedding. KDD, 2016.
- **Xiang Ren**, Wenqi He, Ahmed El-Kishky, Clare R. Voss, Heng Ji, Meng Qu, Jiawei Han. Entity Typing: A Critical Step for Mining Structures from Massive Unstructured Text (Invited Paper). MLG, 2016.
- **Xiang Ren**, A. El-Kishky, C. Wang, F. Tao, C. R. Voss, H. Ji, J. Han. ClusType: Effective Entity Recognition and Typing by Relation Phrase-Based Clustering. KDD, 2015.
- **Xiang Ren**, Tao Cheng. Synonym Discovery for Structured Entities on Heterogeneous Graphs. WWW, 2015.
- Tarique A. Siddiqui*, **Xiang Ren***, Aditya Parameswaran, Jiawei Han. FacetGist: Collective Extraction of Document Facets in Large Technical Corpora. CIKM, 2016.
- Jialu Liu, Jingbo Shang, Chi Wang, **Xiang Ren**, Jiawei Han. Mining Quality Phrases from Massive Text Corpora. SIGMOD, 2015.

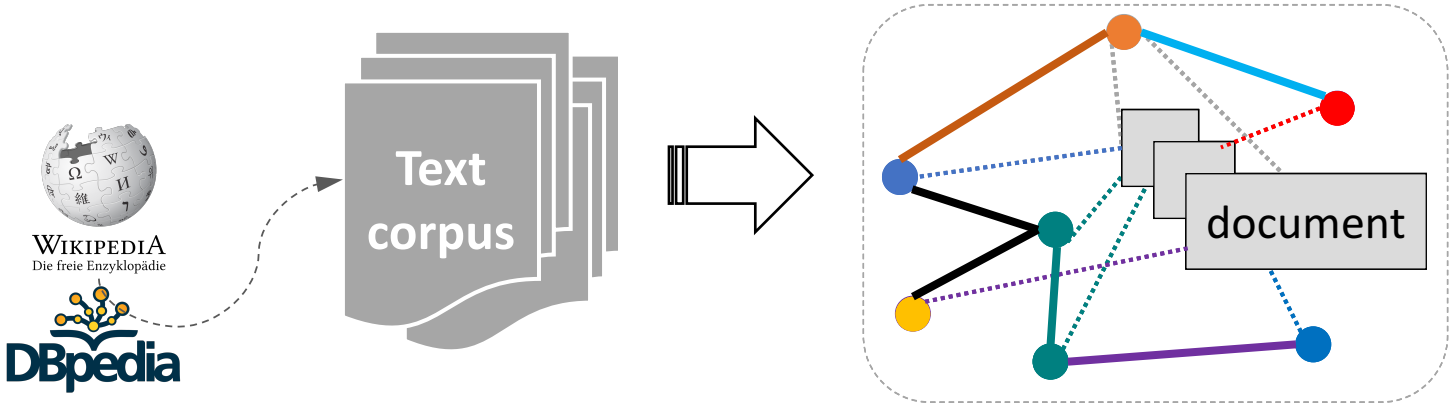
Reference II

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

Backup Slides

Effort-Less StructMine

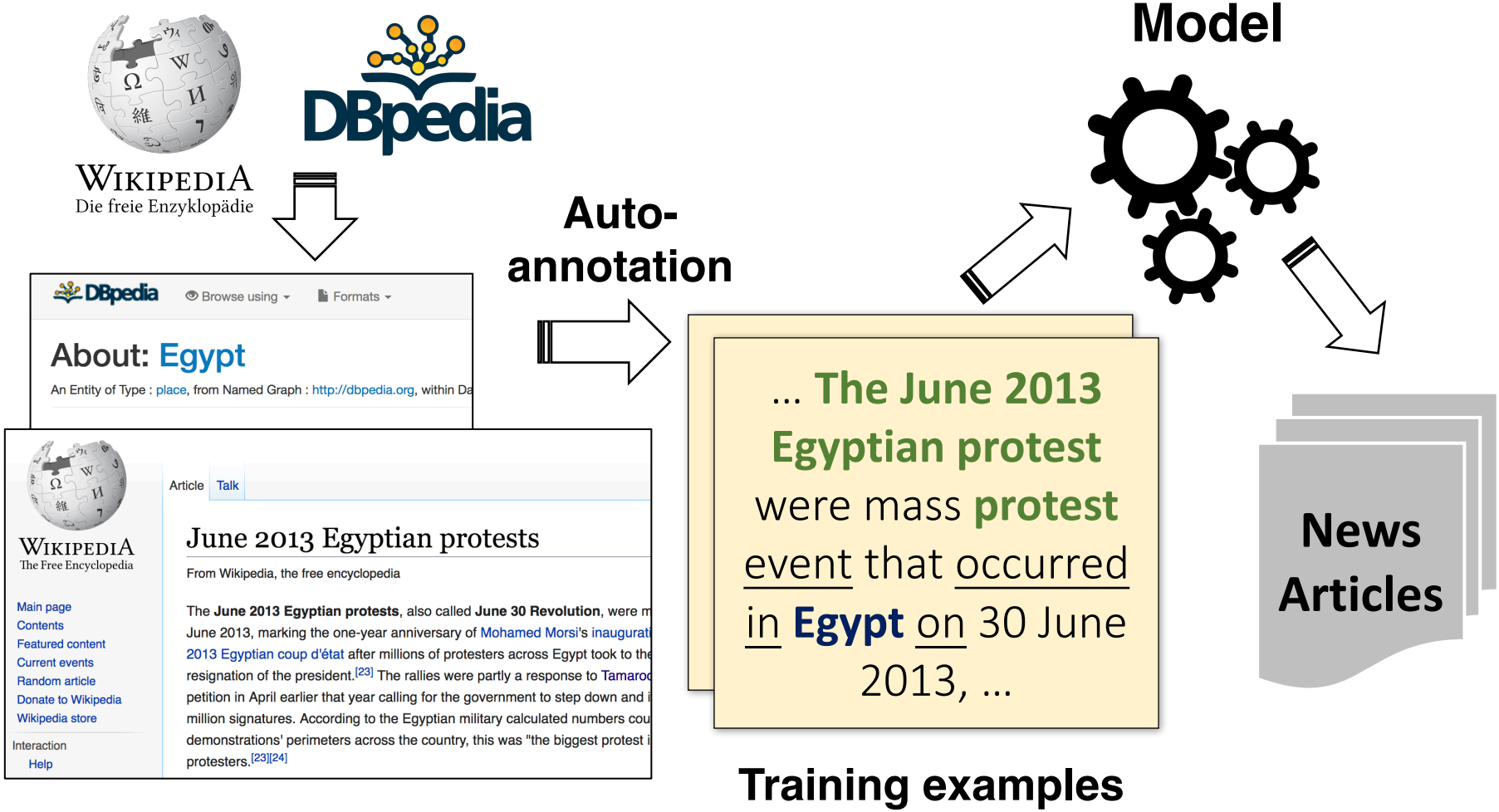
- **StructMine:** mining factual structures from given corpus
- **Minimal-Effort:** (1) no explicit human labeling; (2) light-weight feature engineering



ClusType: Entity Recognition and Typing (KDD'15) → Fine-grained Entity Typing (KDD'16, EMNLP'16) → CoType: Joint Entity and Relation Extraction (WWW'17)

Corpus to Structured Network: The Roadmap

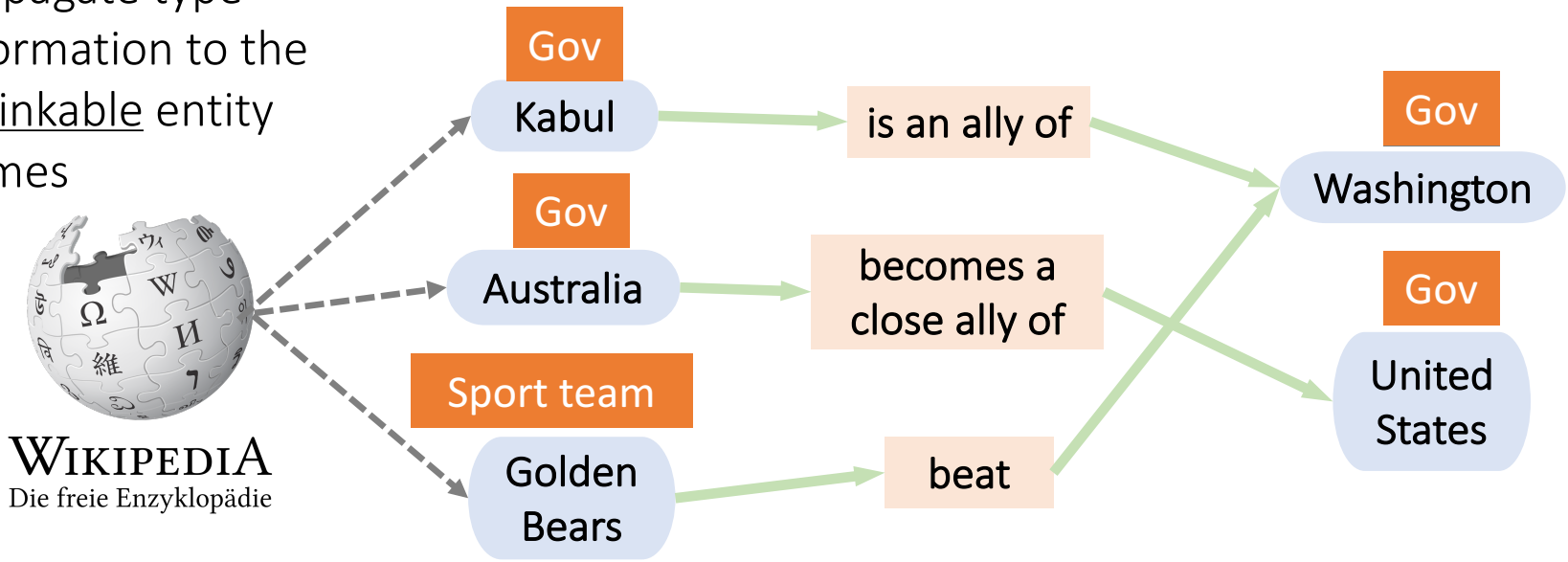
My Work: Mining Factual Structures with **Minimal** Human Involvement



My General Methodology: Distant Supervision with Knowledges Bases

1. Detect entity names from text
2. Link entity names to KB entities
3. Propagate type information to the unlinkable entity names

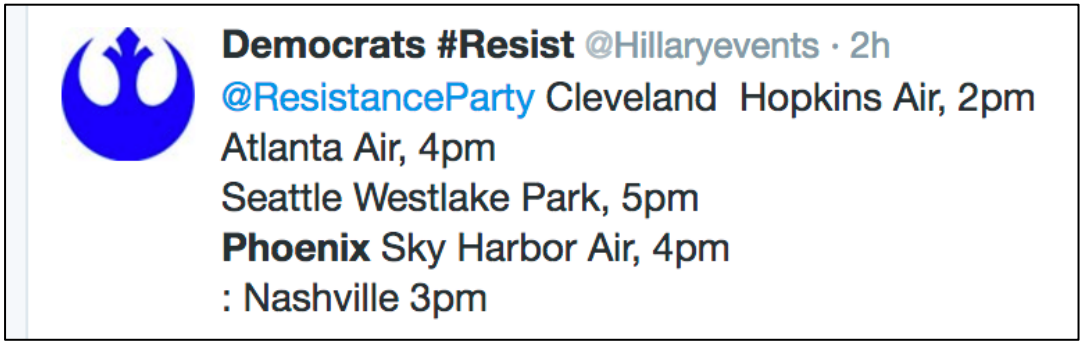
ID	Text
S1	… has concerns whether ***Kabul*** is an ally of ***Washington***.
S2	… ***Australia*** becomes a close ally of the ***United States***.
S3	The ***Cardinal*** will share the title with ***California*** if the ***Golden Bears*** beat ***Washington*** later Saturday.

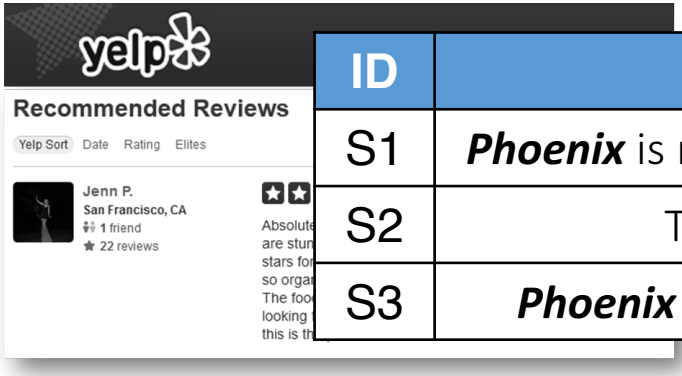


Lin et al. *No noun phrase left behind: detecting and typing unlinkable entities*. EMNLP, 2012.

Current Distant Supervision: Limitation 1

1. Domain restriction:
   - Name detectors trained on one domain/genre (news) are hard to be ported to other corpora (tweets)





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

My Solution: **ClusType** [KDD'15]

Data-driven entity mention detection algorithm

- No human annotated data & less linguistic assumption → ~~Limitation 1: domain restriction~~

Do not merge entity mentions with identical name strings

- Model each entity mention based on its surface name & surrounding context → ~~Limitation 2: name ambiguity~~

Mine synonymous relation phrases simultaneously

- Consolidate "connecting bridges" enables effective type propagation → ~~Limitation 3: context sparsity~~

Ren et al. *ClusType: Effective Entity Recognition and Typing by Relation Phrase-Based Clustering*. KDD, 2015.

ClusType: Comparing with Sequence Model

- How does sequence models trained on general-domain, grammatical corpus perform across different corpora?

Methods	NYT (118k 2013 news articles)	Yelp (230k restaurant reviews)	Tweet (302k tweets)
Stanford NER (2014 version)	0.682	0.240	0.438
ClusType (KDD'15)	0.942	0.594	0.472

- Stanford NER: a linear-chain CRF classifier
  - For three entity types: Person, Location, Organization
  - Trained on a mixture of CoNLL, MUC and ACE corpora (manually-annotated, general-domain, grammatical text).
  - The 2014-10-26 version model is used for comparison
- Challenges: irregular text (tweets, reviews), dynamic domain (news)

Stanford NER: http://nlp.stanford.edu/software/CRF-NER.html

My Solution: **AFET** (EMNLP'16)

- Jointly embed entity mentions and type labels into a low-dimensional vector space (to capture type semantics)

- Design a noise-robust loss function to model "false positive" type labels in noisy training data

- Enforce adaptive margin on entity mentions, to encode type correlation

Ren et al. *AFET: Automatic Fine-Grained Entity Typing by Hierarchical Partial-Label*. EMNLP, 2016.

AFET (EMNLP'16): Framework Overview

Extract Text Features

Partition Training Set

Joint Embedding on Training Sets

Prediction on New Data

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

Text features: HEAD_Donald, CXT_A: television, CXT_A: season, POS: NN, TKN_trump, SHAPE: AA

S1: *Donald Trump*

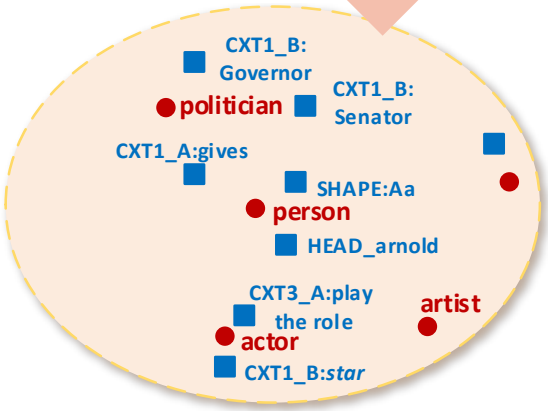
Entity Types: person, artist, actor, author, businessman, politician

Noisy mention

S4: *Ted Cruz*

Entity Types: person, politician

Clean mention



text feature

type label

Performance Comparison on Fine-Grained Typing

Typing Method	Wiki			OntoNotes			BBN		
	Acc	Ma-F1	Mi-F1	Acc	Ma-F1	Mi-F1	Acc	Ma-F1	Mi-F1
CLPL (Cour et al., 2011)	0.162	0.431	0.411	0.201	0.347	0.358	0.438	0.603	0.536
PL-SVM (Nguyen and Caruana, 2008)	0.428	0.613	0.571	0.225	0.455	0.437	0.465	0.648	0.582
FIGER (Ling and Weld, 2012)	0.474	0.692	0.655	0.369	0.578	0.516	0.467	0.672	0.612
FIGER-Min (Gillick et al., 2014)	0.453	0.691	0.631	0.373	0.570	0.509	0.444	0.671	0.613
HYENA (Yosef et al., 2012)	0.288	0.528	0.506	0.249	0.497	0.446	0.523	0.576	0.587
HYENA-Min	0.325	0.566	0.536	0.295	0.523	0.470	0.524	0.582	0.595
ClusType (Ren et al., 2015)	0.274	0.429	0.448	0.305	0.468	0.404	0.441	0.498	0.573
HNM (Dong et al., 2015)	0.237	0.409	0.417	0.122	0.288	0.272	0.551	0.591	0.606
DeepWalk (Perozzi et al., 2014)	0.414	0.563	0.511	0.479	0.669	0.611	0.586	0.638	0.628
LINE (Tang et al., 2015b)	0.181	0.480	0.499	0.436	0.634	0.578	0.576	0.687	0.690
PTE (Tang et al., 2015a)	0.405	0.575	0.526	0.436	0.630	0.572	0.604	0.684	0.695
WSABIE (Yogatama et al., 2015)	0.480	0.679	0.657	0.404	0.580	0.527	0.619	0.670	0.680
AFET-NoCo	0.526	0.693	0.654	0.486	0.652	0.594	0.655	0.711	0.716
AFET-NoPa	0.513	0.675	0.642	0.463	0.637	0.591	0.669	0.715	0.724
AFET-CoH	0.433	0.583	0.551	0.521	0.680	0.609	0.657	0.703	0.712
AFET	**0.533**	**0.693**	**0.664**	**0.551**	**0.711**	**0.647**	**0.670**	**0.727**	**0.735**

- AFET vs. AFET-NoCo → gain from incorporating type correlation
- AFET vs. AFET-NoPa → gain from noise-robust loss function

AFET: Performance of Fine-Grained Entity Typing

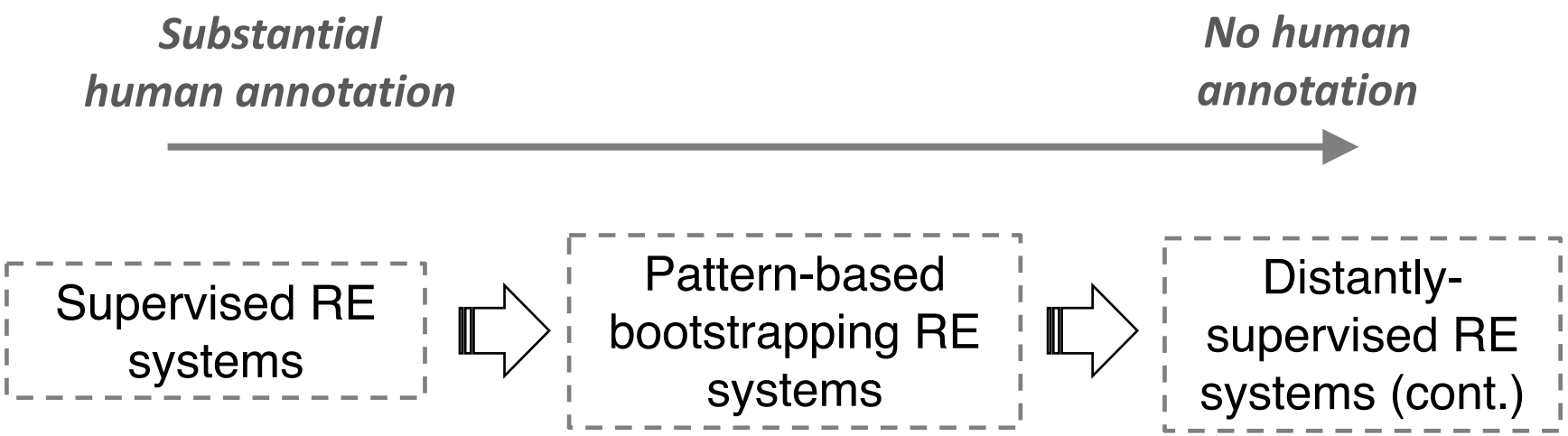
$$\text{Accuracy} = \frac{\#\ mentions\ with\ all\ types\ correctly\ predicted}{\#\ unseen\ entity\ mentions\ in\ the\ test\ set}$$

	Methods	Wikipedia	OntoNotes	BBN
Fine-grained classifiers	**FIGER** (UW, AAAI'12)	0.474	0.369	0.467
	HYENA (Max-Planck, COLING'12)	0.288	0.249	0.523
Embedding-based Methods	**WASABIE** (Google, ACL'14)	0.480	0.404	0.619
	HNM (IBM, EMNLP'15)	0.237	0.122	0.551
	PTE (MSR, WWW'15)	0.405	0.436	0.604
Partial-label Learning	**PL-SVM** (Cornell, KDD'08)	0.428	0.225	0.465
	AFET (EMNLP'16)	0.533	0.551	0.670

- Partial-label loss for modeling noisy labels (vs. fine-grained classifier, embedding methods)
- Adaptive margins for capturing type correlation (vs. PL-SVM, all)

- Wikipedia dataset (Ling & Weld, 2012): 1.5M sentences, 113 types
- OntoNotes dataset (Weischedel et al. 2011, Gillick et al., 2014): 13,109 news articles, 89 types
- BBN dataset (Weischedel & Brunstein, 2005): 2,311 news articles, 93 types

Prior Work of Relation Extraction (RE)

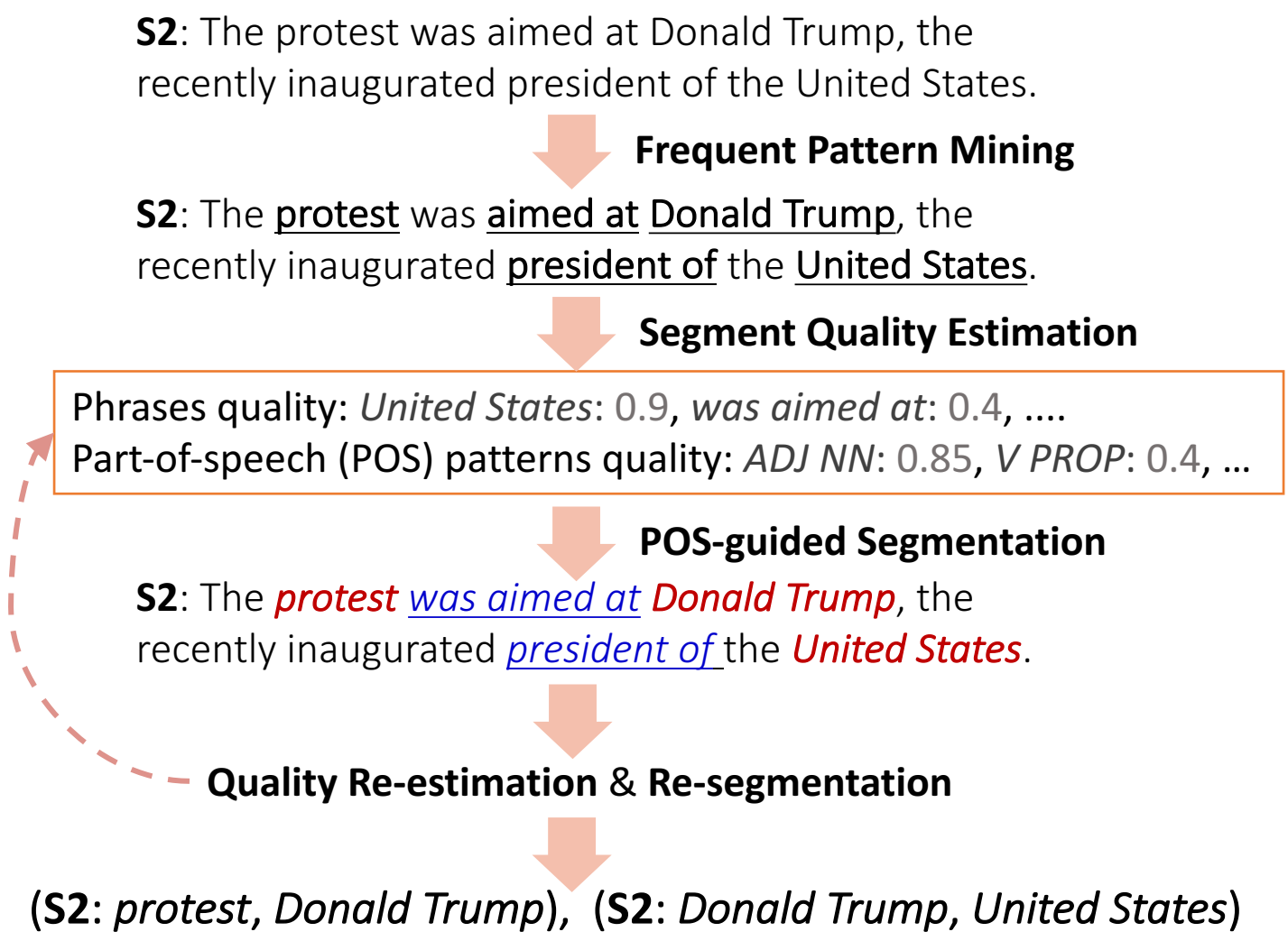


Mintz et al. *Distant supervision for relation extraction without labeled data*. ACL, 2009.

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

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

CoType Step 1: Data-Driven Entity and Relation Detection



Entity Mention Detection: Results

	POS Tag Pattern	Example
Good (high score)	*NNP NNP* *NN NN* *CD NN* *JJ NN*	San Francisco/Barack Obama/United States comedy drama/car accident/club captain seven network/seven dwarfs/2001 census crude oil/nucletic acid/baptist church
Bad (low score)	*DT JJ NND* *CD CD NN IN* *NN IN NNP NNP* *VVD RB IN*	a few miles/the early stages/the late 1980s 2 : 0 victory over/1 : 0 win over rating on rotten tomatoes worked together on/spent much of

	NYT	Wiki-KBP	BioInfer
FIGER segmenter [UW, 2012]	0.751	0.814	0.652
Our Approach	**0.837**	0.833	0.785

Key Insight: Text Co-occurrence Patterns Bring Semantic Power

… a speech was delivered by United States President ***Barack Obama***.

politician

President ***Donald Trump*** delivers a speech during …

President ***Donald Trump*** delivers a speech during …

Entity 1	Entity 2	Co-occurrences
speech	*Barack Obama*	57
United States	*Barack Obama*	217
president	*Barack Obama*	168

CoType: Performance of Entity Recognition and Typing

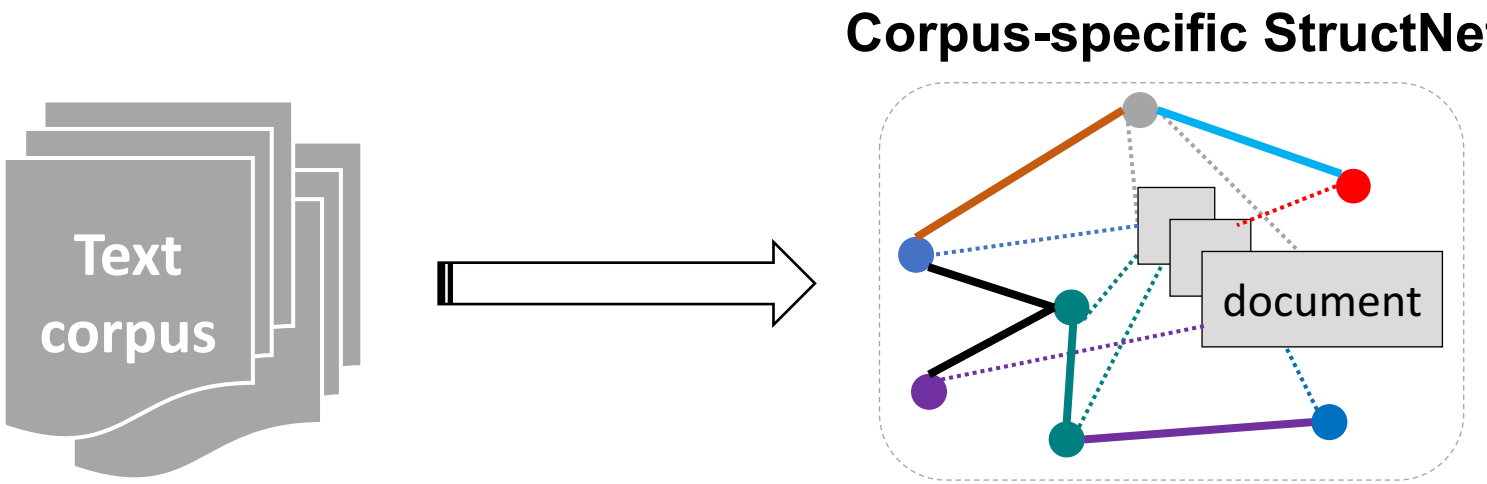
$$\text{Strict-F1 Score} = \frac{\#\ mentions\ with\ all\ types\ \ and\ boundary\ correctly\ predicted}{\#\ entity\ mentions\ in\ the\ test\ set}$$

	Methods	NYT	Wiki-KBP	BioInfer
Fine-grained classifiers	**FIGER** (UW, AAAI'12)	0.40	0.29	0.69
	HYENA (Max-Planck, COLING'12)	0.44	0.26	0.52
Embedding-based Methods	**WASABIE** (Google, ACL'14)	0.53	0.35	0.64
	DeepWalk (StonyBrook, KDD'14)	0.49	0.21	0.58
Noise-robust Embedding	**PLE** (KDD'16)	0.56	0.37	0.70
	CoType (WWW'17)	0.60	0.39	0.74

- Partial-label loss for noise-robust modeling of entities (vs. fine-grained classifiers, embedding-based methods)
- Modeling entity-relation interactions helps entity typing (vs. PLE)

- NYT dataset (Siedel el al., ECML'10): 1.18M sentences, 24 relation types, 47 entity types
- Wiki-KBP dataset (Ling et al. ACL'11, Ellis, TAC'14): 1.5M Wiki sentences, 19 relation types, 126 entity types
- BioInfer dataset (Pyysalo et al., BMC Informatics, 2007): 100k PubMed abstracts, 1,530 annotated sentences as test data, 94 relation types, 2k+ entity types

Method Effectiveness



Method	News	BioInfer
IBM FCM	0.681	0.467
UW MultiR	0.881	0.501
CoType (WWW'17)	**0.939**	**0.617**

relation classification accuracy

Applications of StructNets

Application	Technique	Publications
What are the keyphrases of the documents?	Keyphrase Extraction	WWW'16, SIGMOD'15
What are the commonalities and differences between two documents?	Comparative Document Analysis	WSDM'17
What are the major events in a corpus?	Event-Based Summarization	ICDM'13
What products should I recommend to users?	Entity Recommendation	WSDM'14a, RecSys'13
What are the important references a paper should cite?	Citation Prediction	KDD'14
What are the user search intents when people are using search engines?	Search Intent Learning	WSDM'14b