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
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
- ...

- Social media posts
- Web blogs
- News articles
- ...

- Corporate reports
- News streams
- Customer reviews
- ...

Healthcare

Computational Sociology

Business Intelligence
Prior Work: Mining Factual Structures with Human Efforts

Extraction Rules
Features / Models

Factual Structures
Women’s March
United States
Donald Trump

Human labeling

... 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 June 2013 Egyptian protest were mass protest event that occurred in Egypt on 30 June 2013, ...

Training data
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?

**Human labeling effort**

**Supervised learning systems**
- Stanford CoreNLP, 2005 - present
- UT Austin Dependency Kernel, 2005
- IBM Watson SIRE APIs

**Weakly-supervised learning systems**
- CMU NELL, 2009 - present
- UW KnowItAll, Open IE, 2005 - present
- Max-Planck PROSPERA, 2011

**Distantly-supervised Learning Systems**
- UW FIGER, MultiR, 2012
- Stanford MIML-RE, 2012

**Hand-crafted Systems**
- UCB Hearst Pattern, 1992
- NYU Proteus, 1997

Effort-Light StructMine (KDD’15, KDD’16, EMNLP’16, WWW’17)
My General Approach: Distant Supervision

- Text corpus
- Knowledge Bases as background knowledge

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

1% of 10M sentences → 100K labeled sentences

Publicly available in many domains:
- Common knowledge
- Biomedical sciences
- Arts
- ...

Number of Wikipedia articles

Actively enriched by human crowds

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

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

Effort-Light StructMine: Key Ideas

Data Sparsity

Propagate type information via “textual bridges” + consolidate “similar” bridges

Context-Agnostic Label Assignment

Select “best” label for context with specialized optimization objectives

Challenge

Key Idea

A Joint Optimization Framework
Effort-Light StructMine: Methodology

Text corpus

Data-driven text segmentation (SegPhrase, SIGMOD’15)

Candidate factual structures & text units

Distant supervision

Learning Corpus-specific Models (KDD’15, KDD’16, EMNLP’16, WWW’17)

Partially-labeled corpus

Structures from the unlabeled data

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
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. ...

- food
- location
- person
Traditional Named Entity Recognition (NER) Systems

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

The best [BBQ] I’ve tasted in [Phoenix].

<table>
<thead>
<tr>
<th>The</th>
<th>best</th>
<th>BBQ</th>
<th>I’ve</th>
<th>tasted</th>
<th>in</th>
<th>Phoenix</th>
</tr>
</thead>
<tbody>
<tr>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>0</td>
<td>O</td>
<td>O</td>
</tr>
</tbody>
</table>

Sequence model training

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

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

Seed entities and corpus

Annotate corpus using entities

Select Top patterns

Score candidate patterns

Generate candidate patterns

Apply patterns to find new entities

Patterns for President
the best <X> I’ve tried in their <X> tastes amazing
...

Systems:
UW KnowItAll
CMU NELL
...

Seeds for Food

Pizza
French Fries
Hot Dog
Pancake
...

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

<table>
<thead>
<tr>
<th>ID</th>
<th>Sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td><em>Phoenix</em> is my all-time favorite dive bar in <em>New York City</em>.</td>
</tr>
<tr>
<td>S2</td>
<td>The best <em>BBQ</em> I’ve tasted in <em>Phoenix</em>.</td>
</tr>
<tr>
<td>S3</td>
<td><em>Phoenix</em> has become one of my favorite bars in <em>NY</em>.</td>
</tr>
</tbody>
</table>

Previous Methods: Limitation 1

1. Context-agnostic type prediction
   - Predict types for entity mentions regardless of contexts

2. Textual bridge sparsity

<table>
<thead>
<tr>
<th>ID</th>
<th>Sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td><strong>Phoenix</strong> is my all-time favorite dive bar in <em>New York City</em>.</td>
</tr>
<tr>
<td>S2</td>
<td>The best <em>BBQ</em> I’ve tasted in <strong>Phoenix</strong>.</td>
</tr>
<tr>
<td>S3</td>
<td><strong>Phoenix</strong> has become one of my favorite bars in <em>NY</em>.</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>ID</th>
<th>Sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td><em>Phoenix</em> is my all-time favorite dive bar in <em>New York City</em>.</td>
</tr>
<tr>
<td>S3</td>
<td><em>Phoenix</em> has become one of my favorite bars in <em>NY</em>.</td>
</tr>
</tbody>
</table>
My Solution: **ClusType** (KDD’15)

<table>
<thead>
<tr>
<th>ID</th>
<th>Segmented Sentences</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td><em>Phoenix</em> is my all-time favorite dive bar in <em>New York City</em>.</td>
</tr>
<tr>
<td>S2</td>
<td>The best <em>BBQ</em> I’ve <em>tasted in</em> <em>Phoenix</em>.</td>
</tr>
<tr>
<td>S3</td>
<td><em>Phoenix</em> has become one of my favorite bars in <em>NY</em>.</td>
</tr>
</tbody>
</table>

Jointly optimize two sub-tasks on the graph:

1. Type label propagation
2. Relation phrase clustering

Correlated mentions:

- **S2**: Phoenix
- **S2**: BBQ
- **S3**: Phoenix
- **S1**: New York City
- **S3**: NY

**Represent object interactions**

- **tasted in**
- **is my all-time favorite dive bar in**
- **has become one of my favorite bars in**

**Synonymous relation phrases**

- **Phoenix**
- **New York City**
- **NY**
- **BBQ**
**ClusType**: Data-Driven Entity Mention Detection

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

\[
\rho_x(S_1, S_2) = \frac{\nu(S_1 \oplus S_2) - N \frac{\nu(S_1)}{N} \frac{\nu(S_2)}{N}}{\sqrt{\nu(S_1 \oplus S_2)}} \cdot I_X(S_1 \oplus S_2)
\]

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>(J*)N*</td>
<td>support vector machine</td>
</tr>
<tr>
<td>VP</td>
<td>tasted in, damage on</td>
</tr>
<tr>
<td>VW*(P)</td>
<td>train a classifier with</td>
</tr>
</tbody>
</table>

**Good Concordance**
ClusType: Data-Driven Entity Mention Detection

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

\[
\rho_x(S_1, S_2) = \frac{\nu(S_1 \oplus S_2) - N \frac{\nu(S_1)}{N} \frac{\nu(S_2)}{N}}{\sqrt{\nu(S_1 \oplus S_2)}} \cdot I_x(S_1 \oplus S_2)
\]

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>(J*)N*</td>
<td>support vector machine</td>
</tr>
<tr>
<td>VP</td>
<td>tasted in, damage on</td>
</tr>
<tr>
<td>VW*(P)</td>
<td>train a classifier with</td>
</tr>
</tbody>
</table>

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.

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

\[ O_{\alpha, \gamma, \mu} = F(C, P_L, P_R) + L_{\alpha}(P_L, P_R, \{U^{(v)}, V^{(v)}\}, U^*) + \Omega_{\gamma, \mu}(Y, C, P_L, P_R). \]

Type propagation between entity names and relation phrases

Mention correlation & mention type modeling

Multi-view relation phrases clustering
ClusType: Comparing with State-of-the-Art Systems

<table>
<thead>
<tr>
<th>Methods</th>
<th>NYT (118k 2013 news articles)</th>
<th>Yelp (230k business reviews)</th>
<th>Tweet (302k tweets)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pattern (Stanford, CONLL’14)</td>
<td>0.301</td>
<td>0.199</td>
<td>0.223</td>
</tr>
<tr>
<td>SemTagger (U Utah, ACL’10)</td>
<td>0.407</td>
<td>0.296</td>
<td>0.236</td>
</tr>
<tr>
<td>NNPLB (UW, EMNLP’12)</td>
<td>0.637</td>
<td>0.511</td>
<td>0.246</td>
</tr>
<tr>
<td>APOLO (THU, CIKM’12)</td>
<td>0.795</td>
<td>0.283</td>
<td>0.188</td>
</tr>
<tr>
<td>FIGER (UW, AAAI’12)</td>
<td>0.881</td>
<td>0.198</td>
<td>0.308</td>
</tr>
<tr>
<td>ClusType (KDD’15)</td>
<td><strong>0.939</strong></td>
<td><strong>0.808</strong></td>
<td><strong>0.451</strong></td>
</tr>
</tbody>
</table>

- **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 \( (P) = \frac{\text{#Correctly-typed mentions}}{\text{#System-recognized mentions}} \)

Recall \( (R) = \frac{\text{#Correctly-typed mentions}}{\text{#ground-truth mentions}} \)

F1 score \( = \frac{2(P \times R)}{(P + R)} \)
Fine-Grained Entity Typing

<table>
<thead>
<tr>
<th>ID</th>
<th>Sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td><em>Donald Trump</em> spent 14 television seasons presiding over a business-themed game show, NBC’s <em>The Apprentice</em></td>
</tr>
</tbody>
</table>

A few common types:
- Location
- Person
- Organization

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

A type hierarchy with 100+ types
Current Distant Supervision: Context-Agnostic Labeling

- “Context-agnostic” type assignment in training data
- Prior work: all labels are “perfect” training labels

<table>
<thead>
<tr>
<th>ID</th>
<th>Sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td><em>Donald Trump</em> spent 14 television seasons presiding over a business-themed game show, NBC’s The Apprentice</td>
</tr>
</tbody>
</table>

**S1: Donald Trump**

Entity Types: *person, artist, actor, author, businessman, politician*
Current Distant Supervision: “Type Independence” Assumption

• Entity types are not independent $\rightarrow$ correlated

• Existing studies ignore such correlation information

How to deal with infrequent (fine-grained) entity types?
My Solution: **Partial Label Embedding (KDD’16)**

Extract Text Features

Label Noise Reduction

Train Classifiers on De-noised Data

Prediction on New Data

<table>
<thead>
<tr>
<th>ID</th>
<th>Sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td><em>Donald Trump</em> spent 14 television seasons presiding over a business-themed game show, NBC’s <em>The Apprentice</em></td>
</tr>
</tbody>
</table>

**Text features**: HEAD: *Donald*, CXT_A: television, CXT_A: season, POS: NN, TKN: *trump*, SHAPE: AA

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

“Robust” classifier

PLE: Modeling Clean and Noisy Mentions Separately

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

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

<table>
<thead>
<tr>
<th>ID</th>
<th>Noisy Entity Mention</th>
<th>Types ranked</th>
<th>“Best” candidate type</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>Donald Trump spent 14 television seasons presiding over a business-themed game show, NBC’s The Apprentice</td>
<td>(+) actor 0.88&lt;br&gt;(+) artist 0.74&lt;br&gt;(+) person 0.55&lt;br&gt;(+) author 0.41&lt;br&gt;(+) politician 0.33&lt;br&gt;(+) business 0.31</td>
<td>(+) actor&lt;br&gt;(+) artist&lt;br&gt;(+) person&lt;br&gt;(+) author&lt;br&gt;(+) politician&lt;br&gt;(+) business</td>
</tr>
</tbody>
</table>

**S1: Donald Trump**

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

Measured based on currently estimated embedding space

Type Inference in PLE

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

<table>
<thead>
<tr>
<th>ID</th>
<th>Sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_i$</td>
<td>President <em>Trump gave</em> an all-hands <em>address</em> to troops at the U.S. Central Command headquarters</td>
</tr>
</tbody>
</table>

Test mention: $S_{i-}Trump$

Embedding vectors for text features
PLE: Performance of Fine-Grained Entity Typing

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

Accuracy on different type levels

- **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.
Error propagation cascading down the pipeline

Entity mention detection

Context-aware entity typing

Relation mention detection

Context-aware relation classification

Entity boundary errors:
The Women’s March was a worldwide protest on January 21, 2017.

Entity type errors:
The Women’s March was a worldwide protest on January 21, 2017. → person

Relation mention errors:
(women, protest) ✗
(protest, January 21, 2017)

Relation type errors
(women, protest) → is a ✗
(protest, January 21, 2017)
My Solution: **CoType** (WWW’17)

1. Data-driven detection of entity and relation mentions
   - Data-driven text segmentation
   - POS pattern learning from KBs

2. Joint typing of entity and relation mentions
   - **Noise-robust** type modeling
   - Object “translating” function to model **entity-relation interactions**

---

**Knowledge Bases**

- **Freebase**
- **Wikipedia**

---

CoType: Co-Embedding for Typing Entities and Relations

Modeling Automatically-Labeled Training Corpus

Joint Entity and Relation Embedding

Model entity-relation interactions
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 \( m_i \) and \( m_j \) have a large second-order proximity

- **S7:** \((Donald Trump, United States)\)
- **S6:** \((Trump, US)\)

- `TKN_Trump`
- `BETWEEN_president`
- `EM2_US`
- `EM2_United States`
Current Distant Supervision: Context-Agnostic Labeling

<table>
<thead>
<tr>
<th>ID</th>
<th>Sentence</th>
<th>Type labels for relation mention in S2:</th>
</tr>
</thead>
</table>
| S1 | *Donald Trump* was born in Queens, New York, *USA* on June 14, 1946. | **E1**: *Donald J. Trump*  
**E1 Types**: person, politician, businessman, author, actor  
**E2**: *United States*  
**E2 Types**: location, organization  
**Relations between E1, E2**: president of, citizen of, born in |
| 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*.” |
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.

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

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 $m_1$ and $m_2$, 
$\text{vec}(m_1) \approx \text{vec}(m_2) + \text{vec}(z)$

- Error on an entity-relation triple $(z, m_1, m_2)$: 
  $$\tau(z) = \|m_1 + z - m_2\|^2$$

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

For example:

- Embedding space $m_1$ = “USA”
- $m_2$ = “Washington D.C.”
- $z$ = capital_city_of “France” = “Paris”

Error on an entity-relation triple $(z, m_1, m_2)$:
$$\sum_{z_i \in Z_L} \sum_{v=1}^V \max \{0, 1 + \tau(z_i) - \tau(z_v)\}$$

Positive relation triple
Negative relation triple
Reducing Error Propagation: A Joint Optimization Framework

\[ O_{ZM} = \sum_{z_i \in Z_L} \sum_{v=1}^{V} \max \left\{ 0, 1 + \tau(z_i) - \tau(z_v) \right\} \]

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

\[ O_Z = \mathcal{L}_{ZF} + N_L \sum_{i=1} \ell_i + \frac{\lambda}{2} \sum_{i=1}^{N_L} \| z_i \|_2^2 + \frac{\lambda}{2} \sum_{k=1}^{K_r} \| r_k \|_2^2 \]

\[ O_M = \mathcal{L}_{MF} + \lambda \sum_{i=1}^{N_L'} \ell_i' + \frac{\lambda}{2} \sum_{i=1}^{N_L'} \| m_i \|_2^2 + \frac{\lambda}{2} \sum_{k=1}^{K_y} \| y_k \|_2^2 \]

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/](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

<table>
<thead>
<tr>
<th>LifeNet:</th>
<th>LifeNet by Effort-Less StructMine</th>
</tr>
</thead>
<tbody>
<tr>
<td>A Knowledge Exploration and Analytics System for Life Sciences</td>
<td>Machine-created</td>
</tr>
<tr>
<td>BioInfer Corpus</td>
<td>4 Million+ papers</td>
</tr>
<tr>
<td>Human-created</td>
<td>1,000+ entity types, 400+ relations</td>
</tr>
<tr>
<td>1,100 sentences</td>
<td>&lt;1 hour, single machine</td>
</tr>
<tr>
<td>94 Protein-Protein interactions</td>
<td>10,000x more extractions</td>
</tr>
<tr>
<td>2,500 man-hours</td>
<td></td>
</tr>
<tr>
<td>2,662 facts</td>
<td></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Work done at UIUC</th>
<th>My Research Publications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phrase mining</td>
<td>SDM’14, SIGMOD’15</td>
</tr>
<tr>
<td>Entity recognition and typing</td>
<td>KDD’15, KDD’16, EMNLP’16</td>
</tr>
<tr>
<td>Relation Extraction</td>
<td>WWW’17</td>
</tr>
<tr>
<td>Entity synonym mining</td>
<td>WWW’15</td>
</tr>
<tr>
<td>Facet discovery</td>
<td>CIKM’16</td>
</tr>
<tr>
<td>Automatic summarization</td>
<td>ICDM’13, WWW’16, WSDM’17</td>
</tr>
<tr>
<td>Entity recommendation in text-rich environment</td>
<td>RecSys’13, WSDM’14a, WSDM’14b</td>
</tr>
</tbody>
</table>

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.


- 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.


• **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.


• 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

Corpus to Structured Network: The Roadmap

**ClusType**: Entity Recognition and Typing (KDD’15)

**Fine-grained Entity Typing**: (KDD’16, EMNLP’16)

**CoType**: Joint Entity and Relation Extraction (WWW’17)
My Work: Mining Factual Structures with Minimal Human Involvement

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

Training examples

Model

Auto-annotation

WIKIPEDIA
Die freie Enzyklopädie

About: Egypt
An Entity of Type: place, from Named Graph: http://dbpedia.org, within DBpedia

News Articles

Model Training examples

News Articles

DBpedia

Model

Auto-annotation

WIKIPEDIA
Die freie Enzyklopädie

About: Egypt
An Entity of Type: place, from Named Graph: http://dbpedia.org, within DBpedia

News Articles

Model

Auto-annotation

WIKIPEDIA
Die freie Enzyklopädie

About: Egypt
An Entity of Type: place, from Named Graph: http://dbpedia.org, within DBpedia

News Articles

Model

Auto-annotation

WIKIPEDIA
Die freie Enzyklopädie

About: Egypt
An Entity of Type: place, from Named Graph: http://dbpedia.org, within DBpedia

News Articles

Model

Auto-annotation

WIKIPEDIA
Die freie Enzyklopädie

About: Egypt
An Entity of Type: place, from Named Graph: http://dbpedia.org, within DBpedia

News Articles

Model

Auto-annotation

WIKIPEDIA
Die freie Enzyklopädie

About: Egypt
An Entity of Type: place, from Named Graph: http://dbpedia.org, within DBpedia

News Articles

Model

Auto-annotation

WIKIPEDIA
Die freie Enzyklopädie

About: Egypt
An Entity of Type: place, from Named Graph: http://dbpedia.org, within DBpedia

News Articles

Model

Auto-annotation

WIKIPEDIA
Die freie Enzyklopädie

About: Egypt
An Entity of Type: place, from Named Graph: http://dbpedia.org, within DBpedia

News Articles

Model

Auto-annotation

WIKIPEDIA
Die freie Enzyklopädie

About: Egypt
An Entity of Type: place, from Named Graph: http://dbpedia.org, within DBpedia

News Articles

Model

Auto-annotation

WIKIPEDIA
Die freie Enzyklopädie

About: Egypt
An Entity of Type: place, from Named Graph: http://dbpedia.org, within DBpedia

News Articles

Model

Auto-annotation

WIKIPEDIA
Die freie Enzyklopädie

About: Egypt
An Entity of Type: place, from Named Graph: http://dbpedia.org, within DBpedia

News Articles

Model

Auto-annotation

WIKIPEDIA
Die freie Enzyklopädie

About: Egypt
An Entity of Type: place, from Named Graph: http://dbpedia.org, within DBpedia

News Articles

Model

Auto-annotation

WIKIPEDIA
Die freie Enzyklopädie

About: Egypt
An Entity of Type: place, from Named Graph: http://dbpedia.org, within DBpedia

News Articles

Model

Auto-annotation

WIKIPEDIA
Die freie Enzyklopädie

About: Egypt
An Entity of Type: place, from Named Graph: http://dbpedia.org, within DBpedia

News Articles

Model

Auto-annotation

WIKIPEDIA
Die freie Enzyklopädie

About: Egypt
An Entity of Type: place, from Named Graph: http://dbpedia.org, within DBpedia

News Articles

Model

Auto-annotation

WIKIPEDIA
Die freie Enzyklopädie

About: Egypt
An Entity of Type: place, from Named Graph: http://dbpedia.org, within DBpedia

News Articles

Model

Auto-annotation

WIKIPEDIA
Die freie Enzyklopädie

About: Egypt
An Entity of Type: place, from Named Graph: http://dbpedia.org, within DBpedia

News Articles

Model

Auto-annotation

WIKIPEDIA
Die freie Enzyklopädie

About: Egypt
An Entity of Type: place, from Named Graph: http://dbpedia.org, within DBpedia

News Articles

Model

Auto-annotation

WIKIPEDIA
Die freie Enzyklopädie

About: Egypt
An Entity of Type: place, from Named Graph: http://dbpedia.org, within DBpedia

News Articles

Model

Auto-annotation

WIKIPEDIA
Die freie Enzyklopädie

About: Egypt
An Entity of Type: place, from Named Graph: http://dbpedia.org, within DBpedia

News Articles

Model

Auto-annotation

WIKIPEDIA
Die freie Enzyklopädie

About: Egypt
An Entity of Type: place, from Named Graph: http://dbpedia.org, within DBpedia

News Articles

Model

Auto-annotation

WIKIPEDIA
Die freie Enzyklopädie

About: Egypt
An Entity of Type: place, from Named Graph: http://dbpedia.org, within DBpedia

News Articles

Model

Auto-annotation

WIKIPEDIA
Die freie Enzyklopädie

About: Egypt
An Entity of Type: place, from Named Graph: http://dbpedia.org, within DBpedia

News Articles

Model

Auto-annotation

WIKIPEDIA
Die freie Enzyklopädie

About: Egypt
An Entity of Type: place, from Named Graph: http://dbpedia.org, within DBpedia

News Articles

Model

Auto-annotation

WIKIPEDIA
Die freie Enzyklopädie

About: Egypt
An Entity of Type: place, from Named Graph: http://dbpedia.org, within DBpedia

News Articles

Model

Auto-annotation

WIKIPEDIA
Die freie Enzyklopädie

About: Egypt
An Entity of Type: place, from Named Graph: http://dbpedia.org, within DBpedia

News Articles

Model

Auto-annotation
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

<table>
<thead>
<tr>
<th>ID</th>
<th>Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>... has concerns whether <strong>Kabul</strong> is an ally of <strong>Washington</strong>.</td>
</tr>
<tr>
<td>S2</td>
<td>... <strong>Australia</strong> becomes a close ally of the <strong>United States</strong>.</td>
</tr>
<tr>
<td>S3</td>
<td>The <strong>Cardinal</strong> will share the title with <strong>California</strong> if the <strong>Golden Bears</strong> beat <strong>Washington</strong> later Saturday.</td>
</tr>
</tbody>
</table>

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)

<table>
<thead>
<tr>
<th>ID</th>
<th>Sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td><em>Phoenix</em> is my all-time favorite dive bar in <em>New York City</em>.</td>
</tr>
<tr>
<td>S2</td>
<td>The best <em>BBQ</em> I’ve tasted in <em>Phoenix</em>.</td>
</tr>
<tr>
<td>S3</td>
<td><em>Phoenix</em> has become one of my favorite bars in <em>NY</em>.</td>
</tr>
</tbody>
</table>
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

ClusType: Comparing with Sequence Model

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

<table>
<thead>
<tr>
<th>Methods</th>
<th>NYT (118k 2013 news articles)</th>
<th>Yelp (230k restaurant reviews)</th>
<th>Tweet (302k tweets)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stanford NER (2014 version)</td>
<td>0.682</td>
<td>0.240</td>
<td>0.438</td>
</tr>
<tr>
<td>ClusType (KDD’15)</td>
<td><strong>0.942</strong></td>
<td><strong>0.594</strong></td>
<td><strong>0.472</strong></td>
</tr>
</tbody>
</table>

• 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)
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

AFET (EMNLP’16): Framework Overview

<table>
<thead>
<tr>
<th>ID</th>
<th>Sentence</th>
<th>Text features</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td><em>Donald Trump</em> spent 14 television seasons presiding over a business-themed game show, NBC’s <em>The Apprentice</em></td>
<td>HEAD_Donald, CXT_A: television, CXT_A: season, POS: NN, TKN_trump, SHAPE: AA</td>
</tr>
</tbody>
</table>

**S1: Donald Trump**

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

**S4: Ted Cruz**

**Entity Types:** person, politician
### Performance Comparison on Fine-Grained Typing

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

- AFET vs. AFET-NoCo $\rightarrow$ gain from incorporating type correlation
- AFET vs. AFET-NoPa $\rightarrow$ gain from noise-robust loss function
**AFET**: Performance of Fine-Grained Entity Typing

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

<table>
<thead>
<tr>
<th>Methods</th>
<th>Wikipedia</th>
<th>OntoNotes</th>
<th>BBN</th>
</tr>
</thead>
<tbody>
<tr>
<td>FIGER (UW, AAAI’12)</td>
<td>0.474</td>
<td>0.369</td>
<td>0.467</td>
</tr>
<tr>
<td>HYENA (Max-Planck, COLING’12)</td>
<td>0.288</td>
<td>0.249</td>
<td>0.523</td>
</tr>
<tr>
<td>WASABIE (Google, ACL’14)</td>
<td>0.480</td>
<td>0.404</td>
<td>0.619</td>
</tr>
<tr>
<td>HNM (IBM, EMNLP’15)</td>
<td>0.237</td>
<td>0.122</td>
<td>0.551</td>
</tr>
<tr>
<td>PTE (MSR, WWW’15)</td>
<td>0.405</td>
<td>0.436</td>
<td>0.604</td>
</tr>
<tr>
<td>PL-SVM (Cornell, KDD’08)</td>
<td>0.428</td>
<td>0.225</td>
<td>0.465</td>
</tr>
<tr>
<td><strong>AFET (EMNLP’16)</strong></td>
<td><strong>0.533</strong></td>
<td><strong>0.551</strong></td>
<td><strong>0.670</strong></td>
</tr>
</tbody>
</table>

- 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)

Substantial human annotation

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

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

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

No human annotation

CoType Step 1: Data-Driven Entity and Relation Detection

**S2:** The protest was aimed at Donald Trump, the recently inaugurated president of the United States.

**Frequent Pattern Mining**

**S2:** The protest was aimed at Donald Trump, the recently inaugurated president of the United States.

**Segment Quality Estimation**

Phrases quality: *United States*: 0.9, *was aimed at*: 0.4, ...
Part-of-speech (POS) patterns quality: *ADJ NN*: 0.85, *V PROP*: 0.4, ...

**POS-guided Segmentation**

**S2:** The protest was aimed at Donald Trump, the recently inaugurated president of the United States.

**Quality Re-estimation & Re-segmentation**

(S2: *protest, Donald Trump*), (S2: *Donald Trump, United States*)
# Entity Mention Detection: Results

## POS Tag Pattern

<table>
<thead>
<tr>
<th>Good (high score)</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>NNP NNP</em></td>
<td>San Francisco/Barack Obama/United States comedy drama/car accident/club captain</td>
</tr>
<tr>
<td><em>NN NN</em></td>
<td>seven network/seven dwarfs/2001 census crude oil/nucletic acid/baptist church</td>
</tr>
<tr>
<td><em>CD NN</em></td>
<td></td>
</tr>
<tr>
<td><em>JJ NN</em></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Bad (low score)</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>DT JJ NND</em></td>
<td>a few miles/the early stages/the late 1980s</td>
</tr>
<tr>
<td><em>CD CD NN IN</em></td>
<td>2 : 0 victory over/1 : 0 win over rating on rotten tomatoes worked together on/spent much of</td>
</tr>
<tr>
<td><em>NN IN NNP NNP</em></td>
<td></td>
</tr>
<tr>
<td><em>VVD RB IN</em></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>FIGER segmenter [UW, 2012]</strong></th>
<th>NYT</th>
<th>Wiki-KBP</th>
<th>BioInfer</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Our Approach</em></td>
<td><em>0.837</em></td>
<td>0.833</td>
<td>0.785</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>NYT</th>
<th>Wiki-KBP</th>
<th>BioInfer</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.751</td>
<td>0.814</td>
<td>0.652</td>
</tr>
</tbody>
</table>
Key Insight: Text Co-occurrence Patterns Bring Semantic Power

Training data

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

<table>
<thead>
<tr>
<th>Entity 1</th>
<th>Entity 2</th>
<th>Co-occurrences</th>
</tr>
</thead>
<tbody>
<tr>
<td>speech</td>
<td>Barack Obama</td>
<td>57</td>
</tr>
<tr>
<td>United States</td>
<td>Barack Obama</td>
<td>217</td>
</tr>
<tr>
<td>president</td>
<td>Barack Obama</td>
<td>168</td>
</tr>
</tbody>
</table>

President Donald Trump delivers a speech during ...

President Donald Trump delivers a speech during ...

politician
CoType: Performance of Entity Recognition and Typing

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

<table>
<thead>
<tr>
<th>Methods</th>
<th>NYT</th>
<th>Wiki-KBP</th>
<th>BioInfer</th>
</tr>
</thead>
<tbody>
<tr>
<td>FIGER (UW, AAAI’12)</td>
<td>0.40</td>
<td>0.29</td>
<td>0.69</td>
</tr>
<tr>
<td>HYENA (Max-Planck, COLING’12)</td>
<td>0.44</td>
<td>0.26</td>
<td>0.52</td>
</tr>
<tr>
<td>WASABIE (Google, ACL’14)</td>
<td>0.53</td>
<td>0.35</td>
<td>0.64</td>
</tr>
<tr>
<td>DeepWalk (StonyBrook, KDD’14)</td>
<td>0.49</td>
<td>0.21</td>
<td>0.58</td>
</tr>
<tr>
<td>PLE (KDD’16)</td>
<td>0.56</td>
<td>0.37</td>
<td>0.70</td>
</tr>
<tr>
<td><strong>CoType (WWW’17)</strong></td>
<td><strong>0.60</strong></td>
<td><strong>0.39</strong></td>
<td><strong>0.74</strong></td>
</tr>
</tbody>
</table>

- 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

<table>
<thead>
<tr>
<th>Method</th>
<th>News</th>
<th>BiolInfer</th>
</tr>
</thead>
<tbody>
<tr>
<td>IBM FCM</td>
<td>0.681</td>
<td>0.467</td>
</tr>
<tr>
<td>UW MultiR</td>
<td>0.881</td>
<td>0.501</td>
</tr>
<tr>
<td>CoType (WWW’17)</td>
<td>0.939</td>
<td>0.617</td>
</tr>
</tbody>
</table>

relation classification accuracy
# Applications of StructNets

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