Constructing Structured Information Networks from Massive Text Corpora

Part II: Joint Extraction of Typed Entity and Relation
Effort-Light StructMine: Methodology

Text corpus

Data-driven text segmentation
(SIGMOD'15, WWW'16)

Entity names & context units

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

Partially-labeled corpus

Knowledge bases

Structures from the remaining unlabeled data

Closed-world Assumption vs Open-world Assumption
Effort-Light StructMine: Typing

Text corpus

Data-driven text segmentation (SIGMOD’15, WWW’16)

Entity names & context units

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

Partially-labeled corpus

Knowledge bases

Structures from the remaining unlabeled data

Entity Recognition and Coarse-grained Typing (KDD’15)

Fine-grained Entity Typing (KDD’16)

Joint Entity and Relation Extraction (WWW’17)

Corpus to Structured Network: The Roadmap
Corpus to Structured Network: The Roadmap

Text corpus

Data-driven text segmentation (SIGMOD’15, WWW’16)

Entity Recognition and Coarse-grained Typing (KDD’15)

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

Fine-grained Entity Typing (KDD’16)

Joint Entity and Relation Extraction (WWW’17)

Partially-labeled corpus

Structures from the remaining unlabeled data

Knowledge bases

entity names & context units
Recognizing Entities of Target Types in 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 corpus-specific human labeling
- Training sequence models is slow

A manual annotation interface

The best BBQ I’ve tasted in Phoenix

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

e.g., (McMallum & Li, 2003), (Finkel et al., 2005), (Ratinov & Roth, 2009), ...
Weak-Supervision Systems: Pattern-Based Bootstrapping

- Requires manual seed selection & mid-point checking

Seed entities and corpus → Annotate corpus using entities

Apply patterns to find new entities

Select Top patterns

Generate candidate patterns

Score candidate patterns

Patterns for **Food**

the best <X> I’ve tried in their <X> tastes amazing

Seed entities and corpus

**Seeds for Food**

- Pizza
- French Fries
- Hot Dog
- Pancake
- ...

**Systems:**

- CMU NELL
- UW KnowItAll
- Stanford DeepDive
- Max-Planck PROSPERA
- ...

e.g., (Etzioni et al., 2005), (Talukdar et al., 2010), (Gupta et al., 2014), (Mitchell et al., 2015), ...
Leveraging Distant Supervision

1. **Detect** entity names from text
2. **Match** name strings to KB entities
3. **Propagate** types to the un-matchable names

<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>

(Lin et al., 2012), (Ling et al., 2012), (Nakashole et al., 2013)
Current Distant Supervision: Limitation I

1. Context-agnostic type prediction
   - Predict types for each mention regardless of context

2. Sparsity of contextual bridges

<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>
Current Distant Supervision: Limitation II

1. Context-agnostic type prediction

2. Sparsity of contextual bridges
   • Some relational phrases are infrequent in the corpus ➔ ineffective type propagation

<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>
ClusType: Data-Driven Entity Mention Detection

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

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

<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

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

<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**.
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 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>

Putting two sub-tasks together:

1. Type label propagation
2. Relation phrase clustering

![Diagram](image_url)
Type Propagation in ClusType

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

Vector of scores for single label on nodes

Edge weight / object similarity

Measure of Non-Smoothness

(Belkin & Partha, NIPS’01), (Ren et al., KDD’15)
Relation Phrase Clustering in ClusType

• Two relation phrases should be grouped together if:
  1. Similar string
  2. Similar context
  3. Similar types for entity arguments

“Multi-view” clustering

(Ren et al., KDD’15)
**ClusType: Comparing with State-of-the-Art Systems (F1 Score)**

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

- **vs. bootstrapping**: context-aware prediction on “un-matchable”
- **vs. label propagation**: group similar relation phrases
- **vs. FIGER**: no reliance on complex feature engineering

**NYT**: 118k news articles (1k manually labeled for evaluation); **Yelp**: 230k business reviews (2.5k reviews are manually labeled for evaluation); **Tweet**: 302 tweets (3k tweets are manually labeled for evaluation)

Precision \( (P) = \frac{\text{Correctly-typed mentions}}{\text{System-recognized mentions}} \), Recall \( (R) = \frac{\text{Correctly-typed mentions}}{\text{ground-truth mentions}} \), F1 score \( = \frac{2(P \times R)}{(P + R)} \)
Corpus to Structured Network: The Roadmap

Entity Recognition and Coarse-grained Typing (KDD’15) → Fine-grained Entity Typing (KDD’16) → Joint Entity and Relation Extraction (WWW’17)

Data-driven text segmentation (SIGMOD’15, WWW’16)

Entity names & context units

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

Partially-labeled corpus

Knowledge bases

Text corpus

Structures from the remaining unlabeled data
From Coarse-Grained Typing to Fine-Grained Entity Typing

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

A few common types: Location, Person, Organization

A type hierarchy with 100+ types (from knowledge base)

(Ling et al., 2012), (Nakashole et al., 2013), (Yogatama et al., 2015)
Current Distant Supervision: Context-Agnostic Labeling

- Inaccurate labels in training data
- Prior work: all labels are “perfect”

<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 game show, NBC’s <em>The Apprentice</em></td>
</tr>
</tbody>
</table>

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

Entity: *Donald Trump*
My Solution: **Partial Label Embedding (KDD’16)**

- **Extract Text Features**
- **“Label Noise Reduction” with PLE**
- **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 game show, NBC’s The Apprentice</td>
</tr>
</tbody>
</table>

**Text features**: TOKEN_{Donald}, CONTEXT: television, CONTEXT: season, TOKEN_{trump}, SHAPE: AA

**S1: Donald Trump**

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

“De-noised” labeled data

More effective classifiers

(Ren et al., KDD’16)
**PLE: Modeling Clean and Noisy Mentions Separately**

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

<table>
<thead>
<tr>
<th>ID</th>
<th>Noisy Entity Mention</th>
<th>Types ranked</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td><em>Donald Trump</em> spent 14 television seasons presiding over a game show, NBC’s The Apprentice</td>
<td>(+) actor 0.88, (+) artist 0.74, (+) person 0.55, (+) author 0.41, (+) politician 0.33, (+) business 0.31</td>
</tr>
</tbody>
</table>

*Best* candidate type

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

---

(Ren et al., KDD’16)
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</em> gave an all-hands <em>address</em> to troops at the U.S. Central Command headquarters</td>
</tr>
</tbody>
</table>

Test mention: $S_i \downarrow$ *Trump*

(Ren et al., KDD’16)
**PLE: Performance of Fine-Grained Entity Typing**

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

<table>
<thead>
<tr>
<th>Level</th>
<th>Raw</th>
<th>WSABIE</th>
<th>PTE</th>
<th>PLE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level-1</td>
<td>0.79</td>
<td>0.81</td>
<td>0.45</td>
<td>0.14</td>
</tr>
<tr>
<td>Level-2</td>
<td>0.49</td>
<td>0.51</td>
<td>0.19</td>
<td>0.19</td>
</tr>
<tr>
<td>Level-3</td>
<td>0.05</td>
<td>0.62</td>
<td>0.48</td>
<td>0.48</td>
</tr>
</tbody>
</table>

- **Raw**: candidate types from distant supervision
- **WSABIE** (Google, ACL’15): joint feature and type embedding
- **Predictive Text Embedding** (MSR, WWW’15): joint mention, feature and type embedding
  - Both WASBIE and PTE suffer from “noisy” training labels
- **PLE** (KDD’16): partial-label loss for context-aware labeling

**OntoNotes** public dataset (Weischedel et al. 2011, Gillick et al., 2014): 13,109 news articles, 77 annotated documents, 89 entity types
Corpus to Structured Network: The Roadmap

- Text corpus
- Data-driven text segmentation (SIGMOD’15, WWW’16)
- Entity names & context units
- Partially-labeled corpus
- Learning Corpus-specific Model (KDD’15, KDD’16, EMNLP’16, WWW’17)
- Structures from the remaining unlabeled data
- Knowledge bases

- Entity Recognition and Coarse-grained Typing (KDD’15)
- Fine-grained Entity Typing (KDD’16)
- Joint Entity and Relation Extraction (WWW’17)
Joint 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: 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”

No human annotation

Distantly-supervised RE systems (cont.)

- Error propagation
- Noisy candidate type labels

Prior Work: An “Incremental” System Pipeline

Error propagation cascading down the pipeline

- **Entity mention detection**
- **Context-aware entity typing**
- **Relation mention detection**
- **Context-aware relation typing**

**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*.

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

**Relation type errors:**
- *(women, protest)* → *is a* ***✗***
- *(protest, January 21, 2017)*

(Mintz et al., 2009), (Riedel et al., 2010), (Hoffmann et al., 2011), (Surdeanu et al., 2012), (Nagesh et al., 2014), ...
My Solution: **CoType (WWW’17)**

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

2. **Joint typing of entity and relation mentions**
   - *Context-aware* type modeling
   - Model entity-relation interactions

*Ren et al. WWW’17*
My Solution: **CoType** (WWW’17)

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

2. Joint typing of entity and relation mentions
   - **Context-aware** type modeling
   - Model **entity-relation interactions**

(Ren et al. WWW’17)
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

<table>
<thead>
<tr>
<th>POS Tag Pattern</th>
<th>Example</th>
<th>NYT</th>
<th>Wiki-KBP</th>
<th>BioInfer</th>
</tr>
</thead>
</table>
| **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 | 0.751 | 0.814 | 0.652 |
| **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 | | | |

FIGER segmenter [UW, 2012]  
Our Approach | 0.751 | 0.814 | 0.652 | 0.837 | 0.833 | 0.785 |
CoType: Co-Embedding for Typing Entities and Relations

Object interactions in a heterogeneous graph

Low-dimensional vector spaces

Model entity-relation interactions

(Ren et al. WWW’17)
Modeling Mention-Feature Co-occurrences

- **Second-order Proximity**
  Mentions with similar distributions over text features should have similar types.

Vertex $m_i$ and $m_j$ have a large second-order proximity.

(Tang et al., WWW’15), (Ren et al. WWW’17)
**Challenge: Context-Agnostic Labeling**

<table>
<thead>
<tr>
<th>ID</th>
<th>Sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>S2</td>
<td>The protest was aimed at Donald Trump, the recently inaugurated president of the United States.</td>
</tr>
</tbody>
</table>

**Type labels for relation mention:**

<table>
<thead>
<tr>
<th>E1: Donald J. Trump</th>
<th>E2: United States</th>
</tr>
</thead>
<tbody>
<tr>
<td>E1 Types: person, politician, businessman, author, actor</td>
<td>E2 Types: location, organization</td>
</tr>
</tbody>
</table>

**Relations between E1, E2 in KB:**

- president of
- live in
- born in

Relation types from knowledge base:

- president of
- born in
- live in

Entity 1: Donald Trump

Entity 2: United States
Context-Aware Type Modeling

Partial-label Loss Function

- Vector representation of the relation mention should be more similar to its “best” candidate type, than to any other non-candidate type.

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

Maximal score for non-candidate types

(Ren et al. WWW’17)
Modeling Entity-Relation Interactions

Object “Translating” Assumption
For a relation mention $z$ between entity arguments $m_1$ and $m_2$:

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

Error on a relation triple $(z, m_1, m_2)$:

$$\tau(z) = \|m_1 + z - m_2\|^2_2$$

(Bordes, NIPS’13), (Ren et al., WWW’17)
Reducing Error Propagation: A Joint Optimization Framework

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

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

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

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

Modeling entity-relation interactions

Modeling types of entity mentions

Modeling types of relation mentions

(Ren et al., WWW’17)
**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”

---

<table>
<thead>
<tr>
<th>Method</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>DS+Logistic</td>
<td>Logistic classifier on DS</td>
</tr>
<tr>
<td>MultiR</td>
<td>Handles inappropriate labels in DS</td>
</tr>
<tr>
<td>DeepWalk</td>
<td>StonyBrook, KDD’14: homogeneous graph embedding</td>
</tr>
<tr>
<td>LINE</td>
<td>MSR, WWW’15: joint feature &amp; type embedding</td>
</tr>
<tr>
<td>CoType-RM (WWW’17)</td>
<td>Only models relation mentions</td>
</tr>
<tr>
<td>CoType (WWW’17)</td>
<td>Models entity-relation interactions</td>
</tr>
</tbody>
</table>

**NYT public dataset** (Riedel et al. 2010, Hoffmann et al., 2011): 1.18M sentences in the corpus, 395 manually annotated sentences for evaluation, 24 relation types
An Ongoing Application to Life Sciences

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

**LifeNet:**
A Knowledge Exploration and Analytics System for Life Sciences

**LifeNet by Effort-Light StructMine**
- Machine-created
- 4 Million+ PubMed papers
- 1,000+ entity types
- 400+ relation types
- <1 hour, single machine
- 10,000x more facts

**BioInfer Network by human labeling**
(Pyysalo et al., 2007)
- Human-created
- 1,100 sentences
- 94 protein-protein interactions
- 2,500 man-hours
- 2,662 facts

(Ren et al., ACL’17 demo, under review)
Corpus to Structured Network: The Roadmap

- Text corpus
- Data-driven text segmentation (SIGMOD’15, WWW’16)
- Learning Corpus-specific Model (KDD’15, KDD’16, EMNLP’16, WWW’17)
- Partially-labeled corpus
- Knowledge bases

- Entity Recognition and Coarse-grained Typing (KDD’15)
- Fine-grained Entity Typing (KDD’16)
- Joint Entity and Relation Extraction (WWW’17)

- Structures from the remaining unlabeled data
References 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.
References 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.