Building Structured Databases of Factual Knowledge 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

- **Structures from the remaining unlabeled data**
  - Corpus-specific Structure Discovery (KDD’15, KDD’16, EMNLP’16, WWW’17)
  - Partially-labeled corpus

- **Typing of Entities and Relations**

- **Meta Pattern-Based Attribute Discovery**
Effort-Light StructMine: Typing

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

Entity names & context units

Corpus-specific Structure Discovery (KDD’15, KDD’16, EMNLP’16, WWW’17)

Partially-labeled corpus

Knowledge bases

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)
Corpus to Structured Network: The Roadmap

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

Corpus-specific Structure Discovery (KDD’15, KDD’16, EMNLP’16, WWW’17)

Partially-labeled corpus

Knowledge bases

Text corpus

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)
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

...
Weak-Supervision Systems: Pattern-Based Bootstrapping

- Requires manual seed selection & mid-point checking

Seed entities and corpus → Annotate corpus using entities → Select Top patterns → Generate candidate patterns → Score candidate patterns → Apply patterns to find new entities

Patterns for **Food**
the best <X> I’ve tried in their <X> tastes amazing

**Seed entities and corpus**
- 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

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

<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
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 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
Type Propagation in ClusType

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

(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>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>APOLLO (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><strong>ClusType (KDD’15)</strong></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

Text corpus

Structures from the remaining unlabeled data

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 names & context units

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

Fine-grained Entity Typing (KDD’16)

Joint Entity and Relation Extraction (WWW’17)
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><em>Donald Trump</em> 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>Donald Trump spent 14 television seasons presiding over a game show, NBC’s The Apprentice</td>
</tr>
</tbody>
</table>

**S1: Donald Trump**

**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 <em>The Apprentice</em></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

---

More effective classifiers

“De-noised” labeled data

(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 Trump gave an all-hands address to troops at the U.S. Central Command headquarters</td>
</tr>
</tbody>
</table>

Test mention: $S_i \text{Trump}$

Vectors for text features

Low-dimensional vector space

Type hierarchy (from knowledge base)

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

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

**Accuracy on different type levels**

- **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

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)
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

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*. ➔ *person*

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

Relation type errors
* (women, protest) ➔ * is a X

(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)
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)
The protest was aimed at Donald Trump, the recently inaugurated president of the United States.

Type labels for relation mention:

- **E1**: Donald J. Trump
  - Types: person, politician, businessman, author, actor

- **E2**: United States
  - Types: location, organization

Relations between E1, E2 in KB:
- president of
- live in
- born in
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 $(z, m_1, m_2)$:

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

Low-dimensional vector space

“France” $m_1 =$ “USA” (country)

z = capital_city_of

“Paris” $m_2 =$ “Washington D.C.” (city)

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

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

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

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

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

(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”

- **DeepWalk** (StonyBrook, KDD’14): homogeneous graph embedding
- **LINE** (MSR, WWW’15): joint feature & type embedding
- **DS+Logistic** (Stanford, ACL’09): logistic classifier on DS
- **MultiR** (UW, ACL’11): handles inappropriate labels in DS
- **CoType-RM** (WWW’17): only models relation mentions
- **CoType** (WWW’17): models entity-relation interactions

**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

**BioInfer Network by human labeling**
(Pyysalo et al., 2007)

<table>
<thead>
<tr>
<th>Feature</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human-created</td>
<td></td>
</tr>
<tr>
<td>1,100 sentences</td>
<td></td>
</tr>
<tr>
<td>94 protein-protein interactions</td>
<td></td>
</tr>
<tr>
<td>2,500 man-hours</td>
<td></td>
</tr>
<tr>
<td>2,662 facts</td>
<td></td>
</tr>
</tbody>
</table>

**BioInfer Network by human labeling**
(Pyysalo et al., 2007)

<table>
<thead>
<tr>
<th>Feature</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human-created</td>
<td></td>
</tr>
<tr>
<td>1,100 sentences</td>
<td></td>
</tr>
<tr>
<td>94 protein-protein interactions</td>
<td></td>
</tr>
<tr>
<td>2,500 man-hours</td>
<td></td>
</tr>
<tr>
<td>2,662 facts</td>
<td></td>
</tr>
</tbody>
</table>

**LifeNet by Effort-Light StructMine**

<table>
<thead>
<tr>
<th>Feature</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Machine-created</td>
<td></td>
</tr>
<tr>
<td>4 Million+ PubMed papers</td>
<td></td>
</tr>
<tr>
<td>1,000+ entity types</td>
<td></td>
</tr>
<tr>
<td>400+ relation types</td>
<td></td>
</tr>
<tr>
<td>&lt;1 hour, single machine</td>
<td></td>
</tr>
<tr>
<td>10,000x more facts</td>
<td></td>
</tr>
</tbody>
</table>

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

(Pyysalo et al., BMC Bioinformatics’07)
(Ren et al., ACL’17 demo, under review)
Corpus to Structured Network: The Roadmap

- **Text corpus**
- **Data-driven text segmentation** (SIGMOD’15, WWW’16)
  - Entity names & context units
- **Corpus-specific Structure Discovery** (KDD’15, KDD’16, EMNLP’16, WWW’17)
  - Partially-labeled corpus

**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)

**Knowledge bases**
- DBpedia
- Freebase
- Wikipedia
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**, 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.