CoType: Joint Extraction of Typed Entities and Relations with Knowledge Bases

Xiang Ren¹, Zeqiu Wu¹, Wenqi He¹, Meng Qu¹, Clare R. Voss², Heng Ji³, Tarek F. Abdelzaher¹, Jiawei Han¹

¹UIUC CS, ²ARL CIS, ³RPI CS
Turning Unstructured Text Data into Structures

Unstructured Text Data
(account for ~80% of all data in organizations)

Knowledge & Insights

(Chakraborty, 2016)
This **hotel** is my favorite **Hilton** **property** in **NYC**! It is located right on 42nd street near **Times Square**, it is close to all subways, **Broadways shows**, and next to great restaurants like **Junior’s Cheesecake**, **Virgil’s BBQ** and many others.

--- **TripAdvisor**

**Structured Facts**

1. “Typed” entities
2. “Typed” relationships
Why Text to Structures?

Structured Search & Exploration

- SQL
- SPARQL

Graph Mining & Network Analysis

Pattern / Association Rule Mining

Structured Feature Generation
A Product Use Case: Finding “Interesting Hotel Collections”

Technology Transfer to TripAdvisor

Features for “Catch a Show” collection
1. broadway shows
2. beacon theater
3. broadway dance center
4. broadway plays
5. david letterman show
6. radio city music hall
7. theatre shows

Features for “Near The High Line” collection
1. high line park
2. chelsea market
3. highline walkway
4. elevated park
5. meatpacking district
6. west side
7. old railway

Grouping hotels based on structured facts extracted from the review text

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 Social Sciences

Business Intelligence
Why “Joint Extraction”? 

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
(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**

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**

---

(Ren et al. WWW’17)
Quality Measure for Text Segmentation

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

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

<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**

Markov Blanket Feature Selection for Support Vector Machines.
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)
My Solution: CoType

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

(Ren et al. WWW’17)
CoType: Co-Embedding for Typing Entities and Relations

Object Interactions in a Heterogeneous Graph

Entity Mention
- S1: "Barack Obama"
- S3: "Barack Obama"
- "United States"
- S2: "Dream of My Father"

Relation Mention
- BETWEEN_book
- EM1_Obama
- president_of
- bom_in
- author_of
- travel_to
- CONTEXT_book
- CONTEXT_president
- LOC
- ORG
- TOKEN_States
- HEAD_Obama
- CONTEX

Entity Type
- person
- artist
- book
- politician

Low-dimensional Vector Spaces

Model entity-relation interactions

(Ren et al. WWW’17)
Modeling Mention-Feature Co-Occurrences in Corpus

• **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)
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 business-themed game show, NBC’s The Apprentice</td>
</tr>
</tbody>
</table>

**S1: Donald Trump**

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

Entity: *Donald Trump*
Current Distant Supervision: 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 <em>Donald Trump</em>, the recently inaugurated president of the <em>United States</em>.</td>
</tr>
</tbody>
</table>

Type labels for relation mention:

<table>
<thead>
<tr>
<th>E1: <em>Donald J. Trump</em></th>
<th>E2: <em>United States</em></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: 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

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

(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 $(z, m_1, m_2)$:

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

$\sum \sum_{i \in Z} \max \{0, 1 + \tau(z_i) - \tau(z_v)\}$

Positive relation triple

Negative relation triple

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

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

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

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

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

(Ren et al., WWW’17)
Type Inference in CoType

• 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 address to troops at the U.S. Central Command headquarters</td>
</tr>
</tbody>
</table>

Vectors for text features

Test mention: $S_i_{-Trump}$

(Ren et al., WWW’17), (Ren et al., KDD’16)
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
- **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
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>CoType (WWW’17)</td>
<td>0.60</td>
<td>0.39</td>
<td>0.74</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 et 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
Conclusion

• **Problem**: turning text corpus into structured facts (i.e., typed entities/relationships)

• **Contributions**
  • Model entity & relation spaces jointly
  • Noise-robust modeling of distant supervision

• A principled approach to manage, explore and analyze “Big Text Data”

---

Massive corpus → Corpus-to-network → Structured Network → Network-to-knowledge → Knowledge
Thank you! Q & A

• **Problem**: turning text corpus into structured facts (i.e., typed entities/relationships)

• **Contributions**
  • Model entity & relation spaces jointly
  • Noise-robust modeling of distant supervision

• A principled approach to manage, explore and analyze “Big Text Data”