Using Entity Information from a Knowledge Base to Improve Relation Extraction

Lan Du\textsuperscript{1}, Anish Kumar\textsuperscript{2}, Mark Johnson\textsuperscript{2} and Massimiliano Ciaramita\textsuperscript{3}

\textsuperscript{1}Faculty of Information Technology, Monash University, Australia
\textsuperscript{2}Department of Computing, Macquarie University, Sydney, Australia
\textsuperscript{3}Google, Zurich, Switzerland

Presented at ALTA2015 in Sydney on 08/12/2015
Relation Extraction

Surface Pattern
Michael Bay, the director of Transformers, visited Paris yesterday.

Extracted relation
film_director(Michael Bay, Transformers)

A typical relation extraction system functions as a pipeline:

1. Perform named entity recognition (NER) to identify entities in sentences.
2. Perform entity disambiguation to link the entity mentions found in sentences to their database entries.
3. Use the context in which these entity mentions co-occur to predict the relationship between the entities.
   - The path in a syntactic parse between two mentions in a sentence
Motivation

- Freebase entities have a property called notable type

Example, “Jim Jones”

- organization/organization_founder
- music/composer
- baseball/baseball_player
- government/politician

The notable type information provides much finer-grained information about “Jim Jones” than just the NE category.

- By adding notable type, we want to learn the generalization that
  - the politician Jim Jones is likely to stand for election,
  - while the baseball player is likely to be involved in sport activities.
Relation Extraction as Matrix Completion

**Training data:** 200k tuples from NYT 2000–2010

<table>
<thead>
<tr>
<th>4k syntactic patterns</th>
<th>340k notable type syntactic patterns N</th>
<th>51 FreeBase relations</th>
</tr>
</thead>
<tbody>
<tr>
<td>X-is-producer-of-Y</td>
<td>X-director-of-Y</td>
<td>Person(X)-director-of-Film(Y)</td>
</tr>
<tr>
<td>Year</td>
<td>Year</td>
<td>film/director(X,Y)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Training data:</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>200k tuples from NYT</td>
<td>2000–2010</td>
</tr>
<tr>
<td>⟨Steven Spielberg, E.T.⟩</td>
<td>✓ ✓ ✓ ✓ ✓ ✓</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Testing data:</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>10k tuples from NYT</td>
<td>1990–2000</td>
</tr>
<tr>
<td>⟨Michael Bay, Transformers⟩</td>
<td>✗ ✓ ✗ ✓ ✗ ?</td>
</tr>
</tbody>
</table>

**Figure:** The organisation of the training data $\mathcal{O}$ and the matrix $\Theta$ in a matrix completion formulation of the relation extraction task.

- Entity pair $t$ (row in $\Theta$), e.g., $t = ⟨\text{Michael Bay}, \text{Transformers}⟩$.
- Relation $r$ (column in $\Theta$), e.g., $r = \text{X-the-director-of-Y}$.
- Type information: $n(\text{Michael Bay}) = \text{Person}$ and $n(\text{Transformers}) = \text{Film}$.
- Entry $\theta_{t,r}$ is the log odds of relation $r \in \mathcal{R}$ holding of tuple. $t \in \mathcal{T}$
Riedel et al’s “Universal Schemas” models (Riedel et al. 2013)

- Latent Feature Model ($\Theta^N$):
  $$\theta^F_{r,t} = a_r \cdot v_t,$$
  where $a_r$: latent feature vector for relation $r$, $v_t$: latent feature vector for an entity pair.

- Neighbourhood Model ($\Theta^F$):
  $$\theta^N_{r,t} = \sum_{\langle r',t \rangle \in \mathcal{O} \setminus \{\langle r,t \rangle\}} w_{r,r'},$$
  where $w_{r,r'}$ a real-valued weight corresponds to a directed association strength between relations $r$ and $r'$

- Entity Model ($\Theta^E$):
  $$\theta^E_{r,t} = \sum_{i=1}^{2} d_{r,i} \cdot u_{t_i},$$
  where $d_{r,i}$ and $u_{t_i}$ are $K$-dimensional vectors associated with the $i$th argument slot of relation $r$ and the entity $t_i$ respectively.
Notable Type Extensions of the “Universal Schemas” models

- Latent Feature Model ($\Theta^N$):
  \[ \theta_{r,t}^{F'} = a_r \cdot (v_t + v_{n(t)}'), \]
  where $v_{n(t)}'$: latent feature vector based on the notable types of the entities

- Neighbourhood Model ($\Theta^F$):
  \[ \theta_{r,t}^{N'} = \sum_{\langle r',t \rangle \in \mathcal{O} \setminus \{\langle r,t \rangle \}} w_{r,r'} + w'_{r,\langle r',n(t) \rangle}, \]
  where $w'_{r,\langle r',n(t) \rangle}$: a matrix of weights relating the relations
  $\mathcal{N}' = \{\langle r, n(t) \rangle : \langle r, t \rangle \in \mathcal{O} \}$.

- Entity Model ($\Theta^E$):
  \[ \theta_{r,t}^{E'} = \sum_{i=1}^{2} d_{r,i} \cdot (u_{t_i} + u_{n(t_i)}'), \]
  where $u_{n(t_i)}'$: a latent vector associated with the notable types of the entities.
Inference for Model Parameters

- Assume that if $\langle r, t^+ \rangle \in \mathcal{O}$ (i.e., is observed in the training data) then $\theta_{r,t^+} > \theta_{r,t^-}$ for all $\langle r, t^- \rangle \notin \mathcal{O}$ (i.e., not observed in the training data).
- The training objective is to maximise

$$\ell = \sum_{\langle r,t^+ \rangle \in \mathcal{O}} \sum_{\langle r,t^- \rangle \notin \mathcal{O}} \ell_{\langle r,t^+ \rangle,\langle r,t^- \rangle}$$

where: $\ell_{\langle r,t^+ \rangle,\langle r,t^- \rangle} = \log \sigma(\theta_{r,t^+} - \theta_{r,t^-})$, and $\theta_{r,t} = \theta_{r,t}^N + \theta_{r,t}^F + \theta_{r,t}^E$ or $\theta_{r,t} = \theta_{r,t}^{N'} + \theta_{r,t}^{F'} + \theta_{r,t}^{E'}$, depending on whether the submodels with notable type extensions are used.
- Stochastic gradient ascent + L2 regularisation.
Experiment Dataset & Evaluation Procedure

Dataset: the New York Times corpus (Sandhaus, 2008)
- Training on NYT 2000-10 corpus containing Surface patterns + aligned Freebase relations
- Testing on NYT 1990-99 corpus to predict 19 Freebase relations

Evaluation Procedure: each of the 19 Freebase relations $r$ as a query, and evaluate the ranking of the entity tuples $t$ returned according to $\theta_{r,t}$.
- For each relation $r$ pool the highest-ranked 100 tuples produced by each of the models and manually evaluate their accuracy (e.g., by inspecting the original document if necessary).

Metrics
- Mean Average Precision (MAP)
- Weighted Mean Average Precision (Weighted MAP)
Experiments with Notable Types

Models
- NF: Neighbourhood model + Latent feature model
- NFE: NF + Entity model
- $NF^T$: the type-based extension of NF
- $NFE^T$: the type-based extension of NFE

<table>
<thead>
<tr>
<th>Relation</th>
<th>#</th>
<th>NF</th>
<th>$NF^T$</th>
<th>NFE</th>
<th>$NFE^T$</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAP</td>
<td></td>
<td>0.43</td>
<td>0.49</td>
<td>0.42</td>
<td>0.48</td>
</tr>
<tr>
<td>Weighted MAP</td>
<td></td>
<td>0.55</td>
<td>0.64</td>
<td>0.56</td>
<td>0.62</td>
</tr>
</tbody>
</table>

Table: Averaged precision and mean average precision results.

Improvement (statistically significant ($p < 0.05$) with signed test)
- MAP: 6% higher for both $NF^T$ and $NFE^T$
- Weighted MAP: 9% and 6% higher for $NF^T$ and $NFE^T$ respectively
Ablation Experiments

- **Goal:** study how and where the notable type information improves relation extraction.
- **Method:**
  1. Classified all entities into these four NE classes:
     - If an entity has a FreeBase “people/person” type, then we assigned it to the NE class PERSON;
     - If an entity has a “location/location” type, then its NE class is LOCATION;
     - If an entity has a “organisation/organisation” type, then its NE class is ORGANISATION.
     - All entities not classified as PERSON, LOCATION, or ORGANISATION were labelled MISC.
  2. Ablation: For each NE class $c$ in turn, we replaced the notable type information for entities not classified as $c$ with their NE class.
     - For example, when $c = \text{PERSON}$, only entities with the NE label PERSON had notable type information, and the notable types of all other entities was replaced with their NE labels.
- **Model used:** NEF$^T$
# Ablation Experiments

<table>
<thead>
<tr>
<th>Relation</th>
<th>#</th>
<th>NE</th>
<th>NFE</th>
<th>NE+P</th>
<th>NE+L</th>
<th>NE+O</th>
<th>NE+M</th>
</tr>
</thead>
<tbody>
<tr>
<td>person/place_of_birth</td>
<td>30</td>
<td>0.52</td>
<td><strong>0.57</strong></td>
<td>0.54</td>
<td>0.50</td>
<td>0.50</td>
<td>0.54</td>
</tr>
<tr>
<td>author/works_written</td>
<td>38</td>
<td>0.57</td>
<td>0.53</td>
<td><strong>0.61</strong></td>
<td>0.56</td>
<td>0.57</td>
<td>0.49</td>
</tr>
<tr>
<td>composer/compositions</td>
<td>4</td>
<td>0.35</td>
<td>0.42</td>
<td><strong>0.51</strong></td>
<td>0.37</td>
<td>0.35</td>
<td>0.45</td>
</tr>
<tr>
<td>film/directed_by</td>
<td>5</td>
<td>0.30</td>
<td>0.35</td>
<td><strong>0.41</strong></td>
<td>0.27</td>
<td>0.27</td>
<td><strong>0.41</strong></td>
</tr>
<tr>
<td>film/produced_by</td>
<td>3</td>
<td>0.20</td>
<td>0.26</td>
<td>0.29</td>
<td>0.18</td>
<td>0.19</td>
<td><strong>0.40</strong></td>
</tr>
<tr>
<td>person/religion</td>
<td>5</td>
<td>0.22</td>
<td>0.23</td>
<td>0.21</td>
<td>0.22</td>
<td>0.28</td>
<td><strong>0.53</strong></td>
</tr>
<tr>
<td>sports_team/league</td>
<td>4</td>
<td>0.53</td>
<td>0.64</td>
<td>0.54</td>
<td>0.52</td>
<td><strong>0.75</strong></td>
<td><strong>0.75</strong></td>
</tr>
<tr>
<td>company/founders</td>
<td>7</td>
<td>0.22</td>
<td>0.28</td>
<td>0.28</td>
<td>0.21</td>
<td><strong>0.29</strong></td>
<td>0.22</td>
</tr>
<tr>
<td>person/nationality</td>
<td>51</td>
<td>0.19</td>
<td>0.45</td>
<td>0.23</td>
<td><strong>0.50</strong></td>
<td>0.20</td>
<td>0.21</td>
</tr>
<tr>
<td>broadcast/area_served</td>
<td>8</td>
<td>0.32</td>
<td>0.30</td>
<td>0.33</td>
<td><strong>0.38</strong></td>
<td>0.31</td>
<td>0.29</td>
</tr>
<tr>
<td><strong>MAP</strong></td>
<td></td>
<td>0.45</td>
<td>0.48</td>
<td>0.48</td>
<td>0.46</td>
<td>0.47</td>
<td><strong>0.50</strong></td>
</tr>
<tr>
<td><strong>Weighted MAP</strong></td>
<td></td>
<td>0.57</td>
<td><strong>0.62</strong></td>
<td>0.59</td>
<td>0.60</td>
<td>0.58</td>
<td>0.60</td>
</tr>
</tbody>
</table>

**Table:** Results of ablation experiments on the NFE$^T$ model. **NE**: all entities have only the NE class information. **NE+P**: Only PERSON entities have notable type information. **NE+L**: only LOCATION entities have notable type information. **NE+O**: only ORGANISATION entities have notable type information. **NE+M**: only MISC entities have notable type information.
Conclusion & Future Work

Conclusion:

▶ Adding Freebase types to Relation Extraction significantly improves both the MAP (6%) and weighted MAP scores (7.5%).
▶ Ablation experiments showed that the notable type information improves relation extraction more than NER tags across a wide range of entity types and relations.

Future work:

▶ Develop methods for exploiting other information available in FreeBase to improve a broad range of natural language processing and information extraction tasks.

Thanks very much!