Simple Large-scale Relation Extraction from Unstructured Text

Christos Christodouloupolos and Arpit Mittal
Amazon Research Cambridge
“Alexa, what books did Carrie Fisher write?”

“The books that Carrie Fisher is an author of are Delusions of Grandma, Shockaholic, Surrender the Pink, Postcards from the Edge, The Best Awful There Is and Wishful Drinking.”
Alexa Knowledge Base

Named relations between entities

Carrie Fisher is the author of Postcards from the Edge, which is an instance of book.
Alexa Knowledge Base
Sources of knowledge:
1. Human authorship
2. Structured information
3. Unstructured information
Knowledge from Unstructured Text

The Goal:
Carrie Fisher wrote several semi-autobiographical novels, including Postcards from the Edge.
Knowledge from Unstructured Text

The Goal:

Carrie Fisher wrote several semi-autobiographical novels, including *Postcards from the Edge*. 
Knowledge from Unstructured Text

**The Goal:**
*Carrie Fisher* wrote several semi-autobiographical novels, including *Postcards from the Edge.*
Knowledge from Unstructured Text

The Goal: Carrie Fisher wrote several semi-autobiographical novels, including Postcards from the Edge.

Entity Recognition
Entity Resolution
Relation Extraction
Likelihood Estimation

carrie fisher
postcards from the edge
Knowledge from Unstructured Text

The Goal:

Carrie Fisher wrote several semi-autobiographical novels, including *Postcards from the Edge.*

[carrie fisher] is the author of [postcards from the edge]
Knowledge from Unstructured Text

The Goal:
Carrie Fisher wrote several semi-autobiographical novels, including Postcards from the Edge.

[carrie fisher] [is the author of] [postcards from the edge]

Ontological constraints
Entity embeddings
Distributional information

Entity Recognition
Entity Resolution
Relation Extraction
Likelihood Estimation

98% likelihood
Relation Extraction Approaches

• Rule-based
• Fully supervised
• Unsupervised

• **Distant/weakly supervised**
  • Snow, Jurafsky, Ng, 2005
  • Main assumption: if two entities are linked by a relation, any sentence containing both sentences is *likely* to express that relation
    • [steven spielberg] [is the director of] [saving private ryan]
    • “Spielberg’s film Saving Private Ryan is based on…”
Distant supervision label generation

Wikipedia → Chunking → PoS → Tagging → Entity denotations (surface forms) → Gazetteers → Entity pairs (KB IDs)
Distant supervision label generation

1. Wikipedia
2. Chunking PoS Tagging
3. Entity denotations (surface forms)
4. Gazetteers
5. Entity pairs (KB IDs)
6. Ontological Constraints
7. Check against KB
8. Positive Label
9. Negative Label

Check against KB with Ontological Constraints to generate Positive or Negative Label.
Distant supervision label generation

His studies were interrupted by army service and at the end of the war he was forced to return... [the second world war] [is an instance of] [cause of death]

In the intro to the song, Fred Durst makes reference to... [intro 15367][is an instance of] [song]

Turner also released one album and several singles under the moniker Repeat. [the singles the 2011 album] [is an instance of] [album]
Distant supervision label generation

- **Wikipedia page**
  - URL → KB ID lookup
  - Main entity (KB ID)
  - KB
  - Related entities (x) (KB IDs)
    - rel(x₁, main)
    - rel(main, x₂)
  - KB ID → Denotations lookup
  - Entity denotations (x + main strings)

- **Wikipedia**
  - Chunking PoS Tagging
  - Entity denotations (surface forms)
  - Gazetteers
  - Entity pairs (KB IDs)

- **Ontological Constraints**
  - Check against KB
    - YES → Positive Label
    - NO → Negative Label
Distant supervision label generation

Entity denotations
(x + main strings)

Entity pairs
(KB IDs)

Ontological
Constraints

Check against KB
(Bloom filters)

YES
NO

Positive
Label

Negative
Label
Distant supervision label generation

**Wikipedia page** → **URL → KB ID lookup** → **Main entity (KB ID)** → **KB** → **Related entities (x) (KB IDs)** → **KB ID → Denotations lookup** → **Entity denotations (x + main strings)**

*Call Your Girlfriend* was written by Robyn, Alexander Kronlund and Klas Åhlund, with the latter producing the *song*.

[call your girlfriend 3] [is an instance of] [song]

*Forget Her* is a *song* by Jeff Buckley.

[forget her] [is an instance of] [song]

The *Subei Mongol Autonomous County* is an autonomous *county* within the prefecture-level city of Jiuquan in the northwestern Chinese province of Gansu.

[subei mongol autonomous county] [is an instance of] [chinese county]
Relation extraction

- HypeNET (Shwartz and Goldberg, 2016)
- Hyponyms [is an instance of] only
  - LexNET extends to multiple relations
### Alexa KB

<table>
<thead>
<tr>
<th>Relation</th>
<th>HypeNET</th>
</tr>
</thead>
<tbody>
<tr>
<td>[is an instance of]</td>
<td>94.29 (0.21)</td>
</tr>
<tr>
<td>[is the birthplace of]</td>
<td>85.57 (0.26)</td>
</tr>
<tr>
<td>[applies to]</td>
<td>81.98 (1.78)</td>
</tr>
</tbody>
</table>
## Results

<table>
<thead>
<tr>
<th>Alexa KB</th>
<th>HypeNET</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relation</td>
<td></td>
</tr>
<tr>
<td>[is an instance of]</td>
<td>94.29 (0.21)</td>
</tr>
<tr>
<td>[is the birthplace of]</td>
<td>85.57 (0.26)</td>
</tr>
<tr>
<td>[applies to]</td>
<td>81.98 (1.78)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Wikidata</th>
<th>HypeNET</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relation</td>
<td></td>
</tr>
<tr>
<td>instance of (P31)</td>
<td>93.90 (0.21)</td>
</tr>
<tr>
<td>birthplace of (P19)</td>
<td>92.06 (0.90)</td>
</tr>
<tr>
<td>part of (P527)</td>
<td>48.73 (2.59)</td>
</tr>
</tbody>
</table>
Relation extraction

• fastText (Joulin et al., 2016)
• Linear model
  • One hidden layer
  • Rank constraint
HypeNET **equally good** as the much simpler fastText with the same input features.

<table>
<thead>
<tr>
<th>Relation</th>
<th>HypeNET</th>
<th>fastText</th>
</tr>
</thead>
<tbody>
<tr>
<td>[is an instance of]</td>
<td>94.29 (0.21)</td>
<td>94.31 (0.03)</td>
</tr>
<tr>
<td>[is the birthplace of]</td>
<td>85.57 (0.26)</td>
<td><strong>87.63</strong> (0.01)</td>
</tr>
<tr>
<td>[applies to]</td>
<td>81.98 (1.78)</td>
<td><strong>86.17</strong> (0.01)</td>
</tr>
</tbody>
</table>
Results

HypeNET equally good as the much simpler fastText with the same input features.

<table>
<thead>
<tr>
<th>Relation</th>
<th>HypeNET</th>
<th>fastText</th>
</tr>
</thead>
<tbody>
<tr>
<td>[is an instance of]</td>
<td>94.29 (0.21)</td>
<td>94.31 (0.03)</td>
</tr>
<tr>
<td>[is the birthplace of]</td>
<td>85.57 (0.26)</td>
<td>87.63 (0.01)</td>
</tr>
<tr>
<td>[applies to]</td>
<td>81.98 (1.78)</td>
<td>86.17 (0.01)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Relation</th>
<th>HypeNET</th>
<th>fastText</th>
</tr>
</thead>
<tbody>
<tr>
<td>instance of (P31)</td>
<td>93.90 (0.21)</td>
<td>96.44 (0.01)</td>
</tr>
<tr>
<td>birthplace of (P19)</td>
<td>92.06 (0.90)</td>
<td>93.05 (0.07)</td>
</tr>
<tr>
<td>part of (P527)</td>
<td>48.73 (2.59)</td>
<td>72.87 (0.16)</td>
</tr>
</tbody>
</table>
Results

HypeNET equally good as the much simpler fastText with the same input features.

MaxEnt results show that features alone are not enough. Need to create higher-dimensional representations of discrete features.

<table>
<thead>
<tr>
<th>Relation</th>
<th>Alexa KB</th>
<th>Wikidata</th>
</tr>
</thead>
<tbody>
<tr>
<td>[is an instance of]</td>
<td>HypeNET 94.29 (0.21)</td>
<td>fastText 94.31 (0.03)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[is the birthplace of]</td>
<td>HypeNET 85.57 (0.26)</td>
<td>fastText 87.63 (0.01)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[applies to]</td>
<td>HypeNET 81.98 (1.78)</td>
<td>fastText 86.17 (0.01)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Relation</th>
<th>Alexa KB</th>
<th>Wikidata</th>
</tr>
</thead>
<tbody>
<tr>
<td>instance of (P31)</td>
<td>HypeNET 93.90 (0.21)</td>
<td>fastText 96.44 (0.01)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>birthplace of (P19)</td>
<td>HypeNET 92.06 (0.90)</td>
<td>fastText 93.05 (0.07)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>part of (P527)</td>
<td>HypeNET 48.73 (2.59)</td>
<td>fastText 72.87 (0.16)</td>
</tr>
</tbody>
</table>
Summary

• New method for entity resolution
  • Page-specific gazetteers

• Features are important
  • HypeNET vs fastText

• Feature representation is important
  • fastText vs MaxEnt
Future directions

• Enhanced entity recognition
• Use of human annotation for seeding supervision
• Expanding to multiple sources of text
• Coverage of multiple languages
Thanks!