Neural Relation Extraction from Unstructured Texts

Jinhua Du
ADAPT Centre, Dublin City University, Ireland
Outline

✓ What is Relation Extraction
✓ Distant Supervision for RE
✓ Multi-Level Structured Attention
✓ Application
An Example: What is Relation Extraction

**Company report:** “International Business Machines Corporation (IBM or the company) was incorporated in the State of New York on June 16, 1911, as the Computing-Tabulating-Recording Co. (C-T-R)...”

**Standard Information Extraction Task:**
- Company: IBM
- Location: New York
- Date: June 16, 1911
- Original-Name: Computing-Tabulating-Recording Co.

**Relation Extraction Task:**
- Founding-year (IBM, 1911)
- Founding-location (IBM, New York)
What is Relation Extraction

What is it?

- A fundamental task in Information Extraction
- Definition: Given a sentence $S$ with the annotated pairs of nominals $e_1$ and $e_2$, the goal is to identify the relation $r$ from a predefined relation set $R$ for the entity pair $(e_1, e_2)$, written in the form of a triple $(e_1, r, e_2)$ or $(e_2, r, e_1)$
Automated Content Extraction (ACE): NIST

- 17 relations from 2008 “Relation Extraction Task”
Relation Types (closed domain)

SemEval 2010 Task 8

- 9 relations without directionality (plus ‘other’)
- 19 relations with directionality (including ‘other’)

<table>
<thead>
<tr>
<th>Labels</th>
<th>Train</th>
<th>Dev</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entity-Origin((e_1, e_2))</td>
<td>454</td>
<td>114</td>
<td>211</td>
</tr>
<tr>
<td>Entity-Origin((e_2, e_1))</td>
<td>118</td>
<td>30</td>
<td>47</td>
</tr>
<tr>
<td>Entity-Destination((e_1, e_2))</td>
<td>675</td>
<td>169</td>
<td>291</td>
</tr>
<tr>
<td>Entity-Destination((e_2, e_1))</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Component-Whole((e_1, e_2))</td>
<td>376</td>
<td>94</td>
<td>162</td>
</tr>
<tr>
<td>Component-Whole((e_2, e_1))</td>
<td>376</td>
<td>95</td>
<td>150</td>
</tr>
<tr>
<td>Product-Producer((e_1, e_2))</td>
<td>258</td>
<td>65</td>
<td>108</td>
</tr>
<tr>
<td>Product-Producer((e_2, e_1))</td>
<td>315</td>
<td>79</td>
<td>123</td>
</tr>
<tr>
<td>Instrument-Agency((e_1, e_2))</td>
<td>77</td>
<td>20</td>
<td>22</td>
</tr>
<tr>
<td>Instrument-Agency((e_2, e_1))</td>
<td>325</td>
<td>82</td>
<td>134</td>
</tr>
<tr>
<td>Content-Container((e_1, e_2))</td>
<td>299</td>
<td>75</td>
<td>153</td>
</tr>
<tr>
<td>Content-Container((e_2, e_1))</td>
<td>132</td>
<td>34</td>
<td>39</td>
</tr>
<tr>
<td>Cause-Effect((e_1, e_2))</td>
<td>275</td>
<td>69</td>
<td>134</td>
</tr>
<tr>
<td>Cause-Effect((e_2, e_1))</td>
<td>527</td>
<td>132</td>
<td>194</td>
</tr>
<tr>
<td>Message-Topic((e_1, e_2))</td>
<td>392</td>
<td>98</td>
<td>210</td>
</tr>
<tr>
<td>Message-Topic((e_2, e_1))</td>
<td>115</td>
<td>29</td>
<td>51</td>
</tr>
<tr>
<td>Member-Collection((e_1, e_2))</td>
<td>62</td>
<td>16</td>
<td>32</td>
</tr>
<tr>
<td>Member-Collection((e_2, e_1))</td>
<td>489</td>
<td>123</td>
<td>201</td>
</tr>
<tr>
<td>other</td>
<td>1,128</td>
<td>282</td>
<td>454</td>
</tr>
</tbody>
</table>
## Relation Types (open domain)

- **Freebase:** thousand relations/million entities

<table>
<thead>
<tr>
<th>Relation name</th>
<th>Size</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>/people/person/nationality</td>
<td>281,107</td>
<td>John Dugard, South Africa</td>
</tr>
<tr>
<td>/location/location/contains</td>
<td>253,223</td>
<td>Belgium, Nijlen</td>
</tr>
<tr>
<td>/people/person/profession</td>
<td>208,888</td>
<td>Dusa McDuff, Mathematician</td>
</tr>
<tr>
<td>/people/person/place_of_birth</td>
<td>105,799</td>
<td>Edwin Hubble, Marshfield</td>
</tr>
<tr>
<td>/dining/restaurant/cuisine</td>
<td>86,213</td>
<td>MacAyo’s Mexican Kitchen, Mexican</td>
</tr>
<tr>
<td>/business/business_chain/location</td>
<td>66,529</td>
<td>Apple Inc., Apple Inc., South Park, NC</td>
</tr>
<tr>
<td>/biology/organism_classification_rank</td>
<td>42,806</td>
<td>Scorpaeniformes, Order</td>
</tr>
<tr>
<td>/film/film/genre</td>
<td>40,658</td>
<td>Where the Sidewalk Ends, Film noir</td>
</tr>
<tr>
<td>/film/film/language</td>
<td>31,103</td>
<td>Enter the Phoenix, Cantonese</td>
</tr>
<tr>
<td>/film/film/country</td>
<td>27,217</td>
<td>Calopteryx, Calopterygidae</td>
</tr>
<tr>
<td>/film/writer/film</td>
<td>23,856</td>
<td>Turtle Diary, United States</td>
</tr>
<tr>
<td>/film/director/film</td>
<td>23,539</td>
<td>Irving Shulman, Rebel Without a Cause</td>
</tr>
<tr>
<td>/film/producer/film</td>
<td>22,079</td>
<td>Michael Mann, Collateral</td>
</tr>
<tr>
<td>/people/deceased_person/place_of_death</td>
<td>18,814</td>
<td>Diane Eskenazi, Aladdin</td>
</tr>
<tr>
<td>/music/artist/origin</td>
<td>18,619</td>
<td>John W. Kern, Asheville</td>
</tr>
<tr>
<td>/people/person/religion</td>
<td>17,582</td>
<td>The Octopus Project, Austin</td>
</tr>
<tr>
<td>/book/author/works_written</td>
<td>17,278</td>
<td>Joseph Chartrand, Catholicism</td>
</tr>
<tr>
<td>/soccer/football_position/players</td>
<td>17,244</td>
<td>Paul Auster, Travels in the Scriptorium</td>
</tr>
<tr>
<td>/people/deceased_person/cause_of_death</td>
<td>16,709</td>
<td>Midfielder, Chen Tao</td>
</tr>
<tr>
<td>/film/film/music</td>
<td>14,070</td>
<td>Pony Soldiers, Science fiction</td>
</tr>
<tr>
<td>/business/company/industry</td>
<td>13,805</td>
<td>Stavisky, Stephen Sondheim</td>
</tr>
</tbody>
</table>
Two Sub-tasks in RE

➢ Entities recognition
  • Name entities: Person, Organization, Location, Times, Dates, etc.
  • Domain-specific nouns: genes, proteins, diseases, financial terms, etc.

➢ Relation extraction
  • Located in, employed by, married to, etc.
Two Categories of Modelling

Pipeline Modelling

Text ➔ Entity Recognition ➔ Relation Extraction

Mark Elliot Zuckerberg is an American computer programmer and Internet entrepreneur. He is a co-founder of Facebook, and currently operates as its chairman and chief executive officer.

PERSON: Mark Elliot Zuckerberg
LOCATION: American
ORGANISATION: Facebook

Mark Elliot Zuckerberg is founder of Facebook
Two Categories of Modelling

Joint Modelling

Text → Relation Extraction

Input Sentence: The United States President Trump will visit the Apple Inc founded by Steven Paul Jobs


Final Results: {United States, Country-President, Trump} {Apple Inc, Company-Founder, Steven Paul Jobs}
Outline

✓ What is Relation Extraction

✓ Distant Supervision for RE

✓ Multi-Level Structured Attention

✓ Application
Distant Supervision for Relation Extraction

Problems

- SemEval 2007
- SemEval 2010
- BioNLP Shared Task
- ADE-V2

Human Annotated Data

- Data is always important!
- Labeled data is not enough to train a good RE system with a good generalization capability
Distant Supervision for Relation Extraction

➢ DS-RE:
  • Automatic labeling via knowledge bases, such as Freebase, DBpedia
  • It assumes that if one entity pair appearing in some sentences can be observed in a KB with a certain relationship, then these sentences will be labeled as the context of this entity pair and this relationship.

➢ Advantage
  • Effective and efficient method for automatically labeling large-scale training data

➢ Disadvantage
  • It introduces a severe mislabelling problem due to the fact that a sentence that mentions two entities does not necessarily express their relation in a KB
An Example

Bill Gates

FounderOf

Microsoft

<table>
<thead>
<tr>
<th>Relation</th>
<th>Instance</th>
</tr>
</thead>
<tbody>
<tr>
<td>FounderOf</td>
<td>Microsoft was founded on April 4, 1975, by Bill Gates and Paul Allen in Albuquerque, New Mexico.</td>
</tr>
<tr>
<td>ChairmanOf</td>
<td>In February 2014 Gates stepped down as chairman from Microsoft but continued to serve as a board member.</td>
</tr>
<tr>
<td>Other/NA</td>
<td>Largely on the strength of Microsoft’s success, Gates amassed a huge paper fortune as the company’s largest individual shareholder.</td>
</tr>
</tbody>
</table>
DS-RE is different from the traditional RE
  • It is a multi-instance learning problem
  • It is a multi-label classification problem
  • It contains a lot of noise

Distant Supervision Data
  • New York Times
  • Google’s RE Corpus
  • NIST KBP
  • Portuguese DBpedia
Evaluation Metrics of RE

- **Compute P/R/F1**

  \[
  P = \frac{\# \text{ of correctly extracted relations}}{\text{Total } \# \text{ of extracted relations}}
  \]

  \[
  R = \frac{\# \text{ of correctly extracted relations}}{\text{Total } \# \text{ of gold relations}}
  \]

  \[
  F_1 = \frac{2PR}{P + R}
  \]

- **PR Curve**

- **AUC**

**Precision-Recall Curves**

- CNN+ATT
- PCNN+ATT
- BiGRU+ATT
- BiGRU+ATT
- MLMS-ATT-1
- MLMS-ATT-2
Outline

✓ What is Relation Extraction
✓ Distant Supervision for RE
✓ Multi-Level Structured Attention
✓ Application
Deep Learning for Supervised RE

➢ Neural Networks for RE
  • Convolutional NN
  • Recurrent NN (LSTM, GRU, Bidirectional RNN)
  • Attention mechanism

➢ Like other NLP tasks, neural relation extraction has become the state-of-the-art.
Baseline: RNN with Muti-Level Attentions for DS-RE

MLSSA-ATT-O
1-D word-level and 1-D sentence-level attention

softmax $y$

Output Layer

Weighted Representation Layer

Sentence-level Attention Layer

Weighted Representation Layer

Word-Level Attention

BiLSTM Layer

Embedding Layer

Input Layer

Multiple Instances in a Bag $G$

$S_1$ ... $X_N$ ... $S_J$
Our Work: Multi-Level Structured Self-Attention Mechanism

E-Mail: jinhua.du@adaptcentre.ie
Our Work in Accenture for RE

Relation Extraction with Multi-Level and Multi-Scale Self-Attention

• **Motivation:**
  - fully use contextual knowledge in the input sentence
  - select valid instances, and surpass the noisy instances

• **Results:**

![Precision-Recall Curves](chart.png)
Outline

✓ What is Relation Extraction
✓ Distant Supervision for RE
✓ Multi-Level Structured Attention
✓ Application
Applications of Relation Extraction: A Case Study: Anti-money Laundering Monitoring

➢ Money Laundering: three stages
   • Placement
   • Layering
   • Integration

➢ Transaction Monitoring solutions: attempt to detect high risk or out of character funds transfers that may indicate money laundering activity

E-Mail: jinhua.du@adaptcentre.ie
**Basic Workflow**

- Suspicious Transaction
- Key Info Extraction: Customer, Beneficiary, Organisation, Location
- Relation Extraction from Unstructured texts
- New Knowledge Graph
- Existing Knowledge Graph
- Reasoning
- Hidden relations with money launders
- Augmenting


E-Mail: jinhua.du@adaptcentre.ie
This research is supported by Science Foundation Ireland through the ADAPT Centre (Grant 13/RC/2106) (www.adaptcentre.ie) at Dublin City University and Trinity College Dublin, and by SFI Industry Fellowship Programme 2016 (Grant 16/IFB/4490).