Outline

● Part I. Introduction
● Part II. ML for DI
  ○ ML for entity linkage
  ○ ML for data extraction
  ○ ML for data fusion
  ○ ML for schema alignment
● Part III. DI for ML
● Part IV. Conclusions and research direction
What is Data Extraction?

- Definition: Extract structured information, e.g., (entity, attribute, value) triples, from semi-structured data or unstructured data.
What is Data Extraction?

- Definition: Extract structured information, e.g., (entity, attribute, value) triples, from semi-structured data or unstructured data.
Three Types of Data Extraction

- **Closed-world extraction**: align to existing entities and attributes; e.g., (ID_Obama, place_of_birth, ID_USA)

- **ClosedIE**: align to existing attributes, but extract new entities; e.g., (“Xin Luna Dong”, place_of_birth, “China”)

- **OpenIE**: not limited by existing entities or attributes; e.g., (“Xin Luna Dong”, “was born in”, “China”), (“Luna”, “is originally from”, “China”)
Three Types of Data Extraction

- **Closed-world extraction**: align to existing entities and attributes; e.g., (ID_Obama, place_of_birth, ID_USA)

- **ClosedIE**: align to existing attributes, but extract new entities; e.g., (“Xin Luna Dong”, place_of_birth, “China”)

- **OpenIE**: not limited by existing entities or attributes; e.g., (“Xin Luna Dong”, “was born in”, “China”), (“Luna”, “is originally from”, “China”)
35 Years of Data Extraction

Early Extraction
- Rule-based: Hearst pattern, IBM System T
- Tasks: IS-A, events

2008 (Semi-structured data)
- WebTables: search, extraction
- DOM tree: wrapper induction

2013 (Deep ML)
- Use RNN, CNN, attention for RE
- Revisit DOM extraction
- Data programming / Heterogeneous learning

~2005 (Rel. Ex.)
- Relation extraction from texts
  - NER→EL→RE
    - Feature based: LR, SVM
    - Kernel based: SVM
  - Distant supervision
  - OpenIE

1992 (Early-ML)
35 Years of Data Extraction

Early Extraction
- Rule-based: Hearst pattern, IBM System T
- Tasks: IS-A, events

 Extraction from semi-structured data
- WebTables: search, extraction
- DOM tree: wrapper induction

1992 (Early-ML)

~2005 (Rel. Ex.)

Relation extraction from texts
- NER→EL→RE
  - Feature based: LR, SVM
  - Kernel based: SVM
- Distant supervision

2008 (Semi-stru)

Deep learning
- Use RNN, CNN, attention for RE
- Revisit DOM extraction
- Data programming /

2013 (Deep ML)

Come to our VLDB tutorial for text extraction and OpenIE!!
Why Semi-Structured Data?

- Knowledge Vault @ Google showed big potential from DOM-tree extraction [Dong et al., KDD’14][Dong et al., VLDB’14]

<table>
<thead>
<tr>
<th></th>
<th>Accu</th>
<th>Accu (conf ≥ .7)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TXT</td>
<td>0.36</td>
<td>0.52</td>
</tr>
<tr>
<td>DOM</td>
<td>0.43</td>
<td>0.63</td>
</tr>
<tr>
<td>TBL</td>
<td>0.09</td>
<td>0.62</td>
</tr>
<tr>
<td>ANO</td>
<td>1.1M</td>
<td>1.7M</td>
</tr>
<tr>
<td></td>
<td>110M</td>
<td>3M</td>
</tr>
<tr>
<td></td>
<td>0.3M</td>
<td>1.5M</td>
</tr>
<tr>
<td></td>
<td>1.1M</td>
<td>13K</td>
</tr>
<tr>
<td></td>
<td>145M</td>
<td></td>
</tr>
</tbody>
</table>
Wrapper Induction--Vertex [Gulhane et al., ICDE’11]

Extracted relationships

- (Top Gun, type.object.name, “Top Gun”)
- (Top Gun, film.film.genre, Action)
- (Top Gun, film.film.directed_by, Tony Scott)
- (Top Gun, film.film.starring, Tom Cruise)
- (Top Gun, film.film.runtime, “1h 50min”)
- (Top Gun, film.film.release_Date_s, “16 May 1986”)
Wrapper Induction--Vertex [Gulhane et al., ICDE’11]

● Solution: find XPaths from DOM Trees
Wrapper Induction--Vertex [Gulhane et al., ICDE’11]

- Challenge: slight variations from page to page

Figure 2: Example of XPaths corresponding to the *acted in* predicate on two IMDb pages. They differ at two node indices, and the second path corresponds to the *producer of* predicate from the first page.
Wrapper Induction--Vertex [Gulhane et al., ICDE’11]

One website may use multiple templates (Unsupervised-clustering)

Identify representative webpages for annotation

Learn

Cluster Pages → Annotate Pages → Learn XSLT Rules

Sample pages → Annotations

Sample pages → Monitor Rules

Changed sites

Extract

Web site pages → Extract → Records

Combine attr features and textual features to find a general XPath (LR)
Wrapper Induction--Vertex [Gulhane et al., ICDE’11]

- Sample learned XPaths on IMDb
  - //[@itemprop="name"]
  - //*[@class="bp_item bp_text_only"]/*/*/[@class="bp_heading"]
  - //*[following-sibling::*[position()=3]/@class="subheading"]/*[following-sibling::*[position()=1]/@class="attribute"]
  - //*[@class="bp_heading"]
  - //*[preceeding-sibling::node()[normalize-space(.)!="" ][text()="Language:" ]

Ensure high recall

Ensure high precision
Distantly Supervised Extraction

- **Annotation-based extraction**
  - Pros: high precision and recall
  - Cons: does not scale—annotation per cluster per website

- **Distantly-supervised extraction**
  - Step 1. Use seed data to automatically annotate
  - Step 2. Use the (noisy) annotations for training
  - E.g., DeepDive, Knowledge Vault
Distant Supervision [Mintz et al., ACL’09]

Corpus Text

Bill Gates founded Microsoft in 1975.
Bill Gates, founder of Microsoft, …
Bill Gates attended Harvard from …
Google was founded by Larry Page …

Freebase

(Bill Gates, Founder, Microsoft)
(Larry Page, Founder, Google)
(Bill Gates, CollegeAttended, Harvard)

Training Data

[Adapted example from Luke Zettlemoyer]
Distant Supervision [Mintz et al., ACL’09]

**Corpus Text**

Bill Gates founded Microsoft in 1975.
Bill Gates, founder of Microsoft, …
Bill Gates attended Harvard from …
Google was founded by Larry Page …

**Freebase**

(Bill Gates, Founder, Microsoft)
(Larry Page, Founder, Google)
(Bill Gates, CollegeAttended, Harvard)

**Training Data**

(Bill Gates, Microsoft)
Label: Founder
Feature: X founded Y

[Adapted example from Luke Zettlemoyer]
Distant Supervision [Mintz et al., ACL’09]

Corpus Text
Bill Gates founded Microsoft in 1975.
Bill Gates, founder of Microsoft, …
Bill Gates attended Harvard from …
Google was founded by Larry Page …

FREEBASE
(Bill Gates, Founder, Microsoft)
(Larry Page, Founder, Google)
(Bill Gates, CollegeAttended, Harvard)

Training Data
(Bill Gates, Microsoft)
Label: Founder
Feature: X founded Y
Feature: X, founder of Y

[Adapted example from Luke Zettlemoyer]
Distant Supervision [Mintz et al., ACL’09]

<table>
<thead>
<tr>
<th>Corpus Text</th>
<th>Training Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bill Gates founded Microsoft in 1975.</td>
<td>(Bill Gates, Microsoft)</td>
</tr>
<tr>
<td>Bill Gates, founder of Microsoft, …</td>
<td>Label: Founder</td>
</tr>
<tr>
<td>Bill Gates attended Harvard from …</td>
<td>Feature: X founded Y</td>
</tr>
<tr>
<td>Google was founded by Larry Page ...</td>
<td>Feature: X, founder of Y</td>
</tr>
<tr>
<td>Label: CollegeAttended</td>
<td></td>
</tr>
<tr>
<td>Feature: X attended Y</td>
<td></td>
</tr>
<tr>
<td>(Bill Gates, Microsoft)</td>
<td></td>
</tr>
<tr>
<td>(Bill Gates, Founder, Microsoft)</td>
<td></td>
</tr>
<tr>
<td>(Larry Page, Founder, Google)</td>
<td></td>
</tr>
<tr>
<td>(Bill Gates, CollegeAttended, Harvard)</td>
<td></td>
</tr>
</tbody>
</table>

For negative examples, sample unrelated pairs of entities.

[Adapted example from Luke Zettlemoyer]
Distantly Supervised Extraction--Ceres [Lockard et al., VLDB’18]

Automatic Label Generation

Entity Identification → Relation Annotation → Training

Movie entity

Genre

Release Date

Extracted triples
- (Top Gun, type.object.name, “Top Gun”)
- (Top Gun, film.film.genre, Action)
- (Top Gun, film.film.directed_by, Tony Scott)
- (Top Gun, film.film.starring, Tom Cruise)
- (Top Gun, film.film.runtime, “1h 50min”)
- (Top Gun, film.film.release_Date_s, “16 May 1986”)
Distantly Supervised Extraction--Ceres [Lockard et al., VLDB’18]

- Extraction experiments on SWDE benchmark

<table>
<thead>
<tr>
<th>Vertical</th>
<th>Predicate</th>
<th>Vertex++ P</th>
<th>Vertex++ R</th>
<th>Vertex++ F1</th>
<th>CERES-Full P</th>
<th>CERES-Full R</th>
<th>CERES-Full F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Movie</td>
<td>Title</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>Director</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
</tr>
<tr>
<td></td>
<td>Genre</td>
<td>0.88</td>
<td>0.87</td>
<td>0.87</td>
<td>0.93</td>
<td>0.97</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td>MPAA Rating</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>Average</td>
<td>0.97</td>
<td>0.97</td>
<td>0.97</td>
<td>0.97</td>
<td>0.99</td>
<td>0.98</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Vertical</th>
<th>Predicate</th>
<th>Vertex++ P</th>
<th>Vertex++ R</th>
<th>Vertex++ F1</th>
<th>CERES-Full P</th>
<th>CERES-Full R</th>
<th>CERES-Full F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>University</td>
<td>Name</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>Type</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>0.72</td>
<td>0.80</td>
<td>0.76</td>
</tr>
<tr>
<td></td>
<td>Phone</td>
<td>0.97</td>
<td>0.92</td>
<td>0.94</td>
<td>0.85</td>
<td>0.95</td>
<td>0.90</td>
</tr>
<tr>
<td></td>
<td>Website</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>0.90</td>
<td>1.00</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td>Average</td>
<td>0.99</td>
<td>0.98</td>
<td>0.99</td>
<td>0.87</td>
<td>0.94</td>
<td>0.90</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Vertical</th>
<th>Predicate</th>
<th>Vertex++ P</th>
<th>Vertex++ R</th>
<th>Vertex++ F1</th>
<th>CERES-Full P</th>
<th>CERES-Full R</th>
<th>CERES-Full F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Book</td>
<td>Title</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>1.00</td>
<td>0.90</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td>Author</td>
<td>0.97</td>
<td>0.96</td>
<td>0.96</td>
<td>0.72</td>
<td>0.88</td>
<td>0.79</td>
</tr>
<tr>
<td></td>
<td>Publisher</td>
<td>0.85</td>
<td>0.85</td>
<td>0.85</td>
<td>0.97</td>
<td>0.77</td>
<td>0.86</td>
</tr>
<tr>
<td></td>
<td>Publication Date</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
<td>1.00</td>
<td>0.40</td>
<td>0.57</td>
</tr>
<tr>
<td></td>
<td>ISBN-13</td>
<td>0.94</td>
<td>0.94</td>
<td>0.94</td>
<td>0.99</td>
<td>0.19</td>
<td>0.32</td>
</tr>
<tr>
<td></td>
<td>Average</td>
<td>0.93</td>
<td>0.93</td>
<td>0.93</td>
<td>0.94</td>
<td>0.63</td>
<td>0.70</td>
</tr>
</tbody>
</table>

Very high precision
Competent w. Wrapper induction w. manual annotation
## Distantly Supervised Extraction--Ceres

[Lockard et al., VLDB’18]

- Extraction on long-tail movie websites

<table>
<thead>
<tr>
<th>#Websites / #Webpages</th>
<th>33 / 434K</th>
</tr>
</thead>
<tbody>
<tr>
<td>Language</td>
<td>English and 6 other languages</td>
</tr>
<tr>
<td>Domains</td>
<td>Animated films, Documentary films, Financial performance, etc.</td>
</tr>
<tr>
<td># Annotated pages</td>
<td>70K (16%)</td>
</tr>
<tr>
<td>Annotated : Extracted #entities</td>
<td>1 : 2.6</td>
</tr>
<tr>
<td>Annotated : Extracted #triples</td>
<td>1 : 3.0</td>
</tr>
<tr>
<td># Extractions</td>
<td>1.25 M</td>
</tr>
<tr>
<td>Precision</td>
<td>90%</td>
</tr>
</tbody>
</table>
DISTANTLY SUPERVISED EXTRACTION--CERES [Lockard et al., VLDB’18]

● Which model is the best?
  ○ Logistic regression: best results (20K features on one website)
  ○ Random forest: lower precision and recall
  ○ Deep learning??
Challenges in Applying Deep Learning on Extracting Semi-structured Data

- Web layout is neither 1D sequence nor regular 2D grid, so CNN or RNN does not directly apply
Example System: Fonduer [Wu et al., SIGMOD’18]

Fonduer combines a new bi-directional LSTM with multimodal features and weak supervision (specifically data programming).

Attend the talk in Research Session 13!
New version of code coming soon: https://github.com/HazyResearch/fonduer
WebTable Extraction [Limaye et al., VLDB’10]

- Model table annotation using interrelated random variables, represented by a probabilistic graphical model
  - Cell text (in Web table) and entity label (in catalog)
  - Column header (in Web table) and type label (in catalog)
  - Column type and cell entity (in Web table)
WebTable Extraction [Limaye et al., VLDB’10]

- Model table annotation using interrelated random variables, represented by a probabilistic graphical model
  - Pair of column types (in Web table) and relation (in catalog)
  - Entity pairs (in Web table) and relation (in catalog)
Challenges in Applying ML on DX

- Automatic data extraction cannot reach production quality requirement. How to improve precision?

- Every web designer has her own whim, but there are underlying patterns across websites. How to learn extraction patterns on different websites, especially for semi-structured sources?

- ClosedIE throws away too much data. How to apply OpenIE on all kinds of data?
Recipe for Data Extraction

- **Problem definition:** Extract structure from semi- or un-structured data
- **Short answers**
  - Wrapper induction has high prec/rec
  - Distant supervision is critical for collecting training data
  - LR is often effective; more research is needed for DL