AI-Enhanced Knowledge Harvesting

Lili JIANG
Associate Professor, Department of Computing Science
#frAIday@28 May, 2021
ABOUT ME

Lili, JIANG

Before joining in Umeå University in 2016:
• Research scientist, NEC Laboratories Europe, Germany
• Endeavour Research Fellowship, The University of Queensland, Australia
• Post-doctoral researcher, Max-Planck-Institut für Informatik, Germany.
• PhD from Computer Science, China
ABOUT MY RESEARCH

Data sources

- Database
- Text
- SAS
- Public web

Data preprocessing
- Data cleaning
- Data transformation
- Data extraction
- Data deduction
- ...

Knowledge harvesting
- Entity analysis and retrieval
- Knowledge graph via data federation
- Privacy analysis on texts
- ......
FROM DATA TO KNOWLEDGE
APPLICATION-1: ENTITY DISAMBIGUATION

Same mention, Different entities

Paris (France)  

Paris (Ontario, Canada)

Paris (Prince)
APPLICATION-2: E-DISCOVERY

Product Accident → Lawsuit Investigation → E-Discovery
ENTITY ALIAS DISCOVERY IN ENTERPRISE

Different mentions, Same entity

<table>
<thead>
<tr>
<th>department</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Imaging &amp; Printing</td>
</tr>
<tr>
<td>2 Network Group</td>
</tr>
<tr>
<td>3 Human Resource and Recruitment</td>
</tr>
</tbody>
</table>

Sever log
Script
Partner DB
Customer DB
Email
Memo
Files

HTML:
this is a test, which is a test, test, company html
APPLICATION-3: PERSONALITY DETECTION

Demographic info
- Gender
- Age
- Relation status
- Timzone

Social media activities
- Tweets
- FB status
- Network size

Person

Personality traits
- Extraversion
- Agreeableness
- Conscientiousness
- Openness to experience
- Neuroticism
Looking for a special bday dinner for son's birthday. We love sushi! This hit the spot. Impressive communication with Annie, good food, great sushi rice, love the efficiency. Thank you Annie!!
TODAY

• Information extraction
• Knowledge harvesting
• Responsible knowledge harvesting
TODAY

• Information extraction
• Knowledge harvesting
• Responsible knowledge harvesting
WHAT TYPES OF DATA WE TALK ABOUT?

<table>
<thead>
<tr>
<th>City</th>
<th>Price</th>
<th>Hotel</th>
</tr>
</thead>
<tbody>
<tr>
<td>Berlin</td>
<td>120</td>
<td>GrandCity</td>
</tr>
<tr>
<td>Berlin</td>
<td>65</td>
<td>Ibis</td>
</tr>
<tr>
<td>Munich</td>
<td>90</td>
<td>Smart Stay Hotel</td>
</tr>
</tbody>
</table>

Structured Data

Semi-structured Data

Unstructured Data
WHY UNSTRUCTURED DATA

80% of the world’s data is unstructured
<table>
<thead>
<tr>
<th>Raw Data</th>
<th>Structure Facts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Barack Obama was born on August 4, 1961 and he previously served as the president of USA.</td>
<td>(Barack Obama, birth, 4/8/1961)</td>
</tr>
<tr>
<td></td>
<td>(Barack Obama, job, president)</td>
</tr>
<tr>
<td></td>
<td>(Barack Obama, country, USA)</td>
</tr>
</tbody>
</table>
CASE-1: PEOPLE RELATED INFO EXTRACTION

- Machine learning based methods
  - Named entity recognition (NER)
- Pattern-based method
  email: username@(AT)domain-name
  phone number: local format (e.g., 6258657 )
  international format (e.g., 1-626-780-7552)
Dr. Robert Moore is Distinguished Service Professor of Psychology, in the Graduate Center of the Chicago Theological Seminary where he is the Founding Director of the new Institute for Advanced Studies in Spirituality and Wellness. He is married to Margaret Shanahan and lives in Chicago.

Robert Moore, Ph.D., was born on 13 August 1942 at Little Rock AR, USA. He is an psychoanalyst in Chicago Theological Seminary in Illinois, USA. Contact him via 773.288.7474 or Robert@RobertMoore-PhD.com
### EXAMPLE

<table>
<thead>
<tr>
<th>Tag Type</th>
<th>Extraction Pattern</th>
</tr>
</thead>
<tbody>
<tr>
<td>Birth Date</td>
<td>&lt;name&gt; was born on &lt;birth date&gt;</td>
</tr>
<tr>
<td></td>
<td>&lt;name&gt; was born in &lt;birth date&gt;</td>
</tr>
<tr>
<td></td>
<td>&lt;name&gt; ( &lt;birth date&gt; - ###)</td>
</tr>
<tr>
<td></td>
<td>&lt;name&gt; b. &lt;birth date&gt;</td>
</tr>
<tr>
<td></td>
<td>&lt;name&gt; ( &lt;birth date&gt; )</td>
</tr>
<tr>
<td></td>
<td>&lt;name&gt; (### &lt;birth date&gt;)</td>
</tr>
<tr>
<td></td>
<td>&lt;name&gt; ### birth &lt;birthdate&gt;</td>
</tr>
<tr>
<td>Occupation</td>
<td>&lt;name&gt; ( &lt;occupation&gt;, ###)</td>
</tr>
<tr>
<td></td>
<td>&lt;name&gt;, &lt;occupation&gt;</td>
</tr>
<tr>
<td></td>
<td>&lt;name&gt; was/is a/an &lt;occupation&gt;</td>
</tr>
<tr>
<td>URL domain</td>
<td>^https? : // &lt;domain&gt; (/</td>
</tr>
</tbody>
</table>
### CASE-2: ORGANIZATION AND PRODUCTS EXTRACTION

<table>
<thead>
<tr>
<th>Alias Pattern</th>
<th>Concerned Entity</th>
<th>Alias Example</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Abbreviation</strong></td>
<td><strong>Americas</strong> Finance</td>
<td><strong>AMS</strong> Finance</td>
</tr>
<tr>
<td><strong>Acronym</strong></td>
<td>The <em>Technology for Emerging Markets</em> Research Group</td>
<td><strong>TEM</strong> Group</td>
</tr>
<tr>
<td><strong>Expansion</strong></td>
<td><em>LAN</em> Adapter</td>
<td>Network Adapter <em>Local Area Network</em></td>
</tr>
<tr>
<td><strong>StopWords</strong></td>
<td>PSG AP &amp; Japan</td>
<td>PSG AP and Japan</td>
</tr>
<tr>
<td><strong>TokenVariants</strong></td>
<td><em>Graphics</em> Imaging Business</td>
<td>Graphic and Imaging Business</td>
</tr>
<tr>
<td><strong>Order</strong></td>
<td>PDA <em>iPAQ</em> Pocket</td>
<td><em>iPAQ PDA</em> Pocket</td>
</tr>
<tr>
<td><strong>Hierarchy</strong></td>
<td>Canon <em>Cameras EOS Digital Rebel XT</em></td>
<td>EOS Rebel XT</td>
</tr>
</tbody>
</table>
CASE-3: COREFEERENCE RESOLUTION

Grammy-winning singer Albertina Walker, who was known as the "Queen of Gospel," has died at age 81. Close friend and WVON radio host Pam Morris says Walker died Friday. Morris says Walker was “a living legend”..............
## EXAMPLES OF COREFERENCE RESOLUTION

<table>
<thead>
<tr>
<th>Types</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Appositive</td>
<td>Obama........... the US. president</td>
</tr>
<tr>
<td>Nick name</td>
<td>Albertina Walker .......... Queen of Gospel</td>
</tr>
<tr>
<td>First name</td>
<td>Albertina Walker .......... Walker</td>
</tr>
<tr>
<td>Head word</td>
<td>U. N. Security Council .......... Council</td>
</tr>
<tr>
<td>Appellation</td>
<td>Morton A. Zolotow .......... Mr. Zolotow</td>
</tr>
<tr>
<td>Alias</td>
<td>Chiyoda Mutual Life Insurance Company .......... Chiyoda Life</td>
</tr>
</tbody>
</table>
On March, 2017, Mr. Haruhiko Kuroda, governor of Bank of Japan, announced 0.1 percent growth.
On March, 2017, Mr. Haruhiko Kuroda, governor of Bank of Japan, announced 0.1 percent growth.
**DATA FEDERATION**

<table>
<thead>
<tr>
<th>userid</th>
<th>birth</th>
<th>gender</th>
<th>relation status</th>
<th>location</th>
<th>political view</th>
</tr>
</thead>
<tbody>
<tr>
<td>b76b776cf</td>
<td>12.1991</td>
<td>male</td>
<td>single</td>
<td>Sweden</td>
<td>don’t care</td>
</tr>
</tbody>
</table>

- **DB** Demography
- **Text/XML** Online Social activity (FB)
- **public web**

**Blood Pressure** at young age 27 feeling **depressed** :(. my bp is high always, it some time 180/110 ,160/100" ...
TODAY

• Information extraction
• Knowledge harvesting
• Responsible knowledge harvesting
STUDY: PEOPLE NAME DISAMBIGUATION

Lili Jiang, etc. GRAPE: A Graph-Based Framework for Disambiguating People Appearances in Web Search, ICDM 2009
# RESULTS IN IR SYSTEM

<table>
<thead>
<tr>
<th>Query</th>
<th>Mention Clusters</th>
</tr>
</thead>
</table>
| Robert Moore   | **Robert Moore 1:**  
|                | *Occupation:* manager  
|                | *Birthday:* November 14, 1982  
|                | *Email:* rob.moore@spencer-ogden.com  
|                | *Related People:* Du sicher, Das Ansehen, Neu bei  
|                | *Organization:* Cornell University, LPL Financial  |
|                | **Robert Moore 2:**  
|                | *Birthday:* 1942  
|                | *Related People:* William Morrow, Douglas Gillette  
|                | *Organization:* Male Psyche, Chicago Center for Integrative  |
|                | **Robert Moore 3:**  
|                | *Occupation:* associate professor  
|                | *Phone:* 404-413-0056  
|                | *Email:* rmoore@gsu.edu  
|                | *Related People:* Byrdine F. Lewis, Robert E. Moore  
|                | *Organization:* Young School of Policy Studies, Colledge of Arts and Sciences  |
**STUDY: ALIAS DETECTION AND ENTITY LINKING**

- **Entity alias detection**
  - Graphics Imaging Business <-> Graphic and Imaging Business
  - Canon Cameras EOS Digital Rebel XT <-> EOS Rebel XT
  - Leonardo Viccari <-> 20393751, viccari@sa.com
  - usama_bin_laden <-> usama_bin_muhammad_bin_ladin

- **Entity linking**

  [US news] **Washington** correspondent **Robert Moore** covers all major North America stories for ITV News and led our coverage of the US presidential race in 2012. **Robert** previously held this role before moving to Belgium to......
APPROACH: RANDOM WALK

Mentions

- Washington: 0.184
- Robert Moore: 0.138
- Robert: 0.1112

Entities

- Washington D.C.: 0.015
- George Washington: 0.0005
- Washington (state): 0.0112
- Robert Moore: 0.0023
- Robert Pattinson: 0.000021
- Robert L. Moore: 0.0000021
GRAPH EDGE WEIGHT

- Mention-Mention edge
  - coreference resolution classifier on 30 features
- Mention-Entity edge
  - relevance algorithm based on popularity and context
- Entity-Entity edge
  - relevance computation on anchor link in Knowledge Base (Wiki)
STUDY: HETEROGENEOUS FUSION OF CONTENT & STRUCTURE
RESULTS

Xuan-Son Vu, Abhishek Santra, Sharma Chakravarthy, Lili Jiang. Generic Multilayer Network Data Analysis with the Fusion of Content and Structure. CICLing 2019.
TODAY

• Information extraction
• Knowledge harvesting
• Responsible knowledge harvesting
  o Privacy
  o Explainability
RISK OF PRIVACY BREACHES

• In 2019, **2,013** confirmed data breaches. (–varonis.com).

• There was an **80%** increase in the number of people affected by health data breaches from 2017 to 2019.

• In 2019, **14.4 million** consumers became victims of identity fraud. (–IdentityForce)

• Emotional distress is reported by **77.3 percent** of identity theft victims. (–IdentityForce)
STUDY:
COOKIE MINING FOR PRIVACY PRESERVATION

[Diagram showing cookie data with fields such as Unique Identifier, Time Stamp for First Visit, Time Stamp for Previous Visit, Time Stamp for Current Visit, Domain Hash, Number of Sessions, 2 Year Expiration, Name, Content, Domain, Path, Send For, Accessible to Script, Created, Expires]
FINE-GRAIN COOKIE MINING

Cookie example:
253024271.1453122666.239.16.utmcsr=host.com/search/|
utmccn=(10.0.0.5)|utmcmd=organic| utmctr=ail@host.com

Entity alignment:
{NUM}.{TS}{NUM}.{NUM}.{TEXT}={URL}|
{TEXT}={({IP})|{TEXT}={TEXT}|{TEXT}={EMAIL}

Goal: infer the potential risk for people info re-identification.

STUDY: PRIVACY CONCERN ANALYSIS

Motivation

"Blood Pressure at young age 27 feeling depressed :(. my bp is high always, it some time 180/110 ,160/100" ...

Solution

STUDY: MULTIMODAL REVIEW GENERATION WITH PRIVACY AND FAIRNESS AWARENESS

- A unified review generation framework based on word embedding with user privacy
- The notions of user privacy and sentiment fairness for the task of review generation.

RESULTS

<table>
<thead>
<tr>
<th>Image</th>
<th>PRGen\textsubscript{G-PY}</th>
<th>ShowNTell</th>
<th>MRG</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.jpg" alt="Image" /></td>
<td>i’ve been here for a few times and i have never been disappointed. the service was great and the food was great.</td>
<td>to there have you for forfor personi can order on which crab, they ordered tuna shrimp soup it yum</td>
<td>and spicy tuna roll was in the world i ’m not sure what they ’re in but the chicken was tender and the sauce was just a little bit salty.</td>
</tr>
<tr>
<td><img src="image2.jpg" alt="Image" /></td>
<td>my husband and i went here for a late night and it was a great experience. we ordered the margherita pizza and it was delicious. the pizza was good but the crust was thin and chewy.</td>
<td>was a good night i was n’t you food was n’t to be a i was a year minutefor i!! is i i food was good good it i, were open to open store, but i can see if you ’re in for</td>
<td>and spicy tuna sandwich in vegas loved the atmosphere great the service was very friendly and attentive.</td>
</tr>
<tr>
<td><img src="image3.jpg" alt="Image" /></td>
<td>i’ve been wanting to try this place for a while i’m in vegas. i have to say this place is a great place to eat.</td>
<td>was very pictures order of to had than to was i was good good to other places i’ve had other restaurants restaurants mac was very a highlight of meal of lobster meat opinion</td>
<td>and spicy tuna i ’m hoping it would be more but i ’m not sure if it ‘s a dish however the taste buds were the same thing but i just say it was n’t bad.</td>
</tr>
</tbody>
</table>

Applied techniques: multimodal neural network + differential privacy
STUDY (EXPLAINABILITY): CONTEXTUAL BASED EXPLANATION FOR NEURAL NETWORK

Applied technique: contextual based importance method

Sule Anjomshoae, Daniel Omeiza, Lili Jiang. To submit
CONCLUSION

- Information extraction
- Knowledge harvesting
  - Entity resolution (public news, social media)
  - Personality detection (social media, self-writing essay)
  - Heterogeneous data federation (database, social media)
- Responsible knowledge harvesting
  - Differential privacy (social media, review generation)
  - Explainability (explain images on CNN)
- *Ongoing interdisciplinary projects*
  - AI on climate change and infection
  - Covid pandemic analysis and tracking
  - Anomaly detection in energy consumption