KDD Tutorial T39
Building a Large-scale, Accurate and Fresh Knowledge Graph
Yuqing Gao, Jisheng Liang, Benjamin Han, Mohamed Yakout, Ahmed Mohamed
Satori Group, AI+R, Microsoft
Disclaimers

© 2018 Microsoft Corporation. All rights reserved. This document is provided "as-is." Information and views expressed in this document, including URL and other Internet Web site references, may change without notice. You bear the risk of using it.

Some examples are for illustration only and are fictitious. No real association is intended or inferred. This document does not provide you with any legal rights to any intellectual property in any Microsoft product. You may copy and use this document for your internal, reference purposes.
Outline

• Logistics (5 min)
• Part I: Introduction (30 min)
• Part II: Acquiring Knowledge in the Wild (55 min)
• Break (2:30 – 3:00pm, 30 min)
• Part III: Building Knowledge Graph (70 min)
• Break (20 min)
• Part IV: Serving Knowledge to the World (30 min)
Part I: Introduction

Yuqing Gao
Partner General Engineering Manager, Satori Group, Microsoft AI+R
yuga@microsoft.com
What is Knowledge

• Plato’s definition: Justified true belief

Mission:

• Build the best Knowledge Graph in industry that will provide the highest quality of world's knowledge and personal knowledge measured by correctness, coverage, freshness & usage, to enable Agile, intelligent knowledge experiences
What is a Knowledge Graph?

Knowledge represented as entities, edges and attributes

**Personal entity** showing that Tom watched Ghosts of the Abyss

**Edge (i.e. relationship)** showing that “Ghosts of the Abyss” was ‘directed_by’ and ‘produced_by’ James Cameron

<table>
<thead>
<tr>
<th>Key concepts</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Entity</strong></td>
<td>Represent something in the real world</td>
</tr>
<tr>
<td><strong>Edge</strong></td>
<td>Represent relationship</td>
</tr>
<tr>
<td><strong>Attribute</strong></td>
<td>Represent something about an entity</td>
</tr>
<tr>
<td><strong>Ontology</strong></td>
<td>Definition of possible types of entities, relationships and attributes</td>
</tr>
</tbody>
</table>
State of the art knowledge graphs

Minimum set of characteristics of knowledge graphs:
1. mainly describes real world entities and their interrelations, organized in a graph.
2. defines possible classes and relations of entities in a schema.
3. allows for potentially interrelating arbitrary entities with each other.
4. covers various topical domains.

State of Art KGs:
- Cyc and Open Cyc
- Freebase
- Wikidata
- DBpedia
- YAGO
- NELL
- Google Knowledge Vault
- Google KG
- Microsoft Satori KG

Large vertical KGs
- Facebook (social network)
- LinkedIn (people graph)
- Amazon (product graph)

Large production KGs support Google and Bing Search
Research fields

• Research related to knowledge graph refinement:
  • Ontology learning mainly deals with learning a concept level description of a domain, such as a hierarchy (e.g., Cities are Places)

• Approaches for Completion of Knowledge Graphs
  • Methods for Completing Type Assertions
  • Methods for Predicting Relations

• Approaches for Error Detection in Knowledge Graphs
  • Methods for Finding Erroneous Type Assertions
  • Methods for Finding Erroneous Relations
  • Methods for Finding Erroneous Literal Values

• Knowledge extraction
  • Entity linking and disambiguation
  • Fact extraction and verification
Challenges of scaled KGs

Building a small KG is easy - building a vast system like Satori is a huge challenge

Coverage
Have we got the information we need?

Freshness
Is information up to date?

Correctness
Is our information accurate?

Three forces in constant conflict:

- Increased freshness and coverage → Harder to ensure correctness
- Increased correctness → Harder to ensure freshness and coverage
- Correctness is always hard – what is true and correct?
  Particularly critical in today’s world

Will Smith: Single entity, 108K facts assembled from 41 web sites. There are 200 Will Smiths on Wikipedia alone.
Creating High Quality Web-scale Knowledge
AI: ML + NL + Conflation + Inference

- Raw data sources, structured + unstructured
- Schematized, correlated data and relationships
- High quality, conflated and schematized knowledge
Knowledge flywheel in action: World graph

Search queries, views, click throughs, ...

World graph
- People
- Places
- Things
- Actions
- ...

2B+ entities
130B+ Web pages

Web pages, Web documents, Images, ...

Bing
Knowledge flywheel in action: Domain-specific graph

Knowledge acquisition, search, recommendation ...

Domain-specific graph
- People
- Publications
- Fields of Study
- Venues

1B+ Scholarly articles
48K+ Journals
211M+ Authors

Authors, institutions, articles, conferences ...

Microsoft Academic
Knowledge flywheel in action: Work graph

Messages read/sent, Document author/shared, ...

Work graph
- People
- Groups
- Messages
- Activities

8T+ entities
240+ markets
44+ languages

Emails, Messages, Documents, Meetings, ...
How do we bring knowledge systems to life?

- Raw Data
  - Structured + Unstructured
- Knowledge Production
  - High-quality Semantically Organized
Active research and product efforts in knowledge

- Unsupervised knowledge extraction from unstructured data in open domain
- Knowledge graph semantic embedding
- Autonomous knowledge inference & verification
- Real-time knowledge graph with archiving
- Large scale entity linking and disambiguation
- Ultra-scale knowledge representations
- Knowledge system for multi-lingual
- Knowledge Precision vs Comprehensiveness
- ...

Semantic Ontology

Data Ingestion & Extraction

Entity Linking & Conflation

Knowledge Inference

Publishing & Serving
Infusing knowledge: From search to conversation
Richer Data for Entity Pane, Carousel, and Facts Across Segments
Knowledge powered Q&A

Text-based Q&A

Will I qualify for OSAP if I’m new in Canada?

Selected Passages

“Visit the OSAP website for application deadlines. To get OSAP, you have to be eligible. You can apply using an online form, or you can print off the application forms. If you submit a paper application, you must pay an application fee. The online application is free.”

“To be eligible to apply for financial assistance from the Ontario Student Assistance Program (OSAP), you must be a: 1 Canadian citizen; 2 Permanent resident; or 3 Protected person/convention refugee with a Protected Persons Status Document (PPSD),”

“You will not be eligible for a Canada-Ontario Integrated Student Loan, but can apply for a part-time loan through the Canada Student Loans program. There are also grants, bursaries and scholarships available for both full-time and part-time students.”

Answer
No. You won’t qualify.

Knowledge-based Q&A
Bing – knowledge in answers

area size of switzerland

Switzerland · Area

15,940 sq miles (41,285 km²)
About twice the size of New Jersey
Knowledge graph serves NL fact answers
Knowledge graph serves carousel of information
I want to travel to NY 2 days before Thanksgiving, staying for a week.

Okay, booking a flight to JFK from November 20 to November 27. Where will you be flying from?

From San Francisco, and also non-stop in first class.

Got it, I’ve found some flights for you ...

How about leaving in the afternoon?
Part II: Acquiring Knowledge in the Wild

Benjamin Han
Principal Machine Learning & Data Scientist, Satori Group, Microsoft AI+R
diha@microsoft.com
Goals

• Identify the five pillars of a high-quality Knowledge Acquisition system
• Survey: a whirlwind tour of the proposed approaches
• With our bias and limitation
Knowledge Graph (KG)

• What is a Knowledge Graph? [Paulheim 2016]
  • KG describes real-world entities and their relations, organized in a graph.
  • Possible classes and relations are defined by schemas.
  • Focus on instance aspect of knowledge (A-Box in Description Logic), not the schema aspect (T-Box in DL).

“If someone won a Nobel Prize in Physics, he must be a physicist.”
Knowledge Acquisition (KA) in the Wild

- Heterogeneous sources/formats/modalities.
- Different domains of knowledge.
Knowledge Acquisition (KA) in the Wild

• Hard to ascertain veracity.
• Constantly changing.
• Training data is hard to come by.
Five Pillars of High-Quality KA for KG

• Wide Coverage
• High precision
• Verifiable knowledge
• More efficient human intervention
• High system maintainability
Wide Coverage

• Knowledge can come from many sources and in many forms
  • Structured sources
    • Relational databases
    • Feeds
    • Catalogues, directories etc
  • Unstructured sources
Unstructured Sources: Web Pages

• Web wrappers [Ferrara+ 2014]
  • Procedures for extracting user-designated data from web resources to structured form.
  • Major approaches
    • Rule-based: regular expressions, wrapper programming languages etc.
    • Tree-based: segment DOM in to data regions, then extract with partial alignment.
    • Machine-learning-based.
Degradation of Web Content Extractors

• Web content extractors degrade over time [Weninger+ 2015]
  • Algorithms reflected the state of web at the time.
  • Use of JavaScript and CSS made static HTML much less reliable to extract from.
  • Future: extraction should target visual rendering.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Largest Size Increase (LSI)</td>
<td>[16] 2001</td>
</tr>
<tr>
<td>Link Quota Filter</td>
<td>[21] 2005</td>
</tr>
<tr>
<td>K-Feature Extractor (KFE)</td>
<td>[10] 2005</td>
</tr>
<tr>
<td>Content Code Blurring (CCB)</td>
<td>[14] 2008</td>
</tr>
<tr>
<td>RoadRunner* (RR)</td>
<td>[7] 2008</td>
</tr>
<tr>
<td>Content Extraction via Tag Ratios (CETR)</td>
<td>[31] 2010</td>
</tr>
<tr>
<td>BoilerPipe</td>
<td>[17] 2010</td>
</tr>
<tr>
<td>Eatih</td>
<td>[27] 2015</td>
</tr>
</tbody>
</table>

TABLE IV: Content extraction algorithms, with their citation and publication date. * RoadRunner is a wrapper induction algorithm; all others are heuristic methods.

FIG. 3: Mean average F$_1$ measure per cohort over each lustrum.
• News and forums
  • Continuing from the MUC and ACE, the most important evaluation is TAC KBP (Knowledge Base Population) organized by NIST. [Getman+ 2017]
    • In 2017 five trilingual tracks were offered: Cold Start KB construction, Entity Discovery & Linking, Slot Filling (relation extraction), Event, and Belief and Sentiment.
      • Cold Start KB: builds a knowledge base from scratch using a given document collection and a predefined KB schema.
      • KB schema: entities, Slot filler relations (finding values for pre-defined attributes), event nuggets and arguments, and sentiments.
    • Datasets include newswire and discussion forums, in English, Chinese and Spanish.
Unstructured Sources: Texts – 1b

- TACKBP 2017 CSKB best system: Tinkerbell [Al-Badrashiny+ 2017]
  - First end-to-end trilingual system combining multiple building blocks from member institutions.

<table>
<thead>
<tr>
<th>slot types</th>
<th>#justifications</th>
<th>TinkerBell</th>
<th>Human</th>
<th>% Human</th>
</tr>
</thead>
<tbody>
<tr>
<td>all</td>
<td>3</td>
<td>7.56%</td>
<td>47.1%</td>
<td>16.1%</td>
</tr>
<tr>
<td>all</td>
<td>1</td>
<td>13.32%</td>
<td>59.77%</td>
<td>22.3%</td>
</tr>
<tr>
<td>SF</td>
<td>3</td>
<td>11.43%</td>
<td>40.97%</td>
<td>27.9%</td>
</tr>
<tr>
<td>SF</td>
<td>1</td>
<td>17.30%</td>
<td>41.53%</td>
<td>41.7%</td>
</tr>
</tbody>
</table>
Unstructured Sources: Texts - 2

- Emails & calendars: What can we learn from them?
  - Personal/professional information about people: person entity linking in emails [Gao+ 2017]
  - Information about organization mentions [Gao+ 2016]
  - Linking meeting mentions from emails to calendars [Gao+ 2018]
  - Finding “topics” through clustering and expertise [Tang+ 2014]
  - Extracting problem solving traces in professional emails [Francois+ 2015]
Unstructured Sources: Texts - 3

- Social media: what can we learn from them?
  - Twitter text normalization and named entity recognition [Baldwin+ 2015]
    - Two shared tasks held in 2015

<table>
<thead>
<tr>
<th>Team name</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
<th>Method highlights</th>
</tr>
</thead>
<tbody>
<tr>
<td>NCSU.SAS.NING</td>
<td>0.9061</td>
<td>0.7865</td>
<td>0.8421</td>
<td>Random Forest</td>
</tr>
<tr>
<td>NCSU.SAS.WOOKHEE</td>
<td>0.9136</td>
<td>0.7398</td>
<td>0.8175</td>
<td>Lexicon + LSTM</td>
</tr>
<tr>
<td>NCSU.SAS.SAM</td>
<td>0.9012</td>
<td>0.7437</td>
<td>0.8149</td>
<td>ANN</td>
</tr>
<tr>
<td>IITP</td>
<td>0.9026</td>
<td>0.7191</td>
<td>0.8005</td>
<td>CRF + Rule</td>
</tr>
<tr>
<td>DCU-ADAPT</td>
<td>0.8190</td>
<td>0.5599</td>
<td>0.6887</td>
<td>Generalized Perceptron</td>
</tr>
<tr>
<td>LYSGROUP</td>
<td>0.4640</td>
<td>0.6281</td>
<td>0.5341</td>
<td>Spanish Normalization Adaption</td>
</tr>
</tbody>
</table>

Table 3: Results of the constrained systems for the lexical normalization shared task

<table>
<thead>
<tr>
<th>Team name</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
<th>Method highlights</th>
</tr>
</thead>
<tbody>
<tr>
<td>IHS.RD</td>
<td>0.8469</td>
<td>0.8083</td>
<td>0.8272</td>
<td>Lexicon + CRF + DidYouMean</td>
</tr>
<tr>
<td>USZEGED</td>
<td>0.8606</td>
<td>0.7564</td>
<td>0.8052</td>
<td>CRF + n-gram[ s]</td>
</tr>
<tr>
<td>BKL</td>
<td>0.7743</td>
<td>0.7416</td>
<td>0.7571</td>
<td>Lexicon + Rule + Ranker</td>
</tr>
<tr>
<td>GIGO</td>
<td>0.7593</td>
<td>0.6963</td>
<td>0.7264</td>
<td>N/A</td>
</tr>
<tr>
<td>LYSGROUP</td>
<td>0.4592</td>
<td>0.6296</td>
<td>0.5310</td>
<td>Spanish Normalization Adaption</td>
</tr>
</tbody>
</table>

Table 4: Results of the unconstrained systems for the lexical normalization shared task

<table>
<thead>
<tr>
<th>POS</th>
<th>Orthographic</th>
<th>Gazetteers</th>
<th>Brown clustering</th>
<th>Word embedding</th>
<th>ML</th>
</tr>
</thead>
<tbody>
<tr>
<td>BASELINE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hallym</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>CRFSuite</td>
</tr>
<tr>
<td>ittp</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>CRF++</td>
</tr>
<tr>
<td>lattice</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>CRF++</td>
</tr>
<tr>
<td>multimedialab</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>FFNN</td>
</tr>
<tr>
<td>NLANGP</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>CRF++</td>
</tr>
<tr>
<td>mc</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>semi-Markov MIRA</td>
</tr>
<tr>
<td>basis</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>entity linking</td>
</tr>
<tr>
<td>UNLF</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>CRF++</td>
</tr>
</tbody>
</table>

Table 7: Features and machine learning approach taken by each team.

<table>
<thead>
<tr>
<th>Team name</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>NLANGP</td>
<td>63.62</td>
<td>43.12</td>
<td>51.40</td>
<td>NLANGP</td>
<td>67.74</td>
<td>54.31</td>
</tr>
<tr>
<td>mc</td>
<td>53.24</td>
<td>38.58</td>
<td>44.74</td>
<td>mc</td>
<td>63.81</td>
<td>56.28</td>
</tr>
<tr>
<td>multimedialab</td>
<td>49.52</td>
<td>39.18</td>
<td>43.75</td>
<td>multimedialab</td>
<td>62.93</td>
<td>55.22</td>
</tr>
<tr>
<td>USFD</td>
<td>60.72</td>
<td>39.64</td>
<td>42.46</td>
<td>USFD</td>
<td>61.23</td>
<td>54.61</td>
</tr>
<tr>
<td>ittp</td>
<td>60.68</td>
<td>29.65</td>
<td>39.84</td>
<td>ittp</td>
<td>63.43</td>
<td>51.44</td>
</tr>
<tr>
<td>Hallym</td>
<td>39.59</td>
<td>35.10</td>
<td>37.21</td>
<td>Hallym</td>
<td>58.36</td>
<td>48.5</td>
</tr>
<tr>
<td>lattice</td>
<td>55.17</td>
<td>9.68</td>
<td>16.47</td>
<td>lattice</td>
<td>58.42</td>
<td>25.72</td>
</tr>
<tr>
<td>BASELINE</td>
<td>35.56</td>
<td>29.05</td>
<td>31.97</td>
<td>BASELINE</td>
<td>53.86</td>
<td>46.44</td>
</tr>
</tbody>
</table>

Table 8: Results segmenting and categorizing entities into 10 types.

Table 9: Results on segmentation only (no types).
Unstructured Sources: Texts - 4

- Extracting events and attributes [Wang+ 2015]
- Extracting user profiles [Jiwei+ 2015]
- Extracting computer security events [Ritter+ 2015]
- Extracting emerging entities using seeds [Brambilla+ 2017]
- Quantitative Information Extraction From Social Data [Alonso & Sellam 2018]

<table>
<thead>
<tr>
<th>Victim</th>
<th>Date</th>
<th>Category</th>
<th>Sample Tweet</th>
</tr>
</thead>
<tbody>
<tr>
<td>namecheap</td>
<td>Feb-20-2014</td>
<td>DDoS</td>
<td>My site was down due to a DDoS attack on NameCheap’s DNS server. These are last page hits man...</td>
</tr>
<tr>
<td>bitcoin</td>
<td>Feb-12-2014</td>
<td>DDoS</td>
<td>Bitcoin value dramatically drops as massive #DDOS attack is waged on #Bitcoin <a href="http://t.co/YdogyOGmhw">http://t.co/YdogyOGmhw</a></td>
</tr>
<tr>
<td>europe</td>
<td>Feb-20-2014</td>
<td>DDoS</td>
<td>Record-breaking DDoS attack in Europe hits 400Gbps.</td>
</tr>
<tr>
<td>barcelona</td>
<td>Feb-18-2014</td>
<td>Account Hijacking</td>
<td>Lmao, the official Barcelona account has been hacked.</td>
</tr>
<tr>
<td>adam</td>
<td>Feb-16-2014</td>
<td>Account Hijacking</td>
<td>dadam lambert You’ve been hacked Adam! Argh!</td>
</tr>
<tr>
<td>dubai</td>
<td>Feb-09-2014</td>
<td>Account Hijacking</td>
<td>Dubai police twitter account just got hacked!</td>
</tr>
<tr>
<td>maryland</td>
<td>Feb-20-2014</td>
<td>Data Breach</td>
<td>SSNs Compromised In University of Maryland Data Breach: <a href="https://t.co/jB9VjIC4dw">https://t.co/jB9VjIC4dw</a></td>
</tr>
<tr>
<td>kickstarter</td>
<td>Feb-15-2014</td>
<td>Data Breach</td>
<td>I suspect my card was compromised because of the Kickstarter breach. It’s a card I don’t use often but have used for things like that.</td>
</tr>
<tr>
<td>tesco</td>
<td>Feb-14-2014</td>
<td>Data Breach</td>
<td>#directhex @Tesco thanks to the data breach yesterday it’s clear no-one in Tesco does their sysadmin housekeeping!</td>
</tr>
</tbody>
</table>

Table 6: Example high-confidence events extracted using our system.
Unstructured Sources: Texts - 5

• Catalog: Product Knowledge Graph [Dong 2017]
  • No major sources to curate product knowledge from
  • Wikipedia does not help too much
  • A lot of structured data buried in text descriptions in Catalog
  • Retailers gaming with the system so noisy data
  • Large # of products and categories, changing everyday
  • Many entities are not named
Unstructured Sources: Other Modalities

• Speech, images, video
  • ImageCLEF competition [Ionescu+ 2017]
    • Lifelogging data retrieval and summarization; medical images to textual description/classification; discover unknown info from Earth observation images
  • TACKBP 2018 – Streaming Multimedia Knowledge Base Population [web]
    • Evaluate systems for extracting and aggregating knowledge from heterogeneous sources such as multilingual multimedia sources including text, speech, images, videos, and pdf files, and developing hypotheses interpreting the input.
Coverage: Extracting Entities - 1

• *Joint* entity and relation extraction
  • Incremental joint extraction [Li & Ji 2014]
  • With a novel tagging scheme [Zheng+ 2017]
  • With knowledge bases [Ren+ 2016]
Coverage: Extracting Entities - 2

- **[Ren+ 2016]** Framework CoType
  - Produce candidate entity mentions using POS then candidate relation mentions; generate training set using the labels from KB
  - Jointly embed relation and entity mentions, text features and labels
  - Estimate type labels for test relation mentions and their argument mentions

![Figure 2: Framework Overview of CoTYPE.](image-url)
Coverage/Precision: Entity Linking

- Disastrous result if linking failed, even with perfect extraction
• The best evidence for entity disambiguation is provided by the set of co-occurring entities
  • Extract and disambiguate jointly all entities in a target document
  • Employ both observable attributes (known values, contexts) and latent attributes (e.g. entity relationships, topics)
• Syntax and local context are important - one-sense-per-discourse does not hold
  • Employ both whole-document and local context features

[Cucerzan 2007]
Disambiguation - Intuition
Each entity has multiple vectorial representations

Find the entity assignment that maximizes the similarity between the observable representations and the document context $d$ and between the latent representations of the entities in the assignment.

[Cucerzan 2007]
### NEMO - 3

#### NIST/LDC Evaluations

<table>
<thead>
<tr>
<th>Accuracy</th>
<th>NEMO system (2014)</th>
<th>best result in the TAC evaluation</th>
</tr>
</thead>
<tbody>
<tr>
<td>TAC 2011 test set</td>
<td>89.3 %</td>
<td>86.8% (MSR/NEMO)</td>
</tr>
<tr>
<td>TAC 2012 test set</td>
<td>80.4 %</td>
<td>76.2% (MSR/NEMO)</td>
</tr>
<tr>
<td>TAC 2013 test set</td>
<td>85.2 %</td>
<td>83.2% (MSR/NEMO)</td>
</tr>
<tr>
<td>TAC 2014 test set</td>
<td>86.8 %</td>
<td>86.8% (MSR/NEMO)</td>
</tr>
</tbody>
</table>

Google-Microsoft-Yahoo ERD Challenge (best participating system) [Carmel+ 2014]

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>ERD 2014 train set</td>
<td>83.7%</td>
<td>72.6%</td>
<td>0.778</td>
</tr>
<tr>
<td>ERD 2014 test set</td>
<td>83.3%</td>
<td>69.9%</td>
<td>0.760</td>
</tr>
</tbody>
</table>
Coverage: Extracting Relations

- Predicting relations based on existing ones using Tensor NN [Socher+ 2013]
- Universal Schemas [Riedel+ 2013]
- Type-constrained learning in KG [Krompaß+ 2015]
- Association rules mining [Kolthoff & Dutta 2015]
- Embedding-based methods [Zhao+ 2015] [Bishan+ 2015] [Toutanova 2015] [Goyal & Ferrara 2017] [Shen+ 2016]
- Reinforcement Learning [Feng+ 2018]
- Open IE [Cui+ 2018]
- Web search [West+ 2014]
- Survey of relational ML for Knowledge Graphs [Nickel+ 2015]
Embedding Methods for KB Completion - 1

• Each entity in a KB is represented by an $R^d$ vector
• Predict whether $(e_1, r, e_2)$ is true by $f_r(v_{e_1}, v_{e_2})$
• Work on KB embedding
  • Tensor decomposition
    • RESCAL [Nickel+ ICML-11], TRESICAL [Chang+ EMNLP-14]
  • Neural networks
    • SME [Bordes+ AISTATS-12], NTN [Socher+ NIPS-13], TransE [Bordes+ NIPS-13]
Embedding Methods for KB Completion - 2

- Objective:

\[ \frac{1}{2} \left( \sum_k \| x_k - A R_k A^T \|_F^2 \right) + \frac{1}{2} \left( \| A \|_F^2 + \sum_k \| R_k \|_F^2 \right) \]

Reconstruction Error  \hspace{2cm}  Regularization

\[ f_{\text{BornIn}}(\text{Obama, Hawaii}) = A_{\text{Obama}} \cdot R_{\text{BornIn}} A^T_{\text{Hawaii}} \]

RESCAL \text{ [Nickel + ICML-11]}
Embedding Methods for KB Completion - 3

• Typed tensor decomposition (TRESCAL) [Chang+ EMNLP-14]
  • Only legitimate entities are included in the loss
  • Faster model training time (4.6x speedup), highly scalable, higher accuracy
  • Reconstruction error: \( \frac{1}{2} \sum_k \| X_k - \mathbf{A} \mathbf{R}_k \mathbf{A}^T \|_F^2 \)
  • Training: Alternating Least-Squares (ALS)

Entity Retrieval \((e_i, r_k, ?)\)
Relation Retrieval \((e_i, ?, e_j)\)

Mean Average Precision

\[ \frac{1}{n} \sum_{i=1}^{n} \text{MAP}(e_i) \]
Relation Extraction from Semi-Structured Sources

- Wikipedia tables [Muñoz+ 2013].
- Wikipedia list pages [Paulheim & Ponzetto 2013]
- Web tables [Ritze+ 2015]
- Microsoft Kable: Large scale unsupervised template learning
Verifiable Knowledge - 1

• Not everything accurately extracted is fact
  • Knowledge-based Trust [Dong+ 2015]

• Many recent efforts on assessing truth and finding supports
  • Multilingual answer validation [Rodrigo+ 2009] [Kobayashi+ 2017]
  • FactChecker [Nakashole & Mitchell 2014]
  • PolitiFact [Vlachos & Riedel 2014], [Wang 2017]
  • Fake News challenge [Pomerleau & Rao 2017]
  • Fake news detection via crowd signals [Tschiatschek+ 2018]
  • Fact Verification competition [Thorne+ 2018]
Verifiable Knowledge - 2

• Fact Verification competition (FEVER) [Thorne+ 2018]
  • Goal: given a claim
    • Label claim SUPPORTS, REFUTES, or NOT-ENOUGH-INFO
    • For the first two classes, select relevant sentences from Wikipedia intro sections.
  • Largest annotated fact sets
    • 185,445 annotated claims.
    • Claims generated by mutating Wikipedia sentences: paraphrasing, negation, substitution of entity/relation, generalize/specialize claims.

Claim: The Rodney King riots took place in the most populous county in the USA.

[wiki/Los Angeles Riots]
The 1992 Los Angeles riots, also known as the Rodney King riots were a series of riots, lootings, arsons, and civil disturbances that occurred in Los Angeles County, California in April and May 1992.

[wiki/Los Angeles County]
Los Angeles County, officially the County of Los Angeles, is the most populous county in the USA.

Verdict: Supported
Verifiable Knowledge - 3

• FEVER baseline – sentence classification  

Basic idea: align parts of the text in sentences a and b and then aggregate info to predict the label

Example

Bob is in his room, but because of the thunder and lightning outside, he cannot sleep.

Bob is awake

It is sunny outside

[Thorne+ 2018]
Verifiable Knowledge - 4

• Decomposable Attention model (DA) [Thorne+ 2018]
  • Attend
    • Create soft alignment matrix to produce aligned subphrases between a and b
    • Alignments are learned using feedforward model F
  • Compare
    • Score aligned subphrases using a function G
    • G is a feedforward model which produces comparison vectors
  • Aggregate
    • Sum over comparison vectors and produce final score using feedforward model H

|     | 20   | jamesthorne | FEVER Baseline | 0.1826 | 0.4884 | 0.2745 |
Verifiable Knowledge - 4

<table>
<thead>
<tr>
<th>#</th>
<th>User</th>
<th>Team Name</th>
<th>Evidence F1</th>
<th>Label Accuracy</th>
<th>FEVER Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>chaonan99</td>
<td>UNC-NLP</td>
<td>0.5296</td>
<td>0.6821</td>
<td>0.6421</td>
</tr>
<tr>
<td>2</td>
<td>tyoneda</td>
<td>UCL Machine Reading Group</td>
<td>0.3497</td>
<td>0.6762</td>
<td>0.6252</td>
</tr>
<tr>
<td>3</td>
<td>littsler</td>
<td>Athene UKP TU Darmstadt</td>
<td>0.3697</td>
<td>0.6546</td>
<td>0.6158</td>
</tr>
<tr>
<td>4</td>
<td>papelo</td>
<td></td>
<td>0.6485</td>
<td>0.6108</td>
<td>0.5736</td>
</tr>
<tr>
<td>5</td>
<td>chidey</td>
<td></td>
<td>0.2969</td>
<td>0.5972</td>
<td>0.4994</td>
</tr>
<tr>
<td>6</td>
<td>Tuhin</td>
<td>ColumbiaNLP</td>
<td>0.3533</td>
<td>0.5745</td>
<td>0.4906</td>
</tr>
<tr>
<td>7</td>
<td>nanjiang</td>
<td>The Ohio State University</td>
<td>0.5853</td>
<td>0.5012</td>
<td>0.4342</td>
</tr>
<tr>
<td>8</td>
<td>wotto</td>
<td>gesis cologne</td>
<td>0.1960</td>
<td>0.5415</td>
<td>0.4077</td>
</tr>
<tr>
<td>9</td>
<td>tomoki</td>
<td>Fujixerox</td>
<td>0.1649</td>
<td>0.4713</td>
<td>0.3881</td>
</tr>
<tr>
<td>10</td>
<td>nayeon7lee</td>
<td></td>
<td>0.4912</td>
<td>0.5125</td>
<td>0.3859</td>
</tr>
</tbody>
</table>
Verifiable Knowledge - 5

- Techniques rooted in core NLP fields
  - Textual Entailment [Dagan+ 2006]
  - Natural language inference [Angeli & Manning 2014]
More Efficient Human Intervention - 1

• Slot tagging using search click logs [Kim & Sarikaya 2015]
  • Slot tagging for queries: “when is the new bill murray movie release date?”
  • Weakly supervised: project labels from structured data found in click logs.
More Efficient Human Intervention - 2

- [Kim & Sarikaya 2015] CRF variants to learn from partially labeled sequences

\[
p_{\theta}(y|x) = \frac{\exp(\theta^T \phi(x, y))}{\sum_{y' \in \mathcal{Y}(x)} \exp(\theta^T \phi(x, y'))}
\]

\[
\theta^* = \arg \max_{\theta \in \mathbb{R}^d} \sum_{i=1}^N \log p_{\theta}(y^{(i)}|x^{(i)}) - \frac{\lambda}{2} ||\theta||^2
\]

Initialization:
- Cluster unlabeled data
- Train fully supervised HUCRF with cluster labels
- Keep learned \( \theta \) (between input \( x \) and hidden \( z \)) and start task-specific training

<table>
<thead>
<tr>
<th>Domains</th>
<th>games</th>
<th>music</th>
<th>movies</th>
<th>AVG.</th>
</tr>
</thead>
<tbody>
<tr>
<td>CRF</td>
<td>74.21</td>
<td>37.13</td>
<td>68.58</td>
<td>59.97</td>
</tr>
<tr>
<td>POCRF</td>
<td>77.23</td>
<td>44.55</td>
<td>76.89</td>
<td>66.22</td>
</tr>
<tr>
<td>POHCRF</td>
<td>78.93</td>
<td>46.81</td>
<td>76.46</td>
<td>67.40</td>
</tr>
<tr>
<td>POHCRF+</td>
<td><strong>79.28</strong></td>
<td><strong>47.35</strong></td>
<td><strong>78.33</strong></td>
<td><strong>68.32</strong></td>
</tr>
</tbody>
</table>

F1 scores
More Efficient Human Intervention - 3

• **Distant supervision (DS)** [Mintz+ 2009] [Gerber & Ngomo 2011] [Gerber+ 2013]
  • Enhance DS with dynamic transition matrix **[Luo+ 2017]**
    • Problem of DS: label noise
      • Triple <Donald Trump, born-in, New York> picked “Donald Trump worked in New York City” as positive example.
      • Solution: model noise via a transition matrix $T_{ij}$ indicating the conditional probability for the input sentence to be labeled as relation $j$ by DS, given $i$ as the true relation.

Transition matrix: $T_{ij} = \frac{\exp(w_{ij}^T x_n + b)}{\sum_{j=1}^{|C|} \exp(w_{ij}^T x_n + b)}$

Output: $o = T^T \cdot p$ (p is prediction)
More Efficient Human Intervention - 4

- [Luo+ 2017]
  - Training can be done on sentence level or bag level [Carbonneau+ 2016]
  - How to train transition matrix w/o humans? Curriculum learning. [Bengio+ 2009]
    - \( \text{trace}(T) \): the larger (more similar to identity matrix) the lower the noise – regularize \( \text{trace}(T) \).
    - Training: initially set \( \alpha, \beta = 1 \) to learn \( p \) (prediction) only, then schedule to decrease \( \alpha, \beta \) to learn more about noise.

\[
L = \sum_{i=1}^{N} -((1 - \alpha) \log(o_{i,y_i}) + \alpha \log(p_{i,y_i})) - \beta \text{trace}(T^i)
\]
More Efficient Human Intervention - 5

- DS relation extraction from semi-structured web [Lockard+ 2018]
- Effective crowdsourcing [Chang+ 2017]
- More accessible ML tools [Yang+ 2018]
High Maintainability - 1

• "High Interest Credit Card of Technical Debt" [Sculley+ 2014]
  • Complex Models Erode Boundaries
    • CACG (changing anything changes everything)
    • Hidden feedback loops
    • Undeclared customers
  • Data Dependencies Cost More than Code Dependencies
    • Unstable data dependencies
    • Underutilized data dependencies
    • Difficult to do static analysis of data dependencies
    • Danger in creating error-correction models
  • System-level spaghetti
    • Glue code
    • Pipeline Jungles
    • Dead experiment codepaths
    • Configuration debt
  • Dealing with changing world
High Maintainability - 2

• Classifier error discovery through semantic data exploration [Chen+ 2018]

Figure 1. Overview of AnchorViz interface. The interface has Explore pane (A) that includes the visualization and Items pane (B) which shows a paginated grid of thumbnails of all currently visible items in the left pane. The visualization shows data points (C) within the outer circle (D) where anchors are positioned. The legend for data points (E) also acts as filters. Anchor repository (F) contains unused anchors. The navigator (G) shows which cluster the visualization is displaying in the current navigation stack. Clusters are represented as treemap-style squares (H).
Summary

• Majority of the approaches still relies on textual data
• Providing constant stream of high-quality training data with minimal human intervention is still the key
• Knowledge verification and correction will become even more important
• Model and system maintainability requires a fresh take over the traditional ways of dealing with software engineering tasks
Part III: Building Knowledge Graph

Mohamed Yakout
Principal Applied Science Manager, Satori Group, Microsoft AI+R
myakout@microsoft.com
What is a Knowledge Graph?

Graph: RDF Triples of (Subject, Predicate, Object)
Ontology Basics

• A complete, consistent, non-redundant, machine-readable representation of the world:
  • Allow data from various sources to be merged
  • Allow data to be shared across applications.

• Three elements: entities, properties, and types.
  • **Entities**: individuals, i.e. named objects in the world.
  • **Properties**: relationships between two entities or an entity and a **literal**, e.g. people.person.friends, people.person.employer, people.person.first_name, time.event.start_date, etc.
  • **Types**: sets or classes of entities:
    • **Primary entity types**: represent natural kinds or groupings, e.g. books, films, people, etc.
    • **Enumeration types**: Values that are standard but do not correspond to real objects in the world.
    • **Relationship types**: used to represent associations between more than two things, e.g. marriage (the people involved, when it started, where it began, etc.)
Satori Graph Build

Data Ingestion
- Selection of data sources
- Data Preparation
- Targeted fact extraction by NLP and entity linking

Match & Merge
- Matching Entity Contents
- Detection of Matched Entities
- Scaling & Improve Efficiency of Matching

Knowledge Refinement
- Knowledge Fusion
- Error Detection
- Fact Inference

Publishing & Serve
- Entity search API
- Ranking & Filtering by attributes
- Graph walk
- Semantic Linking & Join
Data Preparation

• Storing the data in a uniform manner.

• **Parsing**: locate, identify and isolate data elements

• **Data Transformation and Standardization**:
  • “44 West Fourth Street” or “44 West 4th St.”
  • 8 inches or 20 cm
  • July 28, 1999 or 07/28/1999 or 28/07/99

• Next, identify which fields to be compared.
Data Preparation

- Schema Matching
- Mapping to Microsoft Ontology

Schema Mapping and Management

- Schema mapping: Declarative language, versioned and managed mappings, validation of mapping with schema change tracking

Example mapping for Music data to Satori ontology

```xml
<ElementMap id='albumEntityPrimary.Album' elementName='Album' className='#Album@1.0'>
  <PropertyMaps>
    <ElementMap expression="music.album" elementName='type.object.type' />
    <ElementMap propertyPath='.Title' elementName='type.object.name' />
    <ElementMap propertyPath='/.ID/ZuneMediaId' elementName='type.object.key' />
    <ElementMap propertyPath='/.ReleaseDate' elementName='music.album.release_date' />
    <ElementMap propertyPath='/.Label' elementName='music.album.record_label' />
    <ElementMap propertyPath='/.Artists/Artist/Title' elementName='music.album.artist' />
    <ElementMap propertyPath='/.Tracks/Track/Title' elementName='music.album.track' multiplicity='MultiValued' />
    <ElementMap expression="SUM(./Tracks/Track/DurationSeconds)" elementName='music.album.length' />
    <ElementMap propertyPath='/.Genres/Genre/Genre' elementName='music.album.genre' multiplicity='MultiValued' />
  </PropertyMaps>
</ElementMap>
```

Closest approach in literature is Beaver: Jin et al, "Beaver: Towards a Declarative Schema Mapping" HILDA 2018
Satori Graph Build

Data Ingestion
- selection of data sources
- Data Preparation
- Targeted fact extraction by NLP and entity linking

Match & Merge
- Matching Entity Contents
- Detection of Matched Entities
- Scaling & Improve Efficiency of Matching

Knowledge Refinement
- Knowledge Fusion
- Error Detection
- Fact Inference

Publishing & Serve
- Entity search API
- Ranking & Filtering by attributes
- Graph walk
- Semantic Linking & Join
Entity Matching

- **Well known problem**: Identify and discover instances referring to the same real-world entity.

- **Objective**:
  - Data Enrichment
  - Improve Data Quality by identifying and removing duplicates
  - Supporting fact correctness by merging duplicate facts from multiple sources

- **Synonyms**: Entity Linking, Entity Resolution, Reference Reconciliation, Deduplication, Match/Merge, Merge/Purge
Integration means more information and enrichment.

Knowledge Evolution.
Entity Matching References

• Book / Survey Articles
  • Duplicate Record Detection [A. Elmagarid, P. Ipeirotis, V. Verykios, TKDE ‘07]
  • An Introduction to Duplicate Detection [F. Naumann, M. Herschel, M&P synthesis lectures 2010]
  • Evaluation of Entity Resolution Approached on Real-world Match Problems [H. Kopke, A. Thor, E. Rahm, PVLDB 2010]
  • A Survey of Indexing Techniques for Scalable Record Linkage and Deduplication [P. Christen TKDE ‘11]
  • Data Matching [P. Christen, Springer 2012]

• Tutorials
  • Record Linkage: Similarity measures and Algorithms [N. Koudas, S. Sarawagi, D. Srivatsava SIGMOD ‘06]
  • Data fusion—Resolving data conflicts for integration [X. Dong, F. Naumann VLDB ‘09]
  • Entity Resolution: Resolution: Theory, Practice Practice and Open Challenges Challenges [L. Getoor, A. Machanavajjhala VLDB ‘12]
  • Entity Resolution in the Web of Data: Tutorial [Kostas Stefanidis CIKM 2013]

• Systems
  • SecondString, http://secondstring.sourceforge.net/
  • Simmetrics: http://sourceforge.net/projects/simmetrics/
Data Quality Challenge


Missing Data
Data error due to IE tech or human errors
Abbreviations and truncation
Open Domain Entity Matching (Disambiguation Challenge)
EM Big Data Challenge

• **Larger Datasets**: Need Faster, Efficient, Parallel techniques.

• **Multi-Domain**: Need different matching methods and a technique to manage executions within and across domains

• **Linked, Connected and Relational data**: Need techniques to leverage the diversity of connections and representation.
Satori Graph Build

Data Ingestion
- selection of data sources
- Data Preparation
- Targeted fact extraction by NLP and entity linking

Match & Merge
- Matching Entity Contents
- Detection of Matched Entities
- Scaling & Improve Efficiency of Matching

Knowledge Refinement
- Knowledge Fusion
- Error Detection
- Fact Inference

Publishing & Serve
- Entity search API
- Ranking & Filtering by attributes
- Graph walk
- Semantic Linking & Join
Matching Entity Contents

• Matching Functions
  • Generic Functions
    • Character Based Functions
    • Token Based Functions
    • Phonetic Based Functions
    • Transformation Rule Based Matching Functions
    • Value-Set Matching Functions
  • Specific Functions
    • Numeric Matching Functions (Numbers, Dates, … etc)
    • Special Matching Functions (Zip codes, Phone Numbers, Address … etc)

Elmagarmid et al, “Duplicate Record Detection” TKDE 2007]
Satori Graph Build

Data Ingestion
- selection of data sources
- Data Preparation
- Targeted fact extraction by NLP and entity linking

Match & Merge
- Matching Entity Contents
- **Detection of Matched Entities**
- Scaling & Improve Efficiency of Matching

Knowledge Refinement
- Knowledge Fusion
- Error Detection
- Fact Inference

Publishing & Serve
- Entity search API
- Ranking & Filtering by attributes
- Graph walk
- Semantic Linking & Join
Detection of Matched Entities

- Probabilistic Matching Models
  - Supervised and Semi-supervised Learning
  - Unsupervised learning
  - Active Learning Based

- Distance Based
  - Threshold
  - Neighborhood exploration

- Declarative Matching Rules and Constraints
  - Disjunction of conjunction
  - Constraint base clustering

- Collective Resolution in Linked Data
  - Similarity signals propagation
  - Entity similarity based on connections
Detection of Matched Entities

• **Probabilistic Matching Models**
  • Supervised and Semi-supervised Learning
  • Unsupervised learning
  • Active Learning Based

• Distance Based
  • Threshold
  • Neighborhood exploration

• Declarative Matching Rules and Constraints
  • Disjunction of conjunction
  • Constraint base clustering

• Collective Resolution in Linked Data
  • Similarity signals propagation
  • Entity similarity based on connections
Detection of Matched Entities

- Compute similarity vector
- Classify the vectors as Match and UnMatch.

<table>
<thead>
<tr>
<th>Name</th>
<th>John Smith</th>
</tr>
</thead>
<tbody>
<tr>
<td>Profession</td>
<td>Software Eng.</td>
</tr>
<tr>
<td>Address</td>
<td>Seattle</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Name</th>
<th>Johan Smith</th>
</tr>
</thead>
<tbody>
<tr>
<td>Profession</td>
<td>Software Dev Eng.</td>
</tr>
<tr>
<td>Address</td>
<td>Seattle</td>
</tr>
</tbody>
</table>

The similarity vector <0.8, 0.7, 1.0>
Detection of Matched Entities: Probabilistic Matching Models

• Supervised and Semi-supervised Learning

  • Map the similarity vector to two classes (M, U)

  •
Detection of Matched Entities: Probabilistic Matching Models

- Supervised and Semi-supervised Learning
  - Map the similarity vector to two classes (M, U)
  - Later on a rejection or uncertain rejoin is considered (M, R, U)
  - Rely on the existence of training data, pair of records pre-labeled match or not. Do we have that?!!
Detection of Matched Entities: Probabilistic Matching Models

• Pairs Sampling for training
  • Random Sample
    • Most of space contains non-matched pairs
  • Sample from blocks
    • Apply blocking
    • Random Sample a set of blocks
    • Get pairs from the randomly sampled blocks
  • Stratified Sample
    • Cluster the similarity vectors
    • Sample from clusters
Detection of Matched Entities: Probabilistic Matching Models

• Active Learning

• Train an initial ML model by an initial small sample
• While (user is not happy with predictions)
  • Foreach Pair \( p \) in all pairs
    • Apply the model on \( p \)
    • Get the prediction probability and compute uncertainty
  • Sort all pairs based on uncertainty
  • Display pairs with the highest uncertainty first to user for labeling
  • Re-train the model.
Detection of Matched Entities: Probabilistic Matching Models

• Active Learning

• Train an initial ML model by an **initial small sample**

• While (user is not happy with predictions)
  • Foreach Pair \( p \) in all pairs
    • Apply the model on \( p \)
    • Get the prediction probability and compute uncertainty
  • Sort all pairs based on uncertainty
  • Display pairs with the highest uncertainty first to user for labeling
  • Re-train the model.
Detection of Matched Entities: Probabilistic Matching Models

- **Effective Active Learning for Entity Matching**
  - Better control on the space of similarities.
  - Clustering for all vectors
  - Offline sample from clusters host locally
  - **Active Learning Guided by the Clusters** through:
    - Focus on clusters with high uncertainty
    - Cover clusters with less training samples
    - From a cluster, sampling positive uncertain cases improves precision
    - From a cluster, sampling negative uncertain cases improves recall.
  - Uncertainty can be computed from the entropy of model’s probability.
Detection of Matched Entities: Probabilistic Matching Models

- **Effective Active Learning for Entity Matching**

- Better control on the space of similarities.
- Clustering for all vectors
- Offline sample from clusters host locally
- **Active Learning Guided by the Clusters** through:
  - Focus on clusters with high uncertainty
  - Cover clusters with less training samples
  - From a cluster, sampling positive uncertain cases improves precision
  - From a cluster, sampling negative uncertain cases improves recall.
- Uncertainty can be computed from the entropy of model’s probability.
Detection of Matched Entities: Probabilistic Matching Models

- **Effective Active Learning for Entity Matching**
  - Better control on the space of similarities.
  - Clustering for all vectors
  - Offline sample from clusters host locally
  - **Active Learning Guided by the Clusters** through:
    - Focus on clusters with high uncertainty
    - Cover clusters with less training samples
    - From a cluster, sampling positive uncertain cases improves precision
    - From a cluster, sampling negative uncertain cases improves recall.
  - Uncertainty can be computed from the entropy of model's probability.
Detection of Matched Entities: Probabilistic Matching Models

• **Effective Active Learning for Entity Matching**
  
  • Better control on the space of similarities.
  
  • Clustering for all vectors
  
  • Offline sample from clusters host locally
  
  • **Active Learning Guided by the Clusters** through:
    • Focus on clusters with high uncertainty
    • Cover clusters with less training samples
    • From a cluster, sampling positive uncertain cases improves precision
    • From a cluster, sampling negative uncertain cases improves recall.
  
  • Uncertainty can be computed from the entropy of model’s probability.
Detection of Matched Entities

- Probabilistic Matching Models
  - Supervised and Semi-supervised Learning
  - Unsupervised learning
  - Active Learning Based

- **Distance Based**
  - Threshold
  - Neighborhood exploration

- Declarative Matching Rules and Constraints
  - Disjunction of conjunction
  - Constraint base clustering

- Collective Resolution in Linked Data
  - Similarity signals propagation
  - Entity similarity based on connections
Detection of Matched Entities: Distance Based

• Threshold
  • If \( w_1 \text{sim}(name_1, name_2) + w_2 \text{sim}(address_1, address_2) > t \).
    Then it a match.

• Neighborhood exploration
  • Matches are “closer” to each other than to others
    • A “Compact Set” criteria
  • The local neighborhood of matched entities is sparse
    • A “Sparse Neighborhood” criteria
  • Requires an overall matching or distance function for two entities

Chaudhuri et al., "Robust Identification of Fuzzy Duplicates", ICDE 2005
Detection of Matched Entities

• Probabilistic Matching Models
  • Supervised and Semi-supervised Learning
  • Unsupervised learning
  • Active Learning Based

• Distance Based
  • Threshold
  • Neighborhood exploration

• **Declarative Matching Rules and Constraints**
  • Disjunction of conjunction
  • Constraint base clustering

• Collective Resolution in Linked Data
  • Similarity signals propagation
  • Entity similarity based on connections
Detection of Matched Entities: Matching Rules and Constraints

- Disjunction of Conjunction (Simple)
  - \( \text{Match(movie_name) AND Match(release_date)} \)
  - \( \text{OR Match(movie_name) AND Match(director) } \rightarrow \text{ Match} \)

- Constraints based clustering and matching (e.g., Dedupalog)
  - Encoding of rules and constrains and then cluster entities to satisfy hard constraints and minimize soft rules violations. Example:
    - No researcher has published more than five AAAI papers in a year
    - If two citations match, then their authors will be matched in order
    - Papers with similar titles should likely be clustered together

- The framework is domain independent. But how realistic is this to compile these rules?

Arasu et al, "Large-Scale Deduplication with Constraints Using Dedupalog", ICDE 2009
Detection of Matched Entities

• Probabilistic Matching Models
  • Supervised and Semi-supervised Learning
  • Unsupervised learning
  • Active Learning Based

• Distance Based
  • Threshold
  • Neighborhood exploration

• Declarative Matching Rules and Constraints
  • Disjunction of conjunction
  • Constraint base clustering

• Collective Resolution in Linked Data
  • Similarity signals propagation
  • Entity similarity based on connections
Detection of Matched Entities: Collective Resolution in Linked Data
Graph Data Model and Conflation

Similarity Signals Propagation
Graph Data Model and Conflation
Graph Data Model and Conflation
Graph Data Model and Conflation
Graph Data Model and Conflation
Graph Data Model and Conflation
Detection of Matched Entities: Collective Resolution in Linked Data

• **Entity similarity based on connections**
  
  • **Adamic/Adar Measure**: Two nodes are more similar if they share more items that are overall less frequent

  $$sim(a, b) = \sum_{i \in \text{shared}} \frac{1}{\log(freq(i))}$$

  • **SimRank**: Two objects are similar if they are related to similar objects

  $$sim(a, b) = \frac{C}{|I(a)||I(b)|} \sum_{i=1}^{|I(a)|} \sum_{j=1}^{|I(b)|} sim(I_i(a), I_j(b))$$

  • **Katz Score**: Two objects are similar if they are connected by shorter paths

  $$sim(a, b) = \sum_{l=1}^{\infty} \beta^l \cdot \text{paths}^{(l)}(a, b)$$
Satori Graph Build

Data Ingestion
- selection of data sources
- Data Preparation
- Targeted fact extraction by NLP and entity linking

Match & Merge
- Matching Entity Contents
- Detection of Matched Entities
- Scaling & Improve Efficiency of Matching

Knowledge Refinement
- Knowledge Fusion
- Error Detection
- Fact Inference

Publishing & Serve
- Entity search API
- Ranking & Filtering by attributes
- Graph walk
- Semantic Linking & Join
Improve Efficiency of Matching

• Matching two data sources each with 1 M entities
• 1M x 1M with an entity pair comparison time of 5 µs
• 160 years
• 300K machines to finish in 5 hrs

• **Solution:** Blocking or Indexing

  • Efficiency or Reduction Ratio = \( \frac{|\text{compared pairs}|}{m \times n} \)

  • Recall or pairs completeness = \( \frac{|\text{True Matches compared}|}{|\text{All existing true matches}|} \)

---

<table>
<thead>
<tr>
<th>Entity</th>
<th>Hashes</th>
</tr>
</thead>
<tbody>
<tr>
<td>E1</td>
<td>h1, h2</td>
</tr>
<tr>
<td>E2</td>
<td>h1, h3, h4</td>
</tr>
<tr>
<td>E3</td>
<td>h3</td>
</tr>
<tr>
<td>E4</td>
<td>h4</td>
</tr>
</tbody>
</table>

---

<table>
<thead>
<tr>
<th>Key</th>
<th>Post List</th>
</tr>
</thead>
<tbody>
<tr>
<td>h1</td>
<td>E1, E2</td>
</tr>
<tr>
<td>h2</td>
<td>E1</td>
</tr>
<tr>
<td>h3</td>
<td>E1, E2, E3</td>
</tr>
<tr>
<td>h4</td>
<td>E2, E4</td>
</tr>
</tbody>
</table>
Improve Efficiency of Matching

• Reduce the number of entities comparisons (Indexing or Blocking)
  1. Identify blocking attributes
  2. Hashing Functions
  3. Retrieval of pairs

<table>
<thead>
<tr>
<th>Entity</th>
<th>Hashes</th>
</tr>
</thead>
<tbody>
<tr>
<td>E1</td>
<td>h1, h2</td>
</tr>
<tr>
<td>E2</td>
<td>h1, h3, h4</td>
</tr>
<tr>
<td>E3</td>
<td>h3</td>
</tr>
<tr>
<td>E4</td>
<td>h4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Key</th>
<th>Post List</th>
</tr>
</thead>
<tbody>
<tr>
<td>h1</td>
<td>E1, E2</td>
</tr>
<tr>
<td>h2</td>
<td>E1</td>
</tr>
<tr>
<td>h3</td>
<td>E1, E2, E3</td>
</tr>
<tr>
<td>h4</td>
<td>E2, E4</td>
</tr>
</tbody>
</table>
Improve Efficiency of Matching

• Reduce the number of entities comparisons (Indexing or Blocking)

1. Identify Blocking Attributes
   • Quality of values in the attributes may directly cause recall loss
   • Frequency and distribution of values directly impact performance and recall.
   • Best practice:
     • Use several attributes with combinations
     • Estimate and/or learn Identity Attributes
       • Movie name and release date –or– movie name, producer and director
       • Person name, date of birth and place of birth –or person name, affiliation and age

<table>
<thead>
<tr>
<th>Entity</th>
<th>Hashes</th>
</tr>
</thead>
<tbody>
<tr>
<td>E1</td>
<td>h1, h2</td>
</tr>
<tr>
<td>E2</td>
<td>h1, h3, h4</td>
</tr>
<tr>
<td>E3</td>
<td>h3</td>
</tr>
<tr>
<td>E4</td>
<td>h4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Key</th>
<th>Post List</th>
</tr>
</thead>
<tbody>
<tr>
<td>h1</td>
<td>E1, E2</td>
</tr>
<tr>
<td>h2</td>
<td>E1</td>
</tr>
<tr>
<td>h3</td>
<td>E1, E2, E3</td>
</tr>
<tr>
<td>h4</td>
<td>E2, E4</td>
</tr>
</tbody>
</table>
Improve Efficiency of Matching

• Reduce the number of entities comparisons (Indexing or Blocking)

2. Hashing Functions
   • PassThrough: \( H(\text{Tom Cruse}) = \{\text{Tom Cruse}\} \)
   • TokenSequence: \( H(\text{Tom Cruse}) = \{\text{crusetom}\} \)
   • Metaphone: \( H(\text{Robert}) = H(\text{Rupert}) \)

   • Q-Gram (a lot of hashes per value)
     • \( H(\text{Smith}) = \{\text{smmiitth, miitth, smith, smmith, smmiit}\} \)
     • (1) Compute grams (2) concat except one
     • 2-Gram(\text{smith}) = \{sm, mi, it, th\}
     • \( H(\text{Smith}) \cap H(\text{Smithy}) \cap H(\text{Smithe}) = \{\text{smmiith}\} \)

   • Suffix Array (a lot of duplicate post list)
     • \( H(\text{Catherine}) = \{\text{catherine, atherine, therine, herine}\} \)

   • Minhash
Improve Efficiency of Matching: Hashing Functions

- **MinHash:** min-wise independent permutations

- Convert the string to a set of elements

- Random function to give a random order for all the elements in the universe

- For two sets of elements \( S_1, S_2 \)

  \[ J(S_1, S_2) = P_r(\text{minhash}(S_1) = \text{minhash}(S_2)) \]

- Example:

  \( s_1 = \{c, d, e, f, g\} \quad s_2 = \{c, d, x, f, g\} \)

  - Order1: a,b,c,d,e,f,g, ...x
    - \( s_1 = \{g, f, e, d, c\} \quad s_2 = \{x, g, f, d, c\} \)
    - \( \text{minhash}(s_1) = c \quad \text{minhash}(s_2) = c \)
  
  - Order2: a,g,d,x,e,b,f,c...
    - \( s_1 = \{c, f, e, d, g\} \quad s_2 = \{c, f, x, d, g\} \)
    - \( \text{minhash}(s_1) = g \quad \text{minhash}(s_2) = g \)

  If \( \text{sim}(s_1, s_2) = 0.6 \), then by generating two minhashes, they will overlap with probability

  \[ 1 - (1 - 0.6)(1 - 0.6) = 1 - (0.4 \times 0.4) = 1 - 0.16 = 0.84 \]
Improve Efficiency of Matching

• Reduce the number of entities comparisons (Indexing or Blocking)

3. Retrieval
   • Within blocks comparison
   • Sorted Neighborhood
   • Canopy Clustering (cluster by random picking centroid, threshold based on distance, and nearest neighbor for cluster identification)
   • Entity Index Join
Improve Efficiency of Matching

• Reduce the number of entities comparisons (Indexing or Blocking)

3. Retrieval
   • **Within blocks comparison**
   • Sorted Neighborhood
   • Canopy Clustering (cluster by random picking centroid, threshold based on distance, and nearest neighbor for cluster identification)
   • Entity Index Join
Improve Efficiency of Matching

• Reduce the number of entities comparisons (Indexing or Blocking)

3. Retrieval
   • Within blocks comparison
   • Sorted Neighborhood
   • Canopy Clustering (cluster by random picking centroid, threshold based on distance, and nearest neighbor for cluster identification)
   • Entity Index Join
Improve Efficiency of Matching

- Reduce the number of entities comparisons (Indexing or Blocking)

3. Retrieval
   - Within blocks comparison
   - Sorted Neighborhood
   - **Canopy Clustering** (cluster by random picking centroid, threshold based on distance, and nearest neighbor for cluster identification)
   - Entity Index Join
Improve Efficiency of Matching

• Reduce the number of entities comparisons (Indexing or Blocking)

3. Retrieval
  • Within blocks comparison
  • Sorted Neighborhood
  • Canopy Clustering (cluster by random picking centroid, threshold based on distance, and nearest neighbor for cluster identification)
  • Entity Index Join

\[
\begin{align*}
E_1 &= \{(h, idf_1(h)) \ldots\} \\
E_2 &= \{(h, idf(h)) \ldots\} \\
L_1(E_1, E_2) &= \sum_{h} idf_1(h) \times idf_2(h)
\end{align*}
\]

Return Top K entities for each other entity
Satori Graph Build

Data Ingestion
- selection of data sources
- Data Preparation
- Targeted fact extraction by NLP and entity linking

Match & Merge
- Matching Entity Contents
- Detection of Matched Entities
- Scaling & Improve Efficiency of Matching

Knowledge Refinement
- Knowledge Fusion
- Error Detection
- Fact Inference

Publishing & Serve
- Entity search API
- Ranking & Filtering by attributes
- Graph walk
- Semantic Linking & Join
Knowledge Fusion (Merging Entities)

- After merging entity nodes in the graph, we end up with conflicting facts and connections
- Resolving facts (and finding truth)
  - Majority Voting
  - Identify Authoritative Sources
  - Fact Checker
    - Gather evidence from different sources
    - Evaluate evidences
    - Model joint interactions
    - Aggregate evidence and predict

Dong et al, "From data fusion to knowledge fusion" VLDB 2014
Dong et al, "Fact checking: theory and practice." KDD 2018
Satori Graph Build

Data Ingestion
- selection of data sources
- Data Preparation
- Targeted fact extraction by NLP and entity linking

Match & Merge
- Matching Entity Contents
- Detection of Matched Entities
- Scaling & Improve Efficiency of Matching

Knowledge Refinement
- Knowledge Fusion
- **Error Detection**
- Fact Inference

Publishing & Serve
- Entity search API
- Ranking & Filtering by attributes
- Graph walk
- Semantic Linking & Join
Error Detection

• Error Detection
  • Data Quality Rules
    • Functional Dependency and its conditional variation
e.g.; Zip $\rightarrow$ City
    • Inconsistency
      Entity cannot be a movie and book
      Date_of_birth $<$ date_of_death
    • Outliers detection

• External signals for relationship validation (e.g.; co-clicks)

• NLP features (e.g.; deadlive)
Satori Graph Build

Data Ingestion
- selection of data sources
- Data Preparation
- Targeted fact extraction by NLP and entity linking

Match & Merge
- Matching Entity Contents
- Detection of Matched Entities
- Scaling & Improve Efficiency of Matching

Knowledge Refinement
- Knowledge Fusion
- Error Detection
- Fact Inference

Publishing & Serve
- Entity search API
- Ranking & Filtering by attributes
- Graph walk
- Semantic Linking & Join
Fact Inference

• Further Enrichment/Data completion
  
  • Internal: Dominant type and Label
  
  • External: search engine method for enriching social links
Satori Graph Build

**Data Ingestion**
- selection of data sources
- Data Preparation
- Targeted fact extraction by NLP and entity linking

**Match & Merge**
- Matching Entity Contents
- Detection of Matched Entities
- Scaling & Improve Efficiency of Matching

**Knowledge Refinement**
- Knowledge Fusion
- Error Detection
- Fact Inference

**Publishing & Serve**
- Entity search API
- Ranking & Filtering by attributes
- Graph walk
- Semantic Linking & Join
Part IV: Serving Knowledge to the World

Ahmed K. A. Mohamed
Senior Applied Data Scientist, Satori Group, Microsoft AI+R
ahmedat@microsoft.com
Why Knowledge Graph Serve

- Documents/Web links
- Precise Answers
- Knowledge
Satori Knowledge Graph Application Areas: Bing&Cortana

Satori data and serve APIs has a tremendous impact on all Bing impressions for e.g.
Satori Knowledge Graph Application Areas: Office

Enriching the Office experience with Satori data

- **Researcher in Word & OneNote**: Get topic information straight into your documents.
- **Project Yellow (Excel)**: Finance and demographic information available based on cell contents.
- **LinkedIn profile information visible in O365 People Card through Satori**

**Challenges**

- Online conflation of people
- Compliance (training, code scanning & fixing, onboarding to new tools, etc.)
- Relevance with sparse profiles w/o access to the raw queries
Serving Knowledge by Answering Questions

• Given:
  • Knowledge graph ingested from unstructured, structured, and semi-structured data sources

• Input:
  • Natural language query

• Output:
  • Answer in the form of knowledge
Serving Knowledge by Answering Questions

Seattle Seahawks · Head coach
Pete Carroll

/American_football_team_current_head_coach
Serving Knowledge by Answering Questions

• Challenges:
  • Matching language
    • There are many ways to ask the same query e.g. {who directed titanic}, {what is the name of the person who directed titanic}, {in the movie titanic, who was the director}, ...etc
  • Scalable entity linking
  • Word sense disambiguation
  • Semantic roles and relationships extraction

• Large search space
  • Every entity can have hundreds of edges and every entity instance can have hundreds of millions of edges/facts

• Compositionality
  • {Movies starring the first wife of tom hanks}, {movies directed by the director of titanic}
Serving Knowledge by Answering Questions

• Approaches:

1. Semantic parsing approaches (serving graph as output):
   1.1 Generic semantic parsing followed by ontology grounding
   1.2 Knowledge base specific semantic parsing
   1.3 Knowledge embedding

2. Information extraction approaches (serving passage outputs):
   2.1 Information retrieval methods with semantic enrichment
1. Semantic Parsing

Semantic Parser

Context + Query

Knowledge

Abstract Meaning Representation

Film Entity \( ?f \)

year(2017)

Film.Actor Entity \( ?a \)

Film Entity \( ?f_b \)

Role Relationship \( ?r \)

Film.Character Entity \( ?c_b \)

“Batman”

Film.Actor

Actor.Performance

Character.Performance

Film.Name

Character.Name

Spider-Man: Homecoming

American Assassin

2017 - Suspense

The Founder

2017 - Biography

2017 movies starring the actor that played Batman in Batman

1.1 Generic Semantic Parsing

• In this approach as in the example provided by [Kwiatkoski 13], we:
  1. Perform a generic semantic parsing of the utterances
  2. Perform ontology matching on relationships

• For e.g. {who is Donald Trump’s Daughter}
  1. $\lambda x. \text{daughter}_o\text{f}(\text{Donald Trump}, x)$
  2. $\lambda x. \text{child}_o\text{f}(\text{Donald Trump}, x)^{\text{gender}} (x, \text{female})$

• This semantic expression can be then compiled into a knowledge graph database query e.g. SPARQL and executed to return the results
Dependency parsers: Arc-standard [Nivre 2004]

Figure 1: An example of transition-based dependency parsing. Above left: a desired dependency tree, above right: an intermediate configuration, bottom: a transition sequence of the arc-standard system.
Arc-standard actions are then learned using for e.g. stack LSTM [Dyer 2015]

Figure 2: Parser state computation encountered while parsing the sentence “an overhasty decision was made.” Here S designates the stack of partially constructed dependency subtrees and its LSTM encoding; B is the buffer of words remaining to be processed and its LSTM encoding; and A is the stack representing the history of actions taken by the parser. These are linearly transformed, passed through a ReLU nonlinearity to produce the parser state embedding $p_t$. An affine transformation of this embedding is passed to a softmax layer to give a distribution over parsing decisions that can be taken.
Ontology Matching on Relationships using DSSM [Shen+ 14]

- Input is mapped into two k dimensional vectors
- Probability is determined by softmax of their cosine similarity
1.2 Knowledge base specific semantic parsing

Constituency parsers:
PCFG Chart Parsing

Grammar is learned independently from an annotated dataset

Fig. 1 of [Bao et al., 2014]
1.3 Knowledge Embedding for e.g. [Bordes 2014]
2. Information extraction approaches

- Extracting and answers on the fly.
- These approaches provide ways to leverage the knowledge graph in cases where the question cannot be covered by the ontology or the data or both.
Information extraction approaches

Question

Understanding

Web Corpus

Who first landed on the Moon?

Type Detection, NER Parsing and Candidate Ranking

Apollo 11 was the spaceflight that landed the first humans on the Moon, Americans Neil Armstrong and Buzz Aldrin, on July 20, 1969, at 20:18 UTC.

Open Domain Question and Answering via Semantic Enrichment [Huan Sun, et al., WWW 2015]
Information Extraction Approach

Question Answering [Dan Jurafsky, Stanford]
Answer Type Detection

• Who first landed on the moon => Person
• Where is the headquarters of Microsoft => Location
• What is the largest country in population => Country
• Highest flying bird => Animal/Bird
Answer Type Detection

• Rules:
  • Grammar for e.g. who be/... => Person
  • Head word for e.g. which city is the largest

• Learned type classifier e.g. SVM utilizing features like question words, phrases, POS tags, headwords, mentioned entities, ...etc [Dan Jurafsky]
Passage Retrieval

• Retrieve documents using expanded query terms + search engine
• Segment the documents into smaller units e.g. passages/paragraphs
• Rank passages using learned model utilizing features like:
  • Number of named entities of the right type in the passage
  • Number of query words in the passage
  • Number of question n-grams in the passage
  • Proximity of query words in the passage
  • Longest sequence of question words
  • Rank of document containing passage,...etc
Process Answer

• Detect answer entity by running NER on the passage
• Mark the answer entity in the passage

• How many bones in an adult human body? (Number)
  • The human skeleton is the internal framework of the body. It is composed of 270 bones at birth – this total decreases to 206 bones by adulthood after some bones have fused together.

Question Answering [Dan Jurafsky, Stanford]
Answer Semantic Enrichment using KB [Huan Sun, et al., WWW 2015]

• 5-20% MRR improvement
Serving Knowledge Through Dialogs

• Approaches:
  • E2E Seq2seq (Ritter et al., 2011; Sordoni et al., 2015; Shang et al., 2015; Vinyals and Le, 2015)
  • Knowledge based ontological slot filling (Dai+ 2017)
  • Knowledge grounded neural approaches (Ghazvininejad+ 2018)
E2E Dialog Systems (e.g. Sordoni et al. 2015)

• Suitable for chitchat kind of bots.
• Predicted target sequences are usually free from facts

*Figure 8:* A computational graph representing the HRED architecture for dialogue over a span of three turns. The major addition to the architecture is a higher-level context-RNN keeping track of past utterances by progressively processing over time each utterance vector and conditioning the decoding on the last hidden state of the context vector (middle).
Knowledge Based Ontological Slot Filling
Knowledge Grounded Neural Approaches e.g. [Ghazvininejad+ 2018]

Figure 3: Knowledge-grounded model architecture.
Enterprise Scenarios

• All the challenges mentioned previously plus the following:
  • Compliance
  • Different data formats: databases, emails, chat logs, discussion forums, web blogs, pdfs, PowerPoint/Word/Excel documents etc.
  • Different schemas: schema mapping and merging, and new schema discovery.
  • Consumption via dialog systems, search interface, mobile devices or other modalities, API.
  • Highly domain-specific models required, bootstrapped by pre-trained models. Need on-prem domain-adaptation.
Questions