Outline

- Motivation - Knowledge base
- Related work - Overview of KBA using Web data
- Data source - Web markup
- Approach - KBA with Web Markup
Knowledge Base
Knowledge bases

In this talk, knowledge bases refers to RDF datasets.

dbr:Forrest_Gump dbo:label "Forrest Gump".
dbr:Forrest_Gump dbo:runtime "142.0"^^ns26:minute.
dbr:Forrest_Gump rdf:type dbo:Film.
dbr:Robert_Zemeckis dbo:label "Robert Zemeckis".
dbr:Robert_Zemeckis dbo:birthDate "1952-05-14"^^xsd:date.
dbr:Robert_Zemeckis dbo:birthPlace dbr:Chicago.
Applications

- Semantic search
- answer entity-centric queries
- question answering system
- rich representation of result
- product search
- product recommendations
- smart assistant (e.g. Amazon Alexa)
- ...

* Xin Luna Dong, Building a Broad Knowledge Graph for Products, Keynote at SigIR eCOM'19
Current state

- attract a lot of attention
- fast growing scale
- diverse domains
- ...

1. https://lod-cloud.net/

However ...

Linked Data: 1239 datasets & > 100 billion statements by March 2019¹

+ comercial KBs such as Google Knowledge Graph
Challenge: Incompleteness of KBs

In particular, there is a large percentage of less popular (long-tail) entities and properties that are under-represented. Some examples:

- Freebase - Person

<table>
<thead>
<tr>
<th>Relation</th>
<th>Percentage unknown</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>All 3M</td>
</tr>
<tr>
<td>PROFESSION</td>
<td>68%</td>
</tr>
<tr>
<td>PLACE OF BIRTH</td>
<td>71%</td>
</tr>
<tr>
<td>NATIONALITY</td>
<td>75%</td>
</tr>
<tr>
<td>EDUCATION</td>
<td>91%</td>
</tr>
<tr>
<td>SPOUSES</td>
<td>92%</td>
</tr>
<tr>
<td>PARENTS</td>
<td>94%</td>
</tr>
<tr>
<td>CHILDREN</td>
<td>94%</td>
</tr>
<tr>
<td>SIBLINGS</td>
<td>96%</td>
</tr>
<tr>
<td>ETHNICITY</td>
<td>99%</td>
</tr>
</tbody>
</table>

Table 1: Incompleteness of Freebase for some relations that apply to entities of type PERSON. Left: all 3M Freebase PERSON entities. Right: only the 100K most frequent PERSON entities.

- DBpedia, freebase, wikidata (2017)
  - Entity description of 30 movie & 30 book
  - The existence of 15 properties

Opportunity

- KBs constitute a small fraction of the Web
- More knowledge on the Web to be discovered!

1. https://lod-cloud.net/
2. https://www.worldwidewebsize.com/
Mining knowledge from the Web
KB enrichment & construction using Web data

- **Internal data from the KBs**
  e.g. inference on existing knowledge [Lehmann et al. 2011, Bühmann et al. 2012+2013, Socher et al. 2013]

- **External data from the Web**
  - unstructured Web documents, e.g.
    - news [Gerber et al. 2013]
    - general webpages: Google knowledge vault [Dong et al. 2014]
    - product description: Amazon Product Graph [Dong 2018]
  - semi-structured, e.g.
    - Web tables [Ritze et al. 2016]
Web markup
Web markup

Embed structured entity information into webpages using standards such as:

- Microformats
- RDFa
- Microdata

WDC datasets: triples extracted from the Web markup of webpages in Common Crawl

WDC @ Nov 2018:
- 37.1% webpages
- 29.3% PLDs
- 31.56 billion RDF quads
advantages & disadvantages

❖ Large scale
❖ up-to-date information
❖ long-tail entities & properties
❖ use ontologies
❖ semantic annotations & relations
❖ easy to parse ...

❖ Common errors
➢ wrong namespace
➢ undefined types & properties
➢ wrong property type (object v.s. datatype)

❖ Unlinked coreferences
❖ High redundancy
❖ Lack of links (isolated nodes)
❖ Semantically misused vocabularies

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coreferences</td>
<td>18,000 results for &lt;&quot;Iphone 6&quot;, type, s:Product&gt; (8,6 quads on average) in CommonCrawl</td>
</tr>
<tr>
<td>Redundancy</td>
<td>&lt;s, schema:name, &quot;Iphone 6&quot;&gt; occurring 1000 times in CC</td>
</tr>
<tr>
<td>Lack of links</td>
<td>Largely unlinked entity descriptions</td>
</tr>
<tr>
<td>Errors (typos &amp; schema violations, see Meusel et al [ESWC2015])</td>
<td>Wrong namespaces, such as <a href="http://schema.org">http://schema.org</a></td>
</tr>
<tr>
<td></td>
<td>Undefined types &amp; predicates: 9,7 %, less common than in LOD</td>
</tr>
<tr>
<td></td>
<td>Confusion of datatype and object properties:</td>
</tr>
<tr>
<td></td>
<td>&lt;s1, s:publisher, &quot;Springer&quot;&gt;, 24,35 % object property issues vs 8% in LOD</td>
</tr>
<tr>
<td></td>
<td>Data property range violations: e.g. literals vs numbers</td>
</tr>
<tr>
<td></td>
<td>(12,6% vs 4,6 in LOD)</td>
</tr>
</tbody>
</table>


Mining Knowledge from Structured Web Markup
KnowMore: KBA using markup

- 0. Noise: data cleansing (node URIs, deduplication etc)
- 1.a) Scale: Blocking (BM25 entity retrieval) on markup index
- 1.b) Relevance: supervised coreference resolution
- 2.) Quality & novelty: data fusion through supervised fact classification (SVM, knn, RF, LR, NB), diverse feature set (authority, relevance etc), considering source- (eg PageRank), entity-, & fact-level

Approach - entity matching

- **Blocking**: BM25 search in Lucene index on name related fields
- **Similarity vector** between KB entity description and markup entity description:
  
  \[ \text{Sim}(v^KB, v) = \{\lambda_{a_1}, \lambda_{a_2}, \ldots, \lambda_{a_n}\} \]

  Where

  \[ \lambda_{a_i} = \text{Sim}(f_{a_i}^{KB}, f_{a_i}) \]

- **Supervised classifier** to classify the relation between a pair of entity descriptions into binary class

---

Approach - data fusion

● **Correctness:**
  ○ Feature extraction
    ■ Relevance: e.g. BM25 score
    ■ Clustering: e.g. cluster size
    ■ Quality: e.g. PLD pagerank
  ○ Supervised classification to classify a fact into binary class [correct, incorrect]

● **Novelty**
  ○ Clustering of near duplicates (only one fact from each cluster selected)
  ○ Remove KB existing fact with datatype-specific similarity metrics
Experiment

- Dataset: WDC 2014
- Entities to enrich: randomly selected 30 movies & 30 books from wikipedia
- Ground truth: Crowdsourced & manually labeled
- KBs to augment: DBpedia, Freebase, Wikidata
- Properties to augment: 15 top populated properties, manually mapped across KBs
Performance of the binary classification (true v.s. false fact)

- Metrics: precision, recall, F1
- Baselines:
  - different configurations of the approach
  - PrecRecCorr [Pochampally et al. 2014] — data fusion baseline with focus on precision and quality of facts
  - CBFS [Yu et al. 2015] — data fusion baseline with focus on novelty

- outperform baselines wrt. F1
- Naive Bayes classifier achieves highest F1 scores
- PrecRecCorr has slightly higher precision but significantly lower recall on Book class.
- Average number of statements: PrecRecCorr 4.88, KnowMore 8.83
Novelty wrt. target KBs

- novelty improved by 0.282 cross KBs and dataset after deduplication wrt KBs, novelty over 90% on average
- able to recall over 90% novel facts
- trade-off between novelty and recall (varied $\tau$ - similarity threshold)
- $\tau$ - do not have a strong influence, as many facts have non-literal values (e.g. numeric, datetime)

### Table: Novelty wrt. KBs

<table>
<thead>
<tr>
<th>KB</th>
<th>Nov($F_{\text{ded}}$)</th>
<th>Nov($F_{\text{nov}}, \tau = 0.3$)</th>
<th>Nov($F_{\text{nov}}, \tau = 0.5$)</th>
<th>Nov($F_{\text{nov}}, \tau = 0.7$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Movie</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DBpedia</td>
<td>0.631</td>
<td>0.963</td>
<td>0.962</td>
<td>0.962</td>
</tr>
<tr>
<td>Freebase</td>
<td>0.527</td>
<td>0.747</td>
<td>0.742</td>
<td>0.742</td>
</tr>
<tr>
<td>Wikidata</td>
<td>0.412</td>
<td>0.929</td>
<td>0.929</td>
<td>0.897</td>
</tr>
<tr>
<td></td>
<td>$R(F_{\text{ded}})$</td>
<td>$R(F_{\text{nov}}, \tau = 0.3)$</td>
<td>$R(F_{\text{nov}}, \tau = 0.5)$</td>
<td>$R(F_{\text{nov}}, \tau = 0.7)$</td>
</tr>
<tr>
<td>DBpedia</td>
<td>1</td>
<td>0.927</td>
<td>0.939</td>
<td>0.939</td>
</tr>
<tr>
<td>Freebase</td>
<td>1</td>
<td>0.942</td>
<td>0.957</td>
<td>0.957</td>
</tr>
<tr>
<td>Wikidata</td>
<td>1</td>
<td>0.963</td>
<td>0.963</td>
<td>0.963</td>
</tr>
<tr>
<td>Book</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DBpedia</td>
<td>0.736</td>
<td>0.962</td>
<td>0.929</td>
<td>0.92</td>
</tr>
<tr>
<td>Freebase</td>
<td>0.639</td>
<td>0.915</td>
<td>0.846</td>
<td>0.825</td>
</tr>
<tr>
<td>Wikidata</td>
<td>0.705</td>
<td>0.944</td>
<td>0.933</td>
<td>0.923</td>
</tr>
<tr>
<td></td>
<td>$R(F_{\text{ded}})$</td>
<td>$R(F_{\text{nov}}, \tau = 0.3)$</td>
<td>$R(F_{\text{nov}}, \tau = 0.5)$</td>
<td>$R(F_{\text{nov}}, \tau = 0.7)$</td>
</tr>
<tr>
<td>DBpedia</td>
<td>1</td>
<td>0.826</td>
<td>0.848</td>
<td>0.870</td>
</tr>
<tr>
<td>Freebase</td>
<td>1</td>
<td>0.833</td>
<td>0.846</td>
<td>0.846</td>
</tr>
<tr>
<td>Wikidata</td>
<td>1</td>
<td>0.791</td>
<td>0.814</td>
<td>0.837</td>
</tr>
</tbody>
</table>

**Metric:** Novelty (percentage of novel fact wrt. KB)
Coverage gain (CG)

- CG 34.75% on average across properties for DBpedia, 39.42% for Freebase and 36.49% for Wikidata
- CG varies strongly between predicates and entity types

More evaluation result in the paper

- coverage gain: proportion of augmented missing facts.
  e.g. y = 40 for genre means: 40% of movies without a genre has been augmented with at least 1 genre
Potential

- Comparison of obtained facts (before deduplication step) with DBpedia
  - "existing": fact already in DBpedia
  - "new": fact not existing in DBpedia (eg a book’s releaseDate in Wiki/DBpedia)
  - "new-p": property not existing in DBpedia (eg a book’s release countries)

- 60-70% new facts for books & movies
- 100% new facts for queried products (not existing in DBpedia apparently)
- Vast potential for KB augmentation
Potential

<table>
<thead>
<tr>
<th>Category</th>
<th>Fact</th>
</tr>
</thead>
<tbody>
<tr>
<td>existing</td>
<td>s:author, “Evelyn Waugh”</td>
</tr>
<tr>
<td>existing</td>
<td>s:inLanguage, “English”</td>
</tr>
<tr>
<td>existing</td>
<td>s:numberOfPages, “402”</td>
</tr>
<tr>
<td>existing</td>
<td>s:copyrightYear, “1945”</td>
</tr>
<tr>
<td>existing</td>
<td>s:description, “Evelyn Waugh’s classic novel retold, and now a major motion picture from Miramax. Academy Award and Tony Award Winner Jeremy Irons—who has starred in films such as Lolita, Die Hard with a Vengeance, and Reversal of Fortune—narrates, invoking memories of his performance in the television miniseries. Irons was nominated at that time for an Emmy Award and a Golden Globe for his role as Charles Ryder.”</td>
</tr>
<tr>
<td>new</td>
<td>s:description, “books.google.com - Adult fiction about a doomed aristocratic family in England between the wars....”</td>
</tr>
<tr>
<td>new</td>
<td>s:datePublished, “1999”</td>
</tr>
<tr>
<td>new</td>
<td>s:publisher, “Back Bay Books”</td>
</tr>
<tr>
<td>new</td>
<td>s:publisher, “Little, Brown &amp; Company”</td>
</tr>
<tr>
<td>new</td>
<td>s:publisher, “BBC Audiobooks America”</td>
</tr>
<tr>
<td>new</td>
<td>s:isbn, “9781904605577”</td>
</tr>
<tr>
<td>new</td>
<td>s:isbn, “9780316216456”</td>
</tr>
<tr>
<td>new</td>
<td>s:isbn, “1904605575”</td>
</tr>
<tr>
<td>new</td>
<td>s:isbn, “9780316926348”</td>
</tr>
<tr>
<td>new</td>
<td>s:isbn, “9781602834866”</td>
</tr>
<tr>
<td>new-p</td>
<td>s:bookFormat, s:Paperback</td>
</tr>
<tr>
<td>new-p</td>
<td>s:copyrightHolder, “Evelyn Waugh”</td>
</tr>
</tbody>
</table>
Conclusions

- The data on the Web can be used to construct/augment machine-readable knowledge bases.
- More factors of a fact need to be considered when populating KBs, e.g. a fact might only be true with a certain time period.