Knowledge Base Augmentation Using Tabular Data

Yoones A. Sekhavat, Denilson Barbosa
University of Alberta

Francesco di Paolo, Paolo Merialdo
Roma Tre University
Outline

• Motivation
  - Need for knowledge bases
  - Exploiting semantics of tabular data

• Triple extraction
  - Architecture
  - Probabilistic model

• Implementation

• Experiments and results

• Limitations and future work

All data used to build and test the models in our work can be found at http://cs.ualberta.ca/~denilson/data/Ldow14_ualberta_data.zip
Knowledge Bases at the Core of the LOD Cloud

- Semantic query answering
- Information integration
- Data cleaning
- Record linkage

Example
- Artists who are also politicians
- Which artists were born in the same place as John Lennon?
YAGO Knowledge Base

Slide from [Weikum, WSDM2009]
Barack Obama is the 44th President of the US. Obama is a graduate of Columbia University and Harvard Law School.

<"Barack Obama", PresidentOf, "United States of America">
<"Barack Obama", GraduateOf, "Columbia University">
<"Barack Obama", GraduateOf, "Harvard Law School">

Documents

Recognize Entities

Resolve Co-references

Split Sentences

Find Relations

Triple store
Relation extraction with dependencies

- Comparison of different relation extraction techniques and varying cost/benefit trade-offs [EMNLP’2013]

EXEMPLAR
https://github.com/U-Alberta/exemplar/
How does a knowledge base get built?

- Reinforcement cycle: find new instances, generate new patterns, test and repeat!
Problem statement

• Augmenting an existing knowledge base with facts expressed in tabular data on the Web

• Why tabular data?
  - Tables have inherent semantics which are often implicit
  - Tables are everywhere!

(Cafarella et al. 2008)

- Over 400,000 Excel spreadsheets in the Clueweb’09
- 154 million relational tables in English documents in Cluweb’09
- 389, 697 raw and geospatial datasets in data.gov
- 147 million relational tables in the 2012 Web Common Crawl
Example

(A snapshot of a table in Wikipedia)

<table>
<thead>
<tr>
<th>Name</th>
<th>Country</th>
<th>Club</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ronaldinho</td>
<td>Brazil</td>
<td>Barcelona FC</td>
</tr>
<tr>
<td>Fabio Cannavaro</td>
<td>Italy</td>
<td>Juventus</td>
</tr>
<tr>
<td>Kaka</td>
<td>Brazil</td>
<td>AC Milan</td>
</tr>
<tr>
<td>Lionel Messi</td>
<td>Argentina</td>
<td>Barcelona FC</td>
</tr>
</tbody>
</table>

• General approach:
  • link the values in each cell to known entities in a KB
  • identify relations between the linked values.

Best case scenario
• Entities are linked to the same KB
• Relation already exists between entities

Our take
• Entities in different or unlinked KB
• Entities are not linked to anything yet.

Table understanding (e.g., Limaye et al., 2010)
Knowledge base augmentation
Some insight into Google’s Knowledge Graph

• Thanks to Xin Luna Dong (Google), from yesterday’s talk at DEOS:

• TXT: text extraction

• DOM: deep-web extraction

• ANO: schema.org annotations

• TBL: Web tables
  ▪ Schema matching/table understanding approach
Goal
Augmenting an existing repository with new instances of relations already defined

Idea
• The fact that someone put some literals together in the same rows indicates that there are relationships between them
• Pairs of entities in different rows of two given columns share the same relation

Method
1. Collect all sentences containing both entities from a large text corpus
2. Extract the text in between them
3. Match those texts against the list of patterns
4. Estimate the posterior probability of all candidate relations.
• Knowledge base
  ▪ We used YAGO with about 10 million entities and over 120 million facts about them.

• Text corpus
  ▪ Our text corpus is the NELL Subject-Verb-Object (SVO) triple corpus, with about 604 million triples extracted from ClueWeb09 dataset.
  ▪ Clueweb09 is a crawl of the Web with about 1 billion web pages in ten different languages.

• Text patterns
  ▪ We used publicly available patterns from the PATTY project
  ▪ We used 4,357 distinct patterns from PATTY having intersection with NELL (intersection with 108,699,400 triples from NELL)

• Ground truth
  ▪ Facts from YAGO relations where both entities can be matched exactly in the NELL corpus.
A relation in a knowledge base can be represented by different textual patterns.

- NELLTriples
- Extract Patterns
- PATTY Patterns
- Patterns_1
- Patterns_2
- Patterns_n

A pattern may represent more than one relation:

- **plays-for**
  - “scored for”
  - “signed contract with”
- **plays-for**
  - (“played in”)
  - (“signed contract with”)
- **performed-at**
  - (“played in”)
We use Bayesian inference to compute the posterior probability of relation $r$ given the observed patterns $p_1,\ldots,p_k$.

\[
Pr(r|p_1,\ldots,p_k) = \frac{Pr(r)Pr(p_1,\ldots,p_k|r)}{Pr(p_1,\ldots,p_k)}
\]

\[
Pr(r) = \frac{|r|}{\sum_{r_i \in R} (|r_i|)}
\]

- $R$ is the set of all relations
- $PT(r)$ is the set of patterns associated with relation $r$
## Results (accuracy)

Number of PATTY patterns and resulting rank obtained by each strategy, for each relation.

<table>
<thead>
<tr>
<th>Relation</th>
<th>Patterns</th>
<th>SP</th>
<th>KE</th>
<th>MR</th>
<th>GR</th>
</tr>
</thead>
<tbody>
<tr>
<td>ismarriedto</td>
<td>1274</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>created</td>
<td>1148</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>haschild</td>
<td>1090</td>
<td>-</td>
<td>-</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>influences</td>
<td>694</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>actedin</td>
<td>624</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>-</td>
</tr>
<tr>
<td>graduatedfrom</td>
<td>472</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>isknownfor</td>
<td>452</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>worksat</td>
<td>447</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>holdspoliticalposition</td>
<td>417</td>
<td>-</td>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>directed</td>
<td>400</td>
<td>2</td>
<td>-</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>playsfor</td>
<td>354</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>diedin</td>
<td>335</td>
<td>3</td>
<td>-</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>wasbornin</td>
<td>273</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>islocatedin</td>
<td>249</td>
<td>-</td>
<td>3</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>livesin</td>
<td>200</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>isleaderof</td>
<td>156</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>iscitizenof</td>
<td>121</td>
<td>1</td>
<td>1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>haswonprize</td>
<td>81</td>
<td>-</td>
<td>-</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>dealswith</td>
<td>59</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>ispoliticianof</td>
<td>49</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>participatedin</td>
<td>33</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>happenedin</td>
<td>12</td>
<td>2</td>
<td>1</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>hascapital</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
Summary of accuracy results

All relations

<table>
<thead>
<tr>
<th></th>
<th>SP</th>
<th>KE</th>
<th>MR</th>
<th>GR</th>
</tr>
</thead>
<tbody>
<tr>
<td>A: Ranked First</td>
<td>10</td>
<td>11</td>
<td>10</td>
<td>7</td>
</tr>
<tr>
<td>B: Ranked Second</td>
<td>8</td>
<td>9</td>
<td>9</td>
<td>12</td>
</tr>
<tr>
<td>C: Ranked Third</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>1</td>
</tr>
</tbody>
</table>

Filtered relations

<table>
<thead>
<tr>
<th></th>
<th>SP</th>
<th>KE</th>
<th>MR</th>
<th>GR</th>
</tr>
</thead>
<tbody>
<tr>
<td>A: Ranked First</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>B: Ranked Second</td>
<td>9</td>
<td>7</td>
<td>9</td>
<td>8</td>
</tr>
<tr>
<td>C: Ranked Third</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>D: Ranked &gt; 3</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

denilson@ualberta.ca --- LDOW'14
Knowledge augmentation pre-experiment

• Test #1:
  - Input table - a spreadsheet including song data available at http://www.aardvarkdjservices.co.uk
  - Our technique found 48 triples for the created relation
  - Among those, 31 were already present in YAGO

• Test #2:
  - Input table - a spreadsheet with data about NBA players extracted from http://www.espn.go.com
  - Found 100 triples for the plays-for relation,
  - Of these, YAGO had 92 triples in the is-affiliated-to relation
Runtime

- The average execution times (ms) for processing a pair of entities (taken over 20 executions) are:

<table>
<thead>
<tr>
<th>SP</th>
<th>KE</th>
<th>MR</th>
<th>GR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1688</td>
<td>1868</td>
<td>1729</td>
<td>1719</td>
</tr>
</tbody>
</table>

- There are no considerable differences among the methods

- The majority of the time is spent on matching the entities against the NELL corpus
Summary and Conclusion

• We described a probabilistic approach for augmenting linked open data repositories using tabular data.

• Unlike prior methods that focus on natural language understanding, we started from the (reasonable) assumption that all entities in the same row of a table are related by definition.

• Unlike previous methods that attempt to understand tabular data, we label pairs of columns in the table with relations coming from an established knowledge base.
Summary and Conclusion

• Limitations
  § Small number of YAGO relations ➔ currently experimenting with Freebase
  § Exact entity matching

• Other applications besides knowledge base augmentation
  § Estimating:
    • How many new triples could be extracted from tabular data on the Web?
    • How accurate are they?
  § Using both quantitative and qualitative metrics to chart which websites provide the best data for knowledge base augmentation
• Knowledge base
  ▪ YAGO ➔ Freebase

• Text corpus
  ▪ Still NELL
  ▪ Musing about indexing all of Clueweb for this

• Text patterns
  ▪ PATTY ➔ Google’s annotated Clueweb with Freebase entities

• Ground truth
  ▪ Facts from YAGO
  ▪ Facts from Freebase (ranging popularity)
Work in progress: (summary of results)

• Filtering
  ▪ Relations associated with less than 1000 or more than 1 million patterns
  ▪ Pattern with length of >12
  ▪ Patterns with frequency below 10

• Ground truth
  ▪ Extracted from freebase facts
  ▪ 50 pairs for each relation
  ▪ Pairs are selected in a way including high and low number of patterns

<table>
<thead>
<tr>
<th>Per-Domain</th>
<th>Ranked First</th>
<th>Ranked Second</th>
<th>Ranked Third</th>
<th>Ranked &gt;3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Location</td>
<td>7</td>
<td>1</td>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td>People</td>
<td>6</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Organization</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>-</td>
</tr>
<tr>
<td>Miscellaneous</td>
<td>9</td>
<td>4</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>All</td>
<td>17</td>
<td>7</td>
<td>6</td>
<td>21</td>
</tr>
</tbody>
</table>

denilson@ualberta.ca --- LDOW’14
Knowledge Base Augmentation Using Tabular Data

Yoones A. Sekhavat, Denilson Barbosa
University of Alberta

Francesco di Paolo, Paolo Merialdo
Roma Tre University

NELL, CMU

Xin Luna Dong, Google

7th International Workshop on Linked Data on the Web
Work in progress (Improved Probabilistic model)

• The types of entities were considered in the model
• We focused on standard NER types: Location, People, Organization and Miscellaneous

\[
Pr(r|p_1, \ldots, p_k, \langle tx, ty \rangle) = \frac{Pr(r)Pr(p_1, \ldots, p_k|r)Pr(\langle tx, ty \rangle|r)}{Pr(p_1, \ldots, p_k)}
\]

• We generated quadruples \(Q = (\text{entity}_1, \text{pattern}, \text{entity}_2, \text{relation})\) from annotated clueweb09 using named entities in Freebase
• Improvement in estimating prior probabilities

\[
Pr(p|r) = |\{q \in Q|\text{pat}(p) \land \text{rel}(r)\}|/|\{q \in Q|\text{rel}(r)\}|
\]
Work in progress (challenges)
Work in progress (challenges)

Frequency of patterns in annotated Clueweb09

Length of pattern

0 500000 1000000 1500000 2000000 2500000 3000000 3500000
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22

denilson@ualberta.ca --- LDOW'14