GRAMS: A Graph-based Approach for Inferring Semantic Descriptions of Wikipedia Tables

Binh Vu, Craig Knoblock, Pedro Szekely, Jay Pujara, Minh Pham

Information Sciences Institute
University of Southern California
Motivating Example

- Wikipedia has 7.5 millions tables covering many domains

List of players won Walter Payton Award

Members of 56th New Brunswick Legislature

<table>
<thead>
<tr>
<th>Name</th>
<th>Party</th>
<th>Riding</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hédard Albert</td>
<td>Liberal</td>
<td>Caraquet</td>
</tr>
<tr>
<td>David Alward</td>
<td>Progressive Conservative</td>
<td>Woodstock</td>
</tr>
<tr>
<td>Donald Arseneault</td>
<td>Liberal</td>
<td>Dalhousie-Restigouche East</td>
</tr>
<tr>
<td>John Betts</td>
<td>Progressive Conservative</td>
<td>Moncton Crescent</td>
</tr>
<tr>
<td>Dereham</td>
<td>Dereham</td>
<td>Regeneron</td>
</tr>
<tr>
<td>Aldeby</td>
<td>Aldeby</td>
<td>F2G</td>
</tr>
<tr>
<td>Ashwellthorpe</td>
<td>Ashwellthorpe</td>
<td>1881 1939</td>
</tr>
<tr>
<td>Hotel Diablo</td>
<td></td>
<td>Great Eastern</td>
</tr>
<tr>
<td>FLYGOD is an Awesome GOD</td>
<td>westside Gunn</td>
<td>Shellharbour 1959</td>
</tr>
<tr>
<td>pericarditis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ebola Virus</td>
<td></td>
<td>invasive mold infections</td>
</tr>
</tbody>
</table>
Source Modeling Problem

• Building semantic descriptions of tables
  – Describing data source using classes and properties in ontologies

<table>
<thead>
<tr>
<th>Name</th>
<th>Entered Office</th>
<th>Left Office</th>
<th>Party</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thomas Prinzhorn</td>
<td>2002</td>
<td>2006</td>
<td>FPÖ</td>
</tr>
<tr>
<td>Eva Glawischnig-Piesczek</td>
<td>2006</td>
<td>2008</td>
<td>Grüne</td>
</tr>
<tr>
<td>Martin Graf</td>
<td>2008</td>
<td>2013</td>
<td>FPÖ</td>
</tr>
</tbody>
</table>

Third Presidents of National Council (Austria)
Source Modeling Problem

• Building semantic descriptions of tables
  – Describing data source using classes and properties in ontologies

<table>
<thead>
<tr>
<th>Name</th>
<th>Entered Office</th>
<th>Left Office</th>
<th>Party</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thomas Prinzhorn</td>
<td>2002</td>
<td>2006</td>
<td>FPÖ</td>
</tr>
<tr>
<td>Eva Glawischnig-Piesczek</td>
<td>2006</td>
<td>2008</td>
<td>Grüne</td>
</tr>
<tr>
<td>Martin Graf</td>
<td>2008</td>
<td>2013</td>
<td>FPÖ</td>
</tr>
</tbody>
</table>

Third Presidents of National Council (Austria)
Main Idea

• Information of entities in KGs can help source modeling
  \[\Rightarrow\text{need little training data}\]

President of the National Council (Austria)

From Wikipedia, the free encyclopedia

List of third presidents  [edit]

<table>
<thead>
<tr>
<th>Name</th>
<th>Entered Office</th>
<th>Left Office</th>
<th>Party</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thomas Prinzhorn</td>
<td>2002</td>
<td>2006</td>
<td>FPÖ</td>
</tr>
<tr>
<td>Eva Glawischnig-Piesczek</td>
<td>2006</td>
<td>2008</td>
<td>Grüne</td>
</tr>
<tr>
<td>Martin Graf</td>
<td>2008</td>
<td>2013</td>
<td>FPÖ</td>
</tr>
</tbody>
</table>

Eva Glawischnig-Piesczek

Austrian politician

Die Grünen

position held

member of political party

Third President of the National Council of Austria

start time

30 October 2006

end time

28 October 2008
Approach

Linked table

Contextual values

Semantic Description
Construct Candidate Graph: Discovering Links

- Create a graph of cells and context

<table>
<thead>
<tr>
<th>Name</th>
<th>Entered Office</th>
<th>Left Office</th>
<th>Party</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thomas Prinzhorn</td>
<td>2002</td>
<td>2006</td>
<td>FPÖ</td>
</tr>
<tr>
<td>Eva Glawischnig-Piesczek</td>
<td>2006</td>
<td>2008</td>
<td>Grüne</td>
</tr>
<tr>
<td>Martin Graf</td>
<td>2008</td>
<td>2013</td>
<td>FPÖ</td>
</tr>
</tbody>
</table>

Third President context
Construct Candidate Graph: Discovering Links

- Add links discovered from knowledge in Wikidata
Construct Candidate Graph: Summarization

- Group links of cells from same source & target columns/context

P39: position held
P580: start time
P582: end time
Construct Candidate Graph: Summarization

- **Second President**
  - Thomas
  - Eva
  - Martin

- **Third President**
  - Party: FPO, Grune, FPO

**Properties:**
- P39: position held
- P580: start time
- P582: end time
Construct Candidate Graph: Summarization

- **Name**
  - Thomas
  - Eva
  - Martin

- **Entered Office**
  - 2002
  - 2006
  - 2008

- **Left Office**
  - 2006
  - 2008
  - 2013

- **Party**
  - FPO
  - Grune
  - FPO

**Labels**
- P39: position held
- P580: start time
- P582: end time
Construct Candidate Graph: Summarization
Construct Candidate Graph: Summarization

- Final candidate graph

P39 : position held
P580: start time
P582: end time
After Building Candidate Graph

• Candidate (n-ary) relationships from the candidate graph
• Candidate columns’ types from entities in table columns

⇒ Need to select the most appropriate relationships and types.
Approach

Inputs
- A target knowledge graph: Wikidata
- A linked relational table $T$
- A set of contextual values $C$

1. Construct candidate graph
2. Infer semantic description

Outputs:
- A semantic description of $(T, C)$
Collective Reasoning Problem

• **Probabilistic Soft Logic (PSL)**
  “A probabilistic graphical models framework using first-order logic”

• **Two main elements: predicates and rules**
  – Predicates have “soft” value in [0, 1]
  – Rules converted to exponential function to approximate $P(x)$
PSL Predicates (examples)

- **CorrectRel\((N_1, N_2, P)\)**: if a relationship is correct
  - CorrectRel(\(Name, stmt_1, P39\))
  - CorrectRel(\(stmt_1, Entered\ Office, P580\))
  - CorrectRel(\(stmt_1, Third\ President, P39\))

- **CorrectType\((N_1, T)\)**: if a column type assignment is correct
  - CorrectType(\(Party, Organization\))
  - CorrectType(\(Party, Political\ Party\))
  - CorrectType(\(Name, Human\))

- ... and more

\(P39\): position held  \(P580\): start time  \(P582\): end time
PSL Rules (examples)

1. By default, relationships/types are incorrect
   1a. $\neg$ CorrectRel($N_1$, $N_2$, P)
   1b. $\neg$ CorrectType($N_1$, T)

2. Relationships/types are correct/incorrect based on evidence
   2a. FreqMatch($N_1$, $N_2$, P) $\rightarrow$ CorrectRel($N_1$, $N_2$, P)
   2b. FreqDiff($N_1$, $N_2$, P) $\rightarrow$ $\neg$ CorrectRel($N_1$, $N_2$, P)
   2c. FreqTypeMatch($N_1$, T) $\rightarrow$ CorrectType($N_1$, T)
   2d. ...and more
PSL Rules (examples)

3. If a statement value is incorrect, then the statement’s qualifiers are also incorrect

4. We prefer fine-grain properties than high-level properties

5. ...and more
Post-Processing

- PSL outputs probability of each relationships and types.
  - Place served (P931): 0.91
  - Located in (P131) p=0.89

- Use BANK algorithm to choose the most probable relationships
  - Avoid unnecessary loops
  - Prefer tree structure if possible

BEFORE
- Airport
- City
- P131
- City
- State
- Contained in (P150)

AFTER
- City
- State
- Country

(P131) contains
(P150)
Evaluation of GRAMS

• Collective reasoning is beneficial
  – Avoid cascading errors from subject column detection phase
  – Handle complex schema: multiple entities’ types and n-ary relationships

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Method</th>
<th>CPA Precision</th>
<th>CPA Recall</th>
<th>CPA F₁</th>
<th>CTA Precision</th>
<th>CTA Recall</th>
<th>CTA F₁</th>
</tr>
</thead>
<tbody>
<tr>
<td>250WT</td>
<td>MantisTable</td>
<td>0.535</td>
<td>0.442</td>
<td>0.484</td>
<td></td>
<td>0.928</td>
<td>0.331</td>
</tr>
<tr>
<td></td>
<td>MantisTable*</td>
<td>0.559</td>
<td>0.569</td>
<td>0.564</td>
<td></td>
<td>0.940</td>
<td>0.394</td>
</tr>
<tr>
<td></td>
<td>BBW</td>
<td>0.796</td>
<td>0.123</td>
<td>0.214</td>
<td></td>
<td>0.850</td>
<td>0.233</td>
</tr>
<tr>
<td></td>
<td>BBW*</td>
<td>0.740</td>
<td>0.559</td>
<td>0.638</td>
<td></td>
<td>0.759</td>
<td>0.777</td>
</tr>
<tr>
<td></td>
<td>GRAMS-ST</td>
<td>0.526</td>
<td>0.681</td>
<td>0.594</td>
<td></td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>GRAMS</td>
<td>0.824</td>
<td>0.650</td>
<td>0.726</td>
<td></td>
<td>0.819</td>
<td>0.813</td>
</tr>
<tr>
<td>SemTab2020</td>
<td>MantisTable</td>
<td>0.985</td>
<td>0.976</td>
<td>0.981</td>
<td></td>
<td>0.977</td>
<td>0.800</td>
</tr>
<tr>
<td></td>
<td>BBW</td>
<td>0.996</td>
<td>0.995</td>
<td>0.995</td>
<td></td>
<td>0.980</td>
<td>0.980</td>
</tr>
<tr>
<td></td>
<td>GRAMS-ST</td>
<td>0.990</td>
<td>0.989</td>
<td>0.990</td>
<td></td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>GRAMS</td>
<td>0.996</td>
<td>0.994</td>
<td>0.995</td>
<td></td>
<td>0.982</td>
<td>0.981</td>
</tr>
</tbody>
</table>

MantisTable* and BBW* are modified to retrieve correct subject column.
## Related Work

<table>
<thead>
<tr>
<th>Method</th>
<th>Data Hungry</th>
<th>Modeling Capabilities</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Handle Literal Columns</td>
</tr>
<tr>
<td>Custom Ontologies</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Taheriyan et al. 2016</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Vu et al. 2019</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Iterative Method</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ritze et al. 2015</td>
<td>–</td>
<td>Y</td>
</tr>
<tr>
<td>Zhang et al. 2017</td>
<td>–</td>
<td>Y</td>
</tr>
<tr>
<td>SemTab systems</td>
<td>–</td>
<td>Y</td>
</tr>
<tr>
<td>KG Ontologies</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Limaye et al. 2010</td>
<td>–</td>
<td>N</td>
</tr>
<tr>
<td>Mulward et al. 2013</td>
<td>–</td>
<td>N</td>
</tr>
<tr>
<td>GRAMS</td>
<td>–</td>
<td>Y</td>
</tr>
</tbody>
</table>
Discussion and Future work

• **Contribution:** A novel graph-based approach, GRAMS, for building semantic descriptions of Wikipedia Tables.
  – The candidate graph makes it easy to represent and discover n-ary relationships.
  – Using PSL to collectively infer correct relationships and types.

• **Future work:**
  – Handle unlinked tables
  – Generate large labeled dataset from Wikipedia tables to train semantic modeling systems