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 albums in 2019 (USA)

<table>
<thead>
<tr>
<th>Date</th>
<th>Album</th>
<th>Artist</th>
<th>Genre(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>Metawar</td>
<td>3Teeth</td>
<td>Industrial · industrial metal</td>
</tr>
<tr>
<td></td>
<td>Stonechild</td>
<td>Jesca Hoop</td>
<td>Folk · blues · pop</td>
</tr>
<tr>
<td></td>
<td>Hotel Diablo</td>
<td>Machine Gun Kelly</td>
<td>Hip hop</td>
</tr>
<tr>
<td></td>
<td>FLYGOD is an Awesome GOD</td>
<td>Westside Gunn</td>
<td>Hip hop</td>
</tr>
</tbody>
</table>
Motivating Example

- Wikipedia has 7.5 millions tables covering many domains
Motivating Example

- Wikipedia has 7.5 millions tables covering many domains
Motivating Example

- Wikipedia has 7.5 million tables covering many domains
Motivating Example

- Wikipedia has 7.5 millions tables covering many domains

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>
</tbody>
</table>

Indications

- Depressive disorder
- Multiple myeloma
- Giant pericarditis

<table>
<thead>
<tr>
<th>Indication</th>
<th>Website</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ebola Virus</td>
<td></td>
</tr>
<tr>
<td>invasive mold infections</td>
<td></td>
</tr>
</tbody>
</table>
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>Inmazeb</td>
</tr>
<tr>
<td>Aldeby</td>
<td>Aldeby</td>
<td>Olorofim</td>
</tr>
<tr>
<td>Ashwellthorpe</td>
<td>Ashwellthorpe</td>
<td>Great Eastern</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Year</th>
<th>Player</th>
<th>School</th>
<th>Position</th>
</tr>
</thead>
<tbody>
<tr>
<td>1987</td>
<td>Kenny Gamble</td>
<td>Colgate</td>
<td>RB</td>
</tr>
<tr>
<td>1988</td>
<td>Dave Meggett</td>
<td>Towson State</td>
<td>RB</td>
</tr>
<tr>
<td>2018</td>
<td>Devlin Hodges</td>
<td>Samford</td>
<td>QB</td>
</tr>
<tr>
<td>2019</td>
<td>Trey Lance</td>
<td>North Dakota State</td>
<td>QB</td>
</tr>
</tbody>
</table>

- FLYGOD is an Awesome GOD

- Hotel Diablo

- westside Gunn

- Hip hop

- Ebola Virus

- Invasive mold infections
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

⟹ 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 Glawischneg-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>
Main Idea

• Information of entities in KGs can help source modeling

$\implies$ 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 (Q93870)

Austrian politician

- **member of political party**
  - Die Grünen

- **position held**
  - Third President of the National Council of Austria
    - **start time**: 30 October 2006
    - **end time**: 28 October 2008
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** (Q93870)

Austrian politician

- member of political party: Die Grünen
- position held: Third President of the National Council of Austria
- start time: 30 October 2006
- end time: 28 October 2008
Main Idea

• Information of entities in KGs can help source modeling
  ⟹ 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 (Q93870)

Austrian politician

member of political party

Die Grünen

position held

Third President of the National Council of Austria

start time

30 October 2006

dead time

28 October 2008
**Approach**

**Linked table**

<table>
<thead>
<tr>
<th>Name</th>
<th>Entered Office</th>
<th>Left Office</th>
<th>Party</th>
</tr>
</thead>
<tbody>
<tr>
<td>Willi Brauneder</td>
<td>1996</td>
<td>1999</td>
<td>FPÖ</td>
</tr>
<tr>
<td>Thomas Prinzhorn</td>
<td>2002</td>
<td>2006</td>
<td>FPÖ</td>
</tr>
<tr>
<td>Eva Glawischneg-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>

**Candidate Graph**

**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>
Construct Candidate Graph: Discovering Links

- Add links discovered from knowledge in Wikidata

- P39 : position held
- P580: start time
- P582: end time
Construct Candidate Graph: Discovering Links

- Add links discovered from knowledge in Wikidata

![Graph showing connections between Thomas Prinzhor, 2002, 2006, and FPO]
Construct Candidate Graph: Discovering Links

- Add links discovered from knowledge in Wikidata

**Thomas Prinzhorn** (Q88195)

Austrian entrepreneur and politician

- **member of political party**
  - P102: FPÖ
  - 0 references

- **position held**
  - P39: Second President of the National Council of Austria
  - P580: start time
    - 29 October 1999
  - P582: end time
    - 20 December 2002
Construct Candidate Graph: Discovering Links

• Add links discovered from knowledge in Wikidata

Thomas Prinzhorn (Q88195)
Austrian entrepreneur and politician

- member of political party: FPÖ
- position held: Second President of the National Council of Austria
  - start time: 29 October 1999
  - end time: 20 December 2002
Construct Candidate Graph: Discovering Links

• Add links discovered from knowledge in Wikidata
Construct Candidate Graph: Discovering Links

• Add links discovered from knowledge in Wikidata
Construct Candidate Graph: Discovering Links

- Add links discovered from knowledge in Wikidata

WD Statement

**Thomas Prinzhorn** (Q88195)

Austrian entrepreneur and politician

- **member of political party**
  - P102
  - **FPÖ**
  - 0 references

- **position held**
  - P39
  - Second President of the National Council of Austria
    - P580: start time
    - 29 October 1999
    - P582: end time
    - 20 December 2002

**fuzzy match**

**entity in WD**

**(entity in WD)**

**entity in WD**
Construct Candidate Graph: Discovering Links

- Add links discovered from knowledge in Wikidata

Diagram:

- Thomas
- 2002
- 2006
- FPO

Fuzzy match:

Thomas Prinzhorn (Q88195)

Member of political party: FPÖ

Position held:
- Second President of the National Council of Austria
  - Start time: 29 October 1999
  - End time: 20 December 2002
Construct Candidate Graph: Discovering Links

- Add links discovered from knowledge in Wikidata

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

• Add links discovered from knowledge in Wikidata


P39: position held
P580: start time
P582: end time
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

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

P39: position held
P580: start time
P582: end time
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

- Thomas: Second President (2002-2006)
- Eva: Second President (2006-2008)
- Martin: Third President (2006-2013)

Properties:
- P39: position held
- P580: start time
- P582: end time

FPO
Grune
FPO
Construct Candidate Graph: Summarization

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

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

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

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

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

- Thomas: 2002-2006
- Eva: 2006-2008
- Martin: 2008-2013

- FPO
- Grune
- FPO

- Third President
- Second President

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

- **Name**: Thomas, Eva, Martin
- **Entered Office**: 2002
- **Left Office**: 2006, 2008, 2013
- **Party**: FPO, Grune, FPO

**Notes**:
- P39: Position held
- P580: Start time
- P582: End time
Construct Candidate Graph: Summarization

Name: Thomas, Eva, Martin
Party: FPO, Grune, FPO

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

• Candidate (n-ary) relationships *from the candidate graph*
• Candidate columns’ types *from entities in table columns*
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.
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, stmt₁, P39)
  - CorrectRel(stmt₁, Entered Office, P580)
  - CorrectRel(stmt₁, 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. \( \neg \text{CorrectRel}(N_1, N_2, P) \)
   1b. \( \neg \text{CorrectType}(N_1, T) \)

2. Relationships/types are correct/incorrect based on evidence
   2a. \( \text{FreqMatch}(N_1, N_2, P) \rightarrow \text{CorrectRel}(N_1, N_2, P) \)
   2b. \( \text{FreqDiff}(N_1, N_2, P) \rightarrow \neg \text{CorrectRel}(N_1, N_2, P) \)
   2c. \( \text{FreqTypeMatch}(N_1, T) \rightarrow \text{CorrectType}(N_1, T) \)
   2d. ...and more
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
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</th>
<th>CTA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Precision</td>
<td>Recall</td>
</tr>
<tr>
<td>250WT</td>
<td>MantisTable</td>
<td>0.535</td>
<td>0.442</td>
</tr>
<tr>
<td></td>
<td>MantisTable*</td>
<td>0.559</td>
<td>0.569</td>
</tr>
<tr>
<td></td>
<td>BBW</td>
<td>0.796</td>
<td>0.123</td>
</tr>
<tr>
<td></td>
<td>BBW*</td>
<td>0.740</td>
<td>0.559</td>
</tr>
<tr>
<td></td>
<td>GRAMS-ST</td>
<td>0.526</td>
<td>0.681</td>
</tr>
<tr>
<td></td>
<td>GRAMS</td>
<td>0.824</td>
<td>0.650</td>
</tr>
<tr>
<td>SemTab2020</td>
<td>MantisTable</td>
<td>0.985</td>
<td>0.976</td>
</tr>
<tr>
<td></td>
<td>BBW</td>
<td>0.996</td>
<td>0.995</td>
</tr>
<tr>
<td></td>
<td>GRAMS-ST</td>
<td>0.990</td>
<td>0.989</td>
</tr>
<tr>
<td></td>
<td>GRAMS</td>
<td>0.996</td>
<td>0.994</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>Zhang et al. 2017</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>
<tr>
<td>Graphical Models</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Ontologies**

**Taheriyan et al. 2016**

**Vu et al. 2019**

**Iterative Method**

**Ritze et al. 2015**

**Zhang et al. 2017**

**SemTab systems**

**KG Ontologies**

**Zhang et al. 2017**

**Mulward et al. 2013**

**Graphical Models**

**Limaye et al. 2010**

**Mulward et al. 2013**
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