Learning to Refine an Automatically Extracted Knowledge Base Using Markov Logic

Shangpu Jiang, Daniel Lowd and Dejing Dou

Department of Computer Information and Science
University of Oregon, USA
Overview

- We show an example of a KB: NELL
- We propose to refine it with Markov Logic
- To improve **scalability & accuracy**
  - Inference: Neighborhood-based method
  - Learning: *calibrated confidence* with an approximate learning objective
### NELL’s Knowledge Base

#### Recently-Learned Facts

<table>
<thead>
<tr>
<th>instance</th>
<th>category</th>
</tr>
</thead>
<tbody>
<tr>
<td>lehman_brothers_holdings_inc is an organization</td>
<td>Categorical fact</td>
</tr>
<tr>
<td>richard_weisner is an actor</td>
<td>Categorical fact</td>
</tr>
<tr>
<td>mycoplasma_glycophilum is a bacterium</td>
<td>Relational fact</td>
</tr>
<tr>
<td>wawa_coffee is a beverage</td>
<td>Relational fact</td>
</tr>
<tr>
<td>taliban_al_qaeda is a terrorist organization</td>
<td>Relational fact</td>
</tr>
<tr>
<td>rogers_farm is a farm in the state or province florida</td>
<td>Relational fact</td>
</tr>
<tr>
<td>tampa_bay_buccaneers is a sports team that plays against green_bay_packers</td>
<td>Relational fact</td>
</tr>
<tr>
<td>the sports league nascar uses the venue richmond_international_raceway</td>
<td>Relational fact</td>
</tr>
<tr>
<td>n2011_pro_bowl is a sports team that plays in the league nfl</td>
<td>Relational fact</td>
</tr>
<tr>
<td>mary is a person who moved to the state new_york</td>
<td>Relational fact</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>iteration</th>
<th>date learned</th>
<th>confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>646</td>
<td>15-oct-2012</td>
<td>100.0</td>
</tr>
<tr>
<td>644</td>
<td>10-oct-2012</td>
<td>98.9</td>
</tr>
<tr>
<td>644</td>
<td>10-oct-2012</td>
<td>98.8</td>
</tr>
<tr>
<td>644</td>
<td>10-oct-2012</td>
<td>98.0</td>
</tr>
<tr>
<td>646</td>
<td>15-oct-2012</td>
<td>92.6</td>
</tr>
<tr>
<td>649</td>
<td>21-oct-2012</td>
<td>100.0</td>
</tr>
<tr>
<td>646</td>
<td>15-oct-2012</td>
<td>93.8</td>
</tr>
<tr>
<td>647</td>
<td>17-oct-2012</td>
<td>100.0</td>
</tr>
<tr>
<td>649</td>
<td>21-oct-2012</td>
<td>93.8</td>
</tr>
<tr>
<td>649</td>
<td>21-oct-2012</td>
<td>100.0</td>
</tr>
</tbody>
</table>
How NELL works

- Bootstrapped Pattern Learning (Brin 98)

**Country:**
- Canada
- Egypt
- France

**Pakistan**
- Sri Lanka
- Argentina

**Planet Earth**
- North Africa
- Student Council

- X is the only country
- Home country of X

- Invasion of X
- Elected president of X
How to Avoid the Drift

* Coupled Learning (Carlson et al., 2009)
  * Use *ontological constraints* to filter out wrong candidates

<table>
<thead>
<tr>
<th>Candidate</th>
<th>Ontology</th>
<th>Existing fact</th>
<th>Operation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Athlete(x)</td>
<td>Subclass(Athlete, Person)</td>
<td>¬Person(x)</td>
<td>Remove</td>
</tr>
<tr>
<td>Athlete(x)</td>
<td>Disjoint(Athlete,City)</td>
<td>City(x)</td>
<td>Remove</td>
</tr>
<tr>
<td>PlaysSport (x, y)</td>
<td>Domain(PlaysSport, Athlete)</td>
<td>¬Athlete(x)</td>
<td>Remove</td>
</tr>
</tbody>
</table>
Our Proposed Method

- **Problem** in NELL:
  - Past decisions never changed
  - Information is integrated using heuristics

- **Solution**: joint inference
  - Markov Logic Networks (Richardson and Domingos 2006)
  - Build upon NELL’s KB
NELL in a MLN

SubClass(Athlete, Person) \rightarrow (Athlete(Bryant) \rightarrow Person(Bryant))

In first-order logic: SubClass(c_1, c_2) \land InClass(x, c_1) \rightarrow InClass(x, c_2)
Inference

- **Problem:** MLN is large (billions of nodes)
- **Solution:** Neighborhood-based inference
  - Select some query facts
  - Only consider 2-hop neighborhood
- Runs in several hours on a single core
Learning

* **Problem**: confidence values from NELL are ad-hoc and uncalibrated
* **Solution**: 2-layer logistic regression to map confidence to *probability*
  * LR on patterns: an additional confidence measure
  * Combined with NELL’s confidence measures using another LR
  * Approximate objective: discard hard ontological constraints in learning
Experiments

* Dataset
  * Take NELL’s 165th iteration
  * Label facts from 13 relations and 10 categories
  * Training set: 9,887 facts from 5 relations and 6 categories
  * Test set: 4,511 facts, about 200 facts for each relation and category
Result: PR-AUC

Area under precision-recall curve (AUC)
Result: F1
Discussion

* Example 1
  * City(Los_Angeles_County)
  * County(Los_Angeles_County)
  * The former was found first, so NELL trusted it

* Example 2
  * ProducesProduct(Adobe, Acrobat_reader_software)
  * ProducesProduct(Adobe, Acrobat_reader_version)
  * Both are acceptable in NELL
Summary

* We propose to use Markov Logic to refine a knowledge base
* Inference on 2-hop neighbors
* Learning with logistic regression to calibrate the confidence values
* Improved NELL without any human labels; did even better with human labels

Acknowledgement We thank NELL’s group, especially Burr Settles, for providing the data. We also thank Ameneh Sarbaziazad for the labeling.