KB-LDA: Jointly Learning a Knowledge Base of Hierarchy, Relations, and Facts

Dana Movshovitz-Attias and William W Cohen

ACL, July 28, 2015
Who is Barack Obama’s wife?

Michelle Obama (m. 1992)
Barack Obama, Spouse

Michelle Obama, née Robinson, the wife of Barack Obama, was born on January 17, 1964, in Chicago, Illinois. She is a lawyer and was a University of Chicago ...
Who is Barack Obama’s wife?

Knowledge Base

Barack Obama married Michelle Obama

wife of

husband of
What is the initial size of an ArrayList?

What is the Python equivalent of a HashMap?

What is the run time of quick sort?
Knowledge Base (KB) Construction

Ontology-Guided Construction

- NELL
- FreeBase
- Yago
- WordNet

Open IE

- ReVerb
- TextRunner

example sentences:
- (Beijing, is the capital of, China)
- (Penticton, has very, warm summers)
- (Goods, can be defined in, a variety of ways)
Reasoning with Ontologies

- Ontologies give information context
- Easy to extract domain-specific information
Reasoning with Ontologies

✔ Ontologies give information context
✔ Easy to extract domain-specific information

✘ Ontologies are
  • expensive
  • require prior knowledge
  • incomplete/outdated

✘ Ontology structure and facts are often drawn from different corpus statistics
Goal: Data-Driven Knowledge Base

- Schema and facts are drawn from corpus and jointly optimized
- Unsupervised: learn latent corpus structure together with best-matching facts
Automatically Learning a KB with Structure and Facts

Corpus

Collection of Documents from StackOverflow

Relation Extraction

KB-LDA

KB Evaluation

Ontology

Documents

Relations

Everything

Barack Obama

Michelle Obama

married

Sasha

Malia

father

mother

Harvard

CMU

Columbia

People

Everything

Universities

Places

Cities

born in

studied at

Honolulu

Chicago

studied at

Columbia

studied at

Harvard
Automatically Learning a KB with Structure and Facts

KB-LDA

Ontology

Documents

Relations

Corpus

Collection of Documents from StackOverflow

Relation Extraction

KB Evaluation

KB-LDA

Everything

People

Places

Universities

Cities

Cities

People

Everything

KB-LDA

Corpus

Relation Extraction

Collection of Documents from StackOverflow

KB Evaluation

Ontology

Documents

Relations

Corpus

KB-LDA

Ontology

Documents

Relations

Corpus

KB-LDA

Ontology

Documents

Relations

Corpus

KB-LDA

Ontology

Documents

Relations

Corpus

KB-LDA

Ontology

Documents

Relations
Pattern-based Relation Extraction

Ontology

“places such as Hawaii”
“places including cities, towns and villages”

Concept(places)

Instance(Hawaii)  Sub-Concept(cities)

Hearst patterns

Y such as X
X is a Y
Y including X

Hypernym-hyponyms from StackOverflow

Languages such as Java
IDEs including Eclipse and NetBeans
Integer and other data types
Pattern-based Relation Extraction

Subject-Verb-Object patterns

“Barack Obama was born in Hawaii”
“Barack Obama was born in a hospital”

Instance(Barack Obama) ➔ born in ➔ Instance(Hawaii)

Concept(hospital)

SVOs from StackOverflow

Class implements interface
Constructor throws exception
Output fills buffer
Open Structure of Extracted Relations

- No distinction between concepts and instances
- Parsing errors: "class not found exception"

**Noise**
- programming language
- magic
- haskell

**Ambiguity**
- ui element
- behavior
- view
- php
- script

- Structure includes cycles:
Automatically Learning a KB with Structure and Facts

Corpus

Collection of Documents from StackOverflow

Relation Extraction

KB-LDA

Ontology

Documents

Relations

KB Evaluation

I successfully done with provision on both side client and server and when I call this code for the first time 1st if(line no. 2) and 2nd if(line no. 8) gets executed and if I call this code 2nd time then 1st else(line no. 4) get executed as client is already provisioned but here it execute 2nd if(line no. 8) instead of executing 2nd else(line no. 10) and throws an exception at line no. 9

server.provision.Apply(); "There is already an object named 'schema_info' in the database."

And if i try to synchronize then it throws an exception at line syncOrchestrator.Synchronize(); "The current operation could not be completed because the database is not provisioned for sync or you not..."
Figure 1: Plate Diagram of KB-LDA.
KB-LDA Model

Figure 1: Plate Diagram of KB-LDA.
KB-LDA Model

Figure 1: Plate Diagram of KB-LDA.
KB-LDA Model

Relations

\[ z_{S_j} \rightarrow S_j \rightarrow O_j \rightarrow z_{O_j} \rightarrow V_j \rightarrow z_{V_j} \rightarrow N_{T_j} \]

\[ \sigma_{t_1} \rightarrow N_{T_j} \rightarrow \gamma_{I} \rightarrow C_i \rightarrow z_{C_i} \rightarrow \pi_{O} \rightarrow \alpha_{O} \]

\[ \pi_{R} \rightarrow \alpha_{R} \rightarrow \text{Verb Topics} \]

\[ \text{Ontology} \rightarrow z_{C_i} \rightarrow \text{Documents} \]

\[ E_{1l} \rightarrow z_{E_{1l}} \rightarrow \pi_{D} \rightarrow \alpha_{D} \]

\[ E_{2l} \rightarrow z_{E_{12}} \rightarrow N_{D} \]

\[ N_{D, I} \rightarrow \text{studied at} \]

\[ N_{D, R} \rightarrow \text{married} \]

\[ \text{entity} \]

\[ \text{person} \]

\[ \text{location} \]

\[ \text{city} \]

\[ \text{university} \]

\[ \text{born in} \]

\[ \text{is husband of} \]

\[ \text{is wife of} \]

\[ \text{married} \]

\[ \text{parent} \]

\[ \text{father} \]

\[ \text{mother} \]

\[ \text{Sasha} \]

\[ \text{Malia} \]

\[ \text{Barack Obama} \]

\[ \text{Michelle Obama} \]

\[ \text{Harvard} \]

\[ \text{Chicago} \]

\[ \text{Honolulu} \]

\[ \text{CMU} \]

\[ \text{studied at} \]

\[ \text{Verb} \]

\[ \text{Topics} \]
KB-LDA Model

Figure 1: Plate Diagram of KB-LDA.
KB-LDA Model

Extracted hypernym-hyponym relations:

- websites → google
- platforms → stackoverflow

Noun Topic 1

websites platforms applications

Noun Topic 2

stackoverflow google facobook
KB-LDA Model

Relations

Ontology

Documents

Extracted SVO relation: Michelle, wife of, Barack
KB-LDA Model

Relations

\[ z_{S_j} \rightarrow S_j \]
\[ z_{O_j} \rightarrow O_j \]
\[ z_{V_j} \rightarrow V_j \]

\[ \sigma_{t_i} \]
\[ N_{T_i} \]
\[ \delta_{t_R} \]
\[ N_{T_R} \]

Ontology

\[ C_i \]
\[ z_{C_i} \]
\[ I_i \]
\[ z_{I_i} \]
\[ N_O \]

Documents

\[ E_{1l} \]
\[ z_{E_{1l}} \]
\[ N_{d,I} \]
\[ E_{2l} \]
\[ z_{E_{12}} \]
\[ N_{d,R} \]
\[ \pi_D \]
\[ \alpha_D \]

\[ \pi_R \]
\[ \alpha_R \]

Documents

Downloads on websites sometimes have an MD5 checksum, allowing people to confirm the integrity of the file. I have heard this is to allow not only corrupted files to be instantly identified before they cause a problem but also for for any malicious changes to be easily detected.
KB-LDA Model

More in paper:
Data-driven topic naming

Documents

Downloads on websites sometimes have an MD5 checksum, allowing people to confirm the integrity of the file. I have heard this is to allow not only corrupted files to be instantly identified before they cause a problem but also for any malicious changes to be easily detected.
Training

Append image file to form data - Cordova/Angular

I am using Angular, Ionic and Cordova in my current project, and I'm trying to POST FormData containing an image file to my server. Right now I'm using the cordova camera plugin to return a file path to the image on the device (ex: file://path/to/img). Once I have the file path I want to append the image file to a FormData object using the images file path. Here is my code right now.

```javascript
var fd = new FormData();
fd.append('attachment', file);
fd.append('uuid', uuid);
fd.append('userRole', userRole);
```

The code above works when appending a file that is taken from an `<input type='file'>` but doesn't work when just given the file path on the device.

Basically the FormData is showing like this right now:

```
-------WebKitFormBoundaryasdf
Content-Disposition: form-data; name="attachment";
file://path/to/img
```

2 Answers

You do need to send file content. With the **HTML5 FileAPI** you need to create a FileReader object.

Some time ago, I developed an application with cordova and I had to read some files, and I made a library called CoFS (first, by Cordova FileSystem, but it’s working in some browsers).

It’s on beta state, but I use it and it works well. You can try to do some like this:

```javascript
var errorHandler = function (err) {

```

Corpus
5.5M documents from StackOverflow

Topics
50
Automatically Learning a KB with Structure and Facts

I successfully done with provision on both side client and server and when I call this code for the first time if(line no. 2) and 2nd if(line no. 8) gets executed and if I call this code 2nd time then 1st else(line no. 4) get executed as client is already provisioned but here it execute 2nd if(line no. 8) instead of executing 2nd else(line no. 10) and throws an exception at line no. 9!

serverprovision.Apply(); “There is already an object named ‘schema_info’ in the database.”

And if i try to synchronize then it throws an exception at line syncOrchestrator.Synchronize(); “The current operation could not be completed because the database is not provisioned for sync or you not
# Top Tokens of Learned Noun Topics

<table>
<thead>
<tr>
<th>Database</th>
<th>Web Browsers</th>
<th>Programming Language</th>
<th>User Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>table</td>
<td>ie</td>
<td>java</td>
<td>name</td>
</tr>
<tr>
<td>query</td>
<td>firefox</td>
<td>python</td>
<td>images</td>
</tr>
<tr>
<td>database</td>
<td>chrome</td>
<td>javascript</td>
<td>id</td>
</tr>
<tr>
<td>sql</td>
<td>safari</td>
<td>lists</td>
<td>number</td>
</tr>
<tr>
<td>column</td>
<td>explorer</td>
<td>ruby</td>
<td>text</td>
</tr>
<tr>
<td>data</td>
<td>ie7</td>
<td>c +</td>
<td>password</td>
</tr>
<tr>
<td>tables</td>
<td>ie6</td>
<td>perl</td>
<td>address</td>
</tr>
<tr>
<td>mysql</td>
<td>ie8</td>
<td>haskell</td>
<td>strings</td>
</tr>
<tr>
<td>index</td>
<td>ie9</td>
<td>jquery</td>
<td>files</td>
</tr>
<tr>
<td>columns</td>
<td>internet explorer</td>
<td>scala</td>
<td>string</td>
</tr>
</tbody>
</table>
M-Turk Evaluation of Noun Topics

“Which words are not related to programming languages?”

java
python
javascript
firefox
ruby
perl

1. Word Intrusion

<table>
<thead>
<tr>
<th>Top Tags</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>0.78</td>
</tr>
<tr>
<td>10</td>
<td>0.78</td>
</tr>
<tr>
<td>15</td>
<td>0.78</td>
</tr>
<tr>
<td>20</td>
<td>0.78</td>
</tr>
<tr>
<td>25</td>
<td>0.78</td>
</tr>
<tr>
<td>30</td>
<td>0.78</td>
</tr>
</tbody>
</table>

2. Group Precision

<table>
<thead>
<tr>
<th>Top Tags</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>0.93</td>
</tr>
<tr>
<td>10</td>
<td>0.93</td>
</tr>
<tr>
<td>15</td>
<td>0.93</td>
</tr>
<tr>
<td>20</td>
<td>0.93</td>
</tr>
<tr>
<td>25</td>
<td>0.93</td>
</tr>
<tr>
<td>30</td>
<td>0.93</td>
</tr>
</tbody>
</table>
Relations

**Users** ➞ interact with ➞ **UI element**

- user
- people
- customer
- client
- player
- clicks
- selects
- submits
- hits
- moves
- button
- form
- link
- item
- file

“Is this a reasonable Software relationship?”

<table>
<thead>
<tr>
<th></th>
<th>KB-LDA</th>
<th>Random</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Avg(Experts)</strong></td>
<td>0.7</td>
<td>0.13</td>
</tr>
<tr>
<td>1 Worker</td>
<td>0.9</td>
<td>0.69</td>
</tr>
<tr>
<td>2 Workers</td>
<td>0.63</td>
<td>0.22</td>
</tr>
<tr>
<td>3 Workers</td>
<td>0.15</td>
<td>0.05</td>
</tr>
</tbody>
</table>
For each document:

1. Find most probable topic (e.g., databases)
2. Aggregate tags for topic

What tags are frequently associated with topics?
Top Tags Associated with Topics

- **databases**
  - sql
  - mysql
  - database
  - query
  - table
  - column

- **regular expressions**
  - string
  - character
  - characters
  - text
  - line

- **html elements**
  - element
  - div
  - css
  - elements
  - http

- **KB-LDA Topics**
  - string
  - character
  - characters
  - text
  - line

- **Human Tags**
  - regex
  - string
  - python
  - php
  - ruby
  - css
  - html
  - jquery
  - html5
  - javascript
Domain-Specific Extraction from Open IE

ReVerb
15m triples

Triples with software entities
5k triples

Subject-verb-object are in KB-LDA Dictionary

SVOs extracted directly from StackOverflow
37k triples

✔ (safari, supports, svg)
✔ (computer, is running, xp)

✗ (view, looks, south)
✗ (people, can read, italian)
Domain-Specific Extraction from Open IE

$p(SVO|\text{KB-LDA})$  
$p(SVO|\text{ReVerb})$

ReVerb  
15m triples

Triples with software entities  
5k triples

Subject-verb-object are in KB-LDA Dictionary

Domain-specific triples?

SVOs extracted directly from StackOverflow  
37k triples
Recall

KB-LDA versus ReVerb Ranking

Recall

Precision

KB-LDA, Best F1=0.73, AUC=0.67

ReVerb, Best F1=0.72, AUC=0.57
KB-LDA: Conclusion

✔ Corpus-driven KB construction: Jointly optimizes schema and facts

✔ Unsupervised: Useful for exploration of new domains

✔ KB-LDA can be used to extract domain-specific facts from an Open IE knowledge base

www.cs.cmu.edu/~dmovshov

Thank you!