Grounded Knowledge Bases for Scientific Domains

Dana Movshovitz-Attias
August 26, 2015

Thesis Oral

Thesis Committee:
William Cohen, Chair
Tom Mitchell
Roni Rosenfeld
Alon Halevy, Google Research
Who is Barack Obama’s wife?
Who is Barack Obama’s wife?

Michelle Obama
1992
Barack Obama, Spouse

Michelle LaVaughn Robinson Obama is an American lawyer and writer. She is married to the 44th and current President of the United States, Barack Obama, and is the first African-American First Lady of the United States. Wikipedia

More about Michelle Obama

Family of Barack Obama - Wikipedia, the free encyclopedia
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 ... Sidwell Friends School - Marian Shields Robinson - Bo - Charles T. Payne

Michelle Obama - Wikipedia, the free encyclopedia
Michelle Obama (born January 17, 1964) is an American lawyer and writer. She is the wife of the 44th and current President of the United ... Craig Robinson (basketball) - Hyde Park, Chicago - Sidney Austin - Valerie Jarrett

Michelle Obama - Biography - U.S. First Lady, Lawyer ...
Explore the life of Michelle Obama, the 44th first lady and wife of President Barack Obama. Learn more at ...

Michelle Obama - First Lady and wife of President Barack ...
Jul 1, 2015
Michelle LaVaughn Obama, First Lady and wife of US President Barack Obama: All the latest news and ...

Barack Obama was asked about 'his first wife' by a woman ...
Jan 22, 2015 - Barack Obama was asked about his 'first wife' by a woman who took a ... when she gave Obama a gift of green lipstick for Michelle Obama, who ...
Who is Barack Obama’s wife?
Software

“What is the **run time** of quick sort?”

QuickSort - Wikipedia, the free encyclopedia
https://en.wikipedia.org/wiki/QuickSort

Animated visualization of the quicksort algorithm. The horizontal lines are pivot values. QuickSort (sometimes called partition-exchange sort) is an efficient sorting algorithm, serving as a systematic method for placing the elements of an array in order.

Tony Hoare - Dutch national flag problem - Robert Sedgewick - Flashsort

Running Time of Quicksort - HackerRank
https://www.hackerrank.com/challenges/quicksort4

Mar 4, 2013 - The running time of Quicksort will depend on how balanced the partitions are. If you are unlucky and select the greatest or the smallest element ...

QuickSort Running Time - Math StackExchange
math.stackexchange.com/.../quicksort-running-time

Feb 18, 2011 - ... the recurrence will be: The above recurrence has the solution (I will prove this later): Hence the running time of QUICKSORT in this case is ...

Quick Sort - Personal.kent.edu
www.personal.kent.edu/~rmuhanna/Algorithms/.../Sorting/quickSort.htm

The running time of quick sort depends on whether partition is balanced or unbalanced, which in turn depends on which elements of an array to be sorted are used for partitioning. A very good partition splits an array up into two equal sized arrays.

...the average case running time is $\Theta(n \log n)$...

Biomedical

“What is the **molecular mass** of BamA?”

[PDF] Quiz 2 answers - with worked solutions - Bama.ua.edu
www.bama.ua.edu/...quiz2_key_with_solutions.pdf

Quizzes 2 answers - with worked solutions. 1. What is average mass, in grams, of one atoms of iron? Use mol wt of iron (= molar mass = mass for one mol of Fe ...

[PDF] Polymers: Introduction - Bama.ua.edu
bama.ua.edu/~kshaughn/ch338/.../poly-lecture.ppt

Monomer: Low molecular weight compound that can be connected together to give a polymer; Oligomer: Short polymer chain; Copolymer: polymer made up of 2 ...

[PDF] Hydrogen/Deuterium exchange mass spectrometry - Bam...
bama.ua.edu/.../Busenlehner_AABBReview_2005.p...

Oct 5, 2004 - exchange mass spectrometry (H/D exchange MS) is emerging as an efficient ... proteins having molecular masses in excess of 50 kDa.

TP0326, a Treponema pallidum β-Barrel Assembly ...

by DC Desrosiers - 2011 - Cited by 32 - Related articles

Apr 27, 2011 - In E. coli, BamA is the central component of a multi-protein complex consisting ..... Native TP0326 forms part of a high molecular mass complex.

The crystal structure of BamB suggests interactions with ...

by N Noinaj - 2011 - Cited by 47 - Related articles

Jan 26, 2011 - It interacts with the periplasmic domain of BamA, an integral outer ..... We determined the molecular mass of BamB in solution using size ...

BAMA Course Survey: ... Use masses listed above to calculate the molecular mass. ... This keeps the shape of the molecule "strang" or "linear" allowing these ...
Challenges in Mining Scientific Text

• Specialized terminology
• Domain-specific language constructs

Affected NLP Techniques

- Class Not Found Exception
- Noun Adverb Verb Noun

POS Tagging

PARSING

(ROOT
  (FRAG
    (NP (NN class) (RB not))
    (VP (VBD found)
      (NP (NN exception)))))

Entity Extraction & Relation Understanding

code

throws

exception

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. Then if this code 2nd time 2nd if (line no. 4) get executed because it 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

code

syncOrchestrator.Synchronize(); "The current operation could not be completed because the database is not provisioned for sync or you not …
Grounding Scientific Entities

Scientific data is not found only in text

Opportunity: Domain-specific resources

Code Repository

ArrayList

HashMap

BRCA1

Alzheimer

Biomedical Ontologies
Grounding Scientific Entities

Scientific data is not found only in text

Opportunity: Domain-specific resources

1. How are objects used?
2. What are their real-world properties?
Thesis Statement

“Grounding entities to specialized data from a scientific domain facilitates improved unsupervised and semi-supervised algorithms for **Knowledge Base construction** for that domain”
Knowledge Base Construction

Open IE

(Beijing, is the capital of, China)
(Penticton, has very, warm summers)
(Goods, can be defined in, a variety of ways)

- ReVerb [Fader et al., 2011]
- TextRunner [Yates et al., 2007]

Ontology-Guided Construction

- NELL [Carlson et al., 2010]
- FreeBase [Google, 2011]
- Yago [Suchanek et al., 2007, 2008]
- Knowledge Vault [Dong et al., 2014]
Reasoning with Ontologies

- Ontologies give information context
- Easy to extract domain-specific information
- Ontologies are expensive
- Require prior knowledge
- Manual ontology does not reflect language statistics
Roadmap

Statistical Language Model for Software Domain Application

/* comment prediction */
Predicting Code Comments

**Model** code with statistical language models

**Predict** class comment

```java
package org.apache.lucene.index;
import java.io.IOException;
import java.io.Closeable;
public abstract class TermEnum implements Closeable {
    public abstract boolean next() throws IOException;
}
```

Abstract class for enumerating terms.

- Term enumerations are always ordered by `Term.compareTo()`. Each term in the enumeration is greater than all that precede it.

**Evaluate** how much typing can we save?

Up to 47% !

Train a named-entity extractor

[Movshovitz-Attias & Cohen, ACL13]
Code Modeling

1. Shallow statistical model
   **N-gram**

2. BOW topic model
   **LDA**

3. Basic domain encoding
   (code vs. text entities)
   **Link-LDA**
Predicting Code Comments

• Comments are highly predictable
  – 47% of characters predicted

• Un-intuitively: Shallow statistical method performed best
  3-gram: Train a named-entity extractor
  link-LDA: Train a named-entity extractor

• Intuitively: Domain encoding improves performance
  Link-LDA > LDA
We Need Deeper Understanding of Domain Entities

/* This method reads the next double from stdin */

- float
- long
- int

Data Type

Numerical Data Type

Signed Floating-Point Type
- double
- float

Signed Integer Type
- long
- int

**Semantics** and **categorical understanding** contribute to language modeling task
We Need Deeper Understanding of Domain Entities

Semantics and categorical understanding contribute to language modeling task
Roadmap

Statistical Language Model for Software Domain Application

/* comment prediction */

Bootstrap KB Learning for the Biomedical Domain
Bootstrap KB Learning for the Biomedical Domain

- Modify KB learning system (NELL) for biomedical domain

![Diagram showing KB Learning process]

**Resources**
- Categories
- Seeds

**Semi-Supervised Bootstrapping**

**KB**
- Learned facts
- KB growth

**Coupled Bootstrapping**

**Biomedical Corpus**

[Movshovitz-Attias & Cohen, BioNLP @ NAACL12]
Semantic Drift in NELL
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.

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.

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.

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.
BioNELL: Lessons

• Biomedical ontologies + filtering bootstrapping seeds lead to:
  – High-precision biomedical KB
  – Improves domain applications (NER)

• Disadvantage: Relies on input ontologies
  – No existing Software ontologies
Reasoning with Ontologies

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

✘ Ontologies are
  • expensive
  • require prior knowledge

✘ Manual ontology does not reflect language statistics

✘ Ontology and facts are often drawn from different sources
Roadmap

Statistical Language Model for Software Domain Application

/* comment prediction */

Bootstrap KB Learning for the Biomedical Domain

Grounded Software Ontology Construction

Corpus

I successfully done with provision and started the job, and when I call this code for the first time line no. 2 and 2nd if(line no. 8) get 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...
Grounded Software Ontology Construction

- We detect coordinate relations (similarity) between Java classes
  - “This method iterates over ArrayLists and HashMaps”

[Movshovitz-Attias & Cohen, 15]
I successfully done with provision on both side client and server and when time (let it in no. 8) gets 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
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
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
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

"...Is a root node an internal node?..."
Grounded Software Ontology Construction

• We detect coordinate relations (similarity) between Java classes
  – “This method iterates over ArrayLists and HashMaps”

[Movshovitz-Attias & Cohen, 15]
Utility Classes

- TreeSet
- SortedMap
- SortedSet
- HashSet
- BitSet
- ArrayList
- HashMap
- ArrayDeque
- PriorityQueue
- BlockingQueue
- ThreadPoolExecutor
- ConcurrentHashMap
- LinkedBlockingQueue
- Future

Nodes: Classes  
Edges: Coordinate relations  
Edge color: community detection (Louvain method)  
Label size: betweenness centrality
Contributions & Lessons

• Linked class entities to code implementation
• Defined distributional similarity for code
• By combining code and text similarities we learned relations (and ontology)

• Advantage: This ontology reflects statistics in language and code
  – In contrast to manually built ontologies

• Grounding to code limits scope of learned ontology to code entities
  – What’s missing?

  Users
  Computer resources
  Design patterns
Roadmap

Statistical Language Model for Software Domain Application

/* comment prediction */

Bootstrap KB Learning for the Biomedical Domain

Grounded Software Ontology Construction

Topic-Model KB Learning

Corpus

I successfully done with provision 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

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 no.

34
Goal: Corpus-Driven Knowledge Base

• Schema and facts are drawn from corpus
• Unsupervised: learn optimal latent corpus structure together with best-matching facts
Reasoning with Ontologies

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

✘ Ontologies are
• expensive
• require prior knowledge

✔ Manual ontology does not reflect language statistics
✔ Ontology and facts are often drawn from different sources
Pattern-based Relation Extraction

1. Hypernym-hyponym

“websites such as stackoverflow”
“websites including google and twitter”

Y such as X
X is a Y
Y including X

2. Subject-Verb-Object

“user clicks button”
“user clicks form”
KB-LDA Model

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.
Hypernym-hyponym relations:

- websites → google
- platforms → stackoverflow
KB-LDA Model

Figure 1: Plate Diagram of KB-LDA.

Relations

Ontology

Documents

Extracted SVO relation: person, clicks, button

S → V → O
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.
KB-LDA Model

More in thesis document: 
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.
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!

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!
<table>
<thead>
<tr>
<th>Subject</th>
<th>Topic</th>
<th>Verb</th>
<th>Topic</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>user</td>
<td>people</td>
<td>clicks</td>
<td>selects</td>
<td>submits</td>
</tr>
<tr>
<td>customer</td>
<td>client</td>
<td>adds</td>
<td>name</td>
<td>images</td>
</tr>
<tr>
<td>client</td>
<td>person</td>
<td>edits</td>
<td>id</td>
<td>password</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>code</th>
<th>function</th>
<th>method</th>
<th>query</th>
</tr>
</thead>
<tbody>
<tr>
<td>server</td>
<td>class</td>
<td>object</td>
<td>client</td>
</tr>
<tr>
<td>program</td>
<td>makes</td>
<td>takes</td>
<td>sends</td>
</tr>
<tr>
<td>uploads</td>
<td>opens</td>
<td>closes</td>
<td>gives</td>
</tr>
<tr>
<td>defined</td>
<td>visits</td>
<td>changes</td>
<td>receives</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>error</th>
<th>exception</th>
<th>request</th>
<th>message</th>
<th>warning</th>
</tr>
</thead>
<tbody>
<tr>
<td>app</td>
<td>page</td>
<td>application</td>
<td>file</td>
<td>site</td>
</tr>
<tr>
<td>name</td>
<td>images</td>
<td>id</td>
<td>password</td>
<td>address</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

2. Group Precision
Domain-Specific Extraction from Open IE

ReVerb

15m SVO triples

Triples with software entities

5k triples

SVOs extracted directly from StackOverflow

37k triples

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

✗ (people, can read, italian)
✗ (view, looks, south)
KB-LDA versus ReVerb Ranking

**Precision**

- KB-LDA, Best F1=0.73, AUC=0.67
- ReVerb, Best F1=0.72, AUC=0.57
KB-LDA Lessons

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

• Unsupervised: Useful for exploration of new domains (Software)

• Can pre-existing domain knowledge improve KB learning? How much? (Biomedical ontologies)
Roadmap

Statistical Language Model for Software Domain Application

/* comment prediction */

Bootstrap KB Learning for the Biomedical Domain

Grounded Software Ontology Construction

Aligning Grounded and Learned Biomedical Relations

Topic-Model KB Learning

Corpus

KB
Aligning Grounded and Learned Relations

• Evaluation of KB-LDA relations compared to known relations

• Investigation of the potential of grounding
Aligning Grounded and Learned Relations

Grounded

- Proteins and Genes (PRGE)
- Chemicals (CHED)
- Diseases and disorders (DISO)
- Living Beings (LIVB)
- Unified Medical Language System (UMLS)
- ...

NER

autoimmune disease
lupus
cutaneous lupus erythematosus

KB-LDA

CalBC

KB

Proteins and Genes (PRGE)
Chemicals (CHED)
Diseases and disorders (DISO)
Living Beings (LIVB)
Unified Medical Language System (UMLS)
…
Possible Alignment Outcomes

1. Does KB-LDA learn existing relations and concepts?
   – Model Validation

2. Does KB-LDA discover “new” relations and concepts?
   – Added value of language statistics

3. What is missing?
   – Can be added through grounding
Grounded and Learned Entities

**Grounded**
- 70k annotated

**KB-LDA**
- 700k extracted

22k common
Grounded and Learned Entities

Grounded

70k annotated

48k common

KB-LDA

700k extracted
Grounded Entities:
Manual evaluation of sample

Partial or incorrect parse
35/100

- bronchial asthma
- pneumoniae
- beta-1,2-mannotriose

Only in incorrect form
20/35

Limitation of grounding

Average corpus frequency
4.87

Frequent entities (f>10)
12

Potential advantage of grounding
Grounded and Learned Entities

**Grounded**
- 70k annotated

**KB-LDA**
- 700k extracted

- 5k In frequent relations
- 20k
Grounded and Learned Entities

**Grounded**
- 70k annotated

**KB-LDA**
- 700k extracted
- In frequent relations
  - 5k
  - 20k
KB-LDA Entities:
Manual evaluation of sample

Correct 97/100

Added value of language

Parse errors 3/100

Experimental Terminology 16/100

✓ techniques
✓ samples
✓ hapten inhibition experiments
✓ sodium dodecyl sulfate-polyacrylamide gel

Biological Entities and Processes 81/100

✓ linkage
✓ leukotoxin
✓ chemotactic response
✓ plasma cell-associated markers
Discovered Ontology Relations

- Known Relations
- New-Entity Relations
- Unknown Relations
- Inferred Relations

- Grounded Entity (found in Bio Ontologies)
- Learned Entity (found by KB-LDA)

Prevalence vs. Top Tokens graph showing the distribution of different types of relations.
Discovered Ontology Relations

Inferred Relations

54% correct

90% correct relations among bio-entities in manual evaluation!

Prevalence

Known Relations

(Un)Known Relations

New-Entity Relations

Grounded Entity (found in Bio Ontologies)

Learned Entity (found by KB-LDA)

Top Tokens
Discovered Ontology Relations

- **antibody**: EMA (Endomysial autoantibodies) and DCS-6 (Anti-Cyclin D1 antibody)
- **tissue**: bone marrow
- **organ**: airway
Contribution of Model Components to Learning Ontology

![Graph showing the contribution of model components to learning ontology precision over different top token numbers.](image-url)
Contribution of Model Components to Learning General Relations

Ontology

Documents

Relations

Subject
strains
e. coli
bacteria

Verb
produced
demonstrated
showed

Object
infection
response
resistance

lupus

c. lupus
Contribution of Model Components to Learning General Relations

![Graph showing the contribution of model components to learning general relations. The graph plots precision against top tokens, with lines for different models including Full Model, Ont + Rel, Doc + Rel, and Rel. The Full Model shows the highest precision across all top tokens.]
Contribution of Model Components to Learning Intra-Topic Relations

- **Ontology**
- **Documents**
- **Noun Topics**
- **Relations**

Nouns: tuberculosis, malaria, mycobacterium, bcg, plasmodium
Lessons

Grounded to Bio-Ontologies

- Potential of grounding
- Model validation

Learned by KB-LDA

- Added value of language statistics
Roadmap

Statistical Language Model for Software Domain Application

/* comment prediction */

Bootstrap KB Learning for the Biomedical Domain

Grounded Software Ontology Construction

Aligning Grounded and Learned Biomedical Relations

Topic-Model KB Learning

Conclusion
“Grounding entities to specialized data from a scientific domain facilitates improved unsupervised and semi-supervised algorithms for Knowledge Base construction for that domain”
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.

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

p(w_i|h_i)

KB-LDA

NELL

BioNELL

Aligning Grounded and Learned Relations

Statistical Language Models

Grounded Ontology Construction

Corpus

KB

Domain Application

73
What's Next?

Grounded Knowledge Base Learning

Learned Coordinate Terms

Figure 2: What's Next?

Figure 3: Plate Diagram of KB-LDA.

Tables

Revisit application improvement using learned KBs

KB

Domain Application