A Compositional-distributional Semantic Model over Structured Data

André Freitas, Edward Curry
Insight @ NUI Galway

Insight Workshop on Latent Space Methods at UCD
Talking to your Data

André Freitas, Edward Curry
Insight @ NUI Galway
Outline

- Motivation & Context
- Compositional-distributional Semantic Model: Distributional Relational Networks (DRNs)
- Use Case: Question Answering over Linked Data
- Conclusions
Big Data

- Vision: More complete *data-based* picture of the world for systems and users.
Shift in the Database Landscape

- Heterogeneous, complex and large-scale Knowledge Bases (KBs).
- Very-large and dynamic “schemas”.

<table>
<thead>
<tr>
<th>EMP_ID</th>
<th>FIRST_NAME</th>
<th>LAST_NAME</th>
<th>PHONE EXT</th>
<th>HIRE_DATE</th>
<th>DEPT</th>
<th>JOB_C.</th>
<th>JOB_GRP.</th>
<th>JOB_COUNT</th>
<th>SALARY</th>
<th>FULL_NAME</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>Robert</td>
<td>Nelson</td>
<td>290</td>
<td>12/28/1989</td>
<td>12:00</td>
<td>600</td>
<td>VP</td>
<td>2</td>
<td>105,900.00</td>
<td>Nelson, Robert</td>
</tr>
<tr>
<td>4</td>
<td>Bruce</td>
<td>Young</td>
<td>233</td>
<td>12/28/1989</td>
<td>12:00</td>
<td>621</td>
<td>Eng</td>
<td>2</td>
<td>97,500.00</td>
<td>Young, Bruce</td>
</tr>
<tr>
<td>5</td>
<td>Kim</td>
<td>Landolt</td>
<td>25</td>
<td>02/06/1999</td>
<td>12:00</td>
<td>110</td>
<td>Eng</td>
<td>2</td>
<td>102,790.00</td>
<td>Landolt, Kim</td>
</tr>
<tr>
<td>8</td>
<td>Leslie</td>
<td>Johnson</td>
<td>410</td>
<td>04/01/1992</td>
<td>12:00</td>
<td>110</td>
<td>Mktg</td>
<td>3</td>
<td>46,435.00</td>
<td>Johnson, Leslie</td>
</tr>
<tr>
<td>9</td>
<td>Phil</td>
<td>Furstor</td>
<td>229</td>
<td>04/17/1992</td>
<td>12:00</td>
<td>110</td>
<td>Eng</td>
<td>3</td>
<td>75,860.00</td>
<td>Furstor, Phil</td>
</tr>
<tr>
<td>11</td>
<td>K, J</td>
<td>Weinstein</td>
<td>34</td>
<td>01/17/1990</td>
<td>12:00</td>
<td>130</td>
<td>Eng</td>
<td>4</td>
<td>66,292.00</td>
<td>Weinstein, K, J</td>
</tr>
<tr>
<td>12</td>
<td>Cesare</td>
<td>Lee</td>
<td>295</td>
<td>06/01/1992</td>
<td>12:00</td>
<td>200</td>
<td>Eng</td>
<td>6</td>
<td>57,793.00</td>
<td>Lee, Cesare</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>First_Name</th>
<th>Last_Name</th>
<th>Phone Ext</th>
<th>Hire_Date</th>
<th>Dept</th>
<th>Job_C.</th>
<th>Job_GRP.</th>
<th>Job_Count</th>
<th>Salary</th>
<th>Full_Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Robert</td>
<td>Nelson</td>
<td>290</td>
<td>12/28/1989</td>
<td>12:00</td>
<td>600</td>
<td>VP</td>
<td>2</td>
<td>105,900.00</td>
<td>Nelson, Robert</td>
</tr>
<tr>
<td>Bruce</td>
<td>Young</td>
<td>233</td>
<td>12/28/1989</td>
<td>12:00</td>
<td>621</td>
<td>Eng</td>
<td>2</td>
<td>97,500.00</td>
<td>Young, Bruce</td>
</tr>
<tr>
<td>Kim</td>
<td>Landolt</td>
<td>25</td>
<td>02/06/1999</td>
<td>12:00</td>
<td>110</td>
<td>Eng</td>
<td>2</td>
<td>102,790.00</td>
<td>Landolt, Kim</td>
</tr>
<tr>
<td>Leslie</td>
<td>Johnson</td>
<td>410</td>
<td>04/01/1992</td>
<td>12:00</td>
<td>110</td>
<td>Mktg</td>
<td>3</td>
<td>46,435.00</td>
<td>Johnson, Leslie</td>
</tr>
<tr>
<td>Phil</td>
<td>Furstor</td>
<td>229</td>
<td>04/17/1992</td>
<td>12:00</td>
<td>110</td>
<td>Eng</td>
<td>3</td>
<td>75,860.00</td>
<td>Furstor, Phil</td>
</tr>
<tr>
<td>K, J</td>
<td>Weinstein</td>
<td>34</td>
<td>01/17/1990</td>
<td>12:00</td>
<td>130</td>
<td>Eng</td>
<td>4</td>
<td>66,292.00</td>
<td>Weinstein, K, J</td>
</tr>
<tr>
<td>Cesare</td>
<td>Lee</td>
<td>295</td>
<td>06/01/1992</td>
<td>12:00</td>
<td>200</td>
<td>Eng</td>
<td>6</td>
<td>57,793.00</td>
<td>Lee, Cesare</td>
</tr>
</tbody>
</table>

circa 2000
10s-100s attributes
circa 2013
1,000s-1,000,000s attributes
Semantic Heterogeneity

- Decentralized content generation.
- Multiple perspectives (conceptualizations) of the reality.
- Ambiguity, vagueness, inconsistency.
Databases for a Complex World

How do you **query** and do **reasoning** over data on this scenario?
Vocabulary Problem for Databases

Query: Who is the daughter of Bill Clinton married to?

Semantic Gap  Semantic approximation

Possible representations = Commonsense Knowledge

A: PresidentsOfTheUnitedStates
   :type
   :Bill Clinton
      :child
      :spouse
      :Chelsea_Clinton
      :Marc Mezvinsky

B: UnitedStates
   :president
   :Bill Clinton
      :fatherOf
      :husband
      :Chelsea_Clinton
      :Marc Mezvinsky
      ...

C: AmericanPresidents
   :type
   :Bill Clinton
      :numberOfKids
         1
Semantics for a Complex World

- “Most semantic models have dealt with particular types of constructions, and have been carried out under very simplifying assumptions, in true lab conditions.”

- “If these idealizations are removed it is not clear at all that modern semantics can give a full account of all but the simplest sentences.”

Baroni et al. Frege in Space

Formal World

Real World
Compositional-Distributional Semantic Models

- Principled and effective semantic models for coping with real world semantic conditions.
- Based on vector space models.
- Practical way to automatically harvest word “meanings” on a large-scale.
- Focus on semantic approximation.

Applications
- Semantic search.
- Approximate semantic inference.
- Paraphrase detection.
- Semantic anomaly detection.
Addressing the Vocabulary Problem for Databases (with Distributional Semantics)
Solution (Video)
More Complex Queries (Video)
Treo Answers Jeopardy Queries (Video)
Evaluation

- 102 natural language queries (Test Collection: QALD 2011).
- Avg. query execution time: 8.53 s.

Dataset (DBpedia 3.7 + YAGO): 45,767 predicates, 5,556,492 classes and 9,434,677 instances

<table>
<thead>
<tr>
<th>Relevance</th>
<th>Avg. Precision</th>
<th>Avg. Recall</th>
<th>MRR</th>
<th>% of queries answered</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.62</td>
<td>0.81</td>
<td>0.49</td>
<td>80%</td>
</tr>
</tbody>
</table>
## Evaluation

<table>
<thead>
<tr>
<th>System</th>
<th>Avg. R</th>
<th>MAP</th>
<th>% answered queries</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treo</td>
<td>0.79</td>
<td>0.63</td>
<td>79%</td>
</tr>
<tr>
<td>PowerAqua</td>
<td>0.54</td>
<td>0.63</td>
<td>48%</td>
</tr>
<tr>
<td>FREyA</td>
<td>0.48</td>
<td>0.52</td>
<td>54%</td>
</tr>
<tr>
<td>Unger et al.</td>
<td>0.63</td>
<td>0.61</td>
<td>-</td>
</tr>
</tbody>
</table>
Compositional-Distributional Model: Distributional Relational Networks (DRNs)
Distributional Hypothesis

“Words occurring in similar (linguistic) contexts are semantically similar.”

- If we can equate meaning with context, we can simply record the contexts in which a word occurs in a collection of texts (a corpus).
- This can then be used as a surrogate of its semantic representation.
Vector Space Model

function (number of times that the words occur in $c_1$)

$0.7$  

$0.5$  

husband  
spouse  
child  

$c_1$  
$c_2$  
$c_n$
Semantic Similarity/Relatedness

\[ \theta \]

\[ c_1 \]

\[ c_2 \]

\[ c_n \]

husband

spouse

child
Approach Overview

Distributional Relational Network (DRN)

Distributional semantics

Core semantic approximation & composition operations

Structured/logic Knowledge Bases

Querying & Reasoning

Large-scale unstructured data

Commonsense knowledge
Approach Overview

Distributional Relational Network (DRN)

Querying & Reasoning

Core semantic approximation & composition operations

Graph Data (RDF)

Explicit Semantic Analysis (ESA)

Wikipedia

Commonsense knowledge
Approach Overview

**Requirements**
- Ability to cope with:
  - lexical expression differences.
  - abstraction level differences.
  - structural differences.
- Comprehensive semantic matching.
- Performance and scalability.

**Querying & Reasoning**

**Core semantic approximation & composition operations**

**Distributional Relational Networks**

**Structured/logic Knowledge Bases**

**Commonsense knowledge**

**Large-scale unstructured data**
DRN (T-Space)

```prolog
childOf(katehudson, goldiehawn).
childOf(chrisrobinson, stanleyrobinson).
spouse(katehudson, chrisrobinson).
isanActress(goldiehawn).
motherInLaw(A, B) ← spouse(B, C) ∧ childOf(C, A)
```
DRN (T-Space)

\[ \overrightarrow{P}_{VS_{dist}} = \{ \overrightarrow{p} : \overrightarrow{p} = \sum_{i=1}^{t} v_i^p \overrightarrow{c}_i \text{, for each } p \in P \} \]

\[ \overrightarrow{E}_{VS_{dist}} = \{ \overrightarrow{e} : \overrightarrow{e} = \sum_{i=1}^{t} v_i^e \overrightarrow{c}_i \text{, for each } e \in E \} \]

atom vector representation \( \overrightarrow{r} \)

\[ p(e_1) \ (\overrightarrow{p} - \overrightarrow{e_1}) \]

\[ p(e_1, e_2) \ (\overrightarrow{p} - \overrightarrow{e_1}, \overrightarrow{e_2} - \overrightarrow{p}) \]

Space is segmented by the instances
Core Operations

Query

\(< q'_0, q'_1, \ldots, q'_n >\)

DRN

symbol/term space

concept space
Core Operations
Core Principles

- Minimize the impact of Ambiguity, Vagueness, Synonymy.
- Address the simplest matchings first (heuristics).
- Semantic Relatedness as a primitive operation.
- Distributional semantics as commonsense knowledge.
Specificity Ordering

Use specificity (grammatical class + (IDF)) as a heuristic measure

Ambiguity, vagueness, synonymy

Low  High

∀i ∈ [0, n], h_{specificity}(q_i) ≥ h_{specificity}(q_{i+1})
Core Operations

- **Instance search**
  - Proper nouns
  - String similarity + node cardinality

- **Class (unary predicate) search**
  - Nouns, adjectives and adverbs
  - String similarity + Distributional semantic relatedness

- **Property (binary predicate) search**
  - Nouns, adjectives, verbs and adverbs
  - Distributional semantic relatedness

\[
sr(q_1', p_0) \geq \eta
\]
Core Operations

- **Navigation**

  
  \[
  < (\vec{q}'_1 - \vec{p}_1), (\vec{q}'_2 - \vec{p}_2), \ldots, (\vec{q}'_n - \vec{p}_n) >
  \]

- **Extensional expansion**
  - Expands the instances associated with a class.

- **Disambiguation dialog (instance, predicate)**
Use Case: Question
Answering over Linked Data
Question Analysis

Transform natural language queries into triple patterns

“Who is the daughter of Bill Clinton married to?”

Bill Clinton → daughter → married to

(INSTANCE) → (PREDICATE) → (PREDICATE)

PODS

Query Features
Query Plan

Map *query features* into a *query plan*.

A *query plan* contains a sequence of *core operations*.

- (1) INSTANCE SEARCH (Bill Clinton)
- (2) $p_1 < -$ SEARCH PREDICATE (Bill Clinton, daughter)
- (3) $e_1 < -$ NAVIGATE (Bill Clinton, $p_1$)
- (4) $p_2 < -$ SEARCH PREDICATE ($e_1$, married to)
- (5) $e_2 < -$ NAVIGATE ($e_1$, $p_2$)
Instance Search

Query: Bill Clinton -> daughter -> married to

Linked Data: :Bill_Clinton
Predicate Search

Query:

Bill Clinton → daughter → married to

Linked Data:

:Bill_Clinton

:child

:Chelsea_Clinton

:religion

:Baptists

:almaMater

:Yale_Law_School

(PIVOT ENTITY)

(ASSOCIATED TRIPLES)
LinkedList Search

Query: Bill Clinton → daughter → married to

Linked Data:
- :Bill_Clinton
  - :child
    - :Chelsea_Clinton
  - :religion
    - :Baptists
  - :almaMater
    - :Yale_Law_School

Which properties are semantically related to ‘daughter’?

- sem_rel(daughter, child) = 0.054
- sem_rel(daughter, child) = 0.004
- sem_rel(daughter, alma mater) = 0.001
Query: Bill Clinton -> daughter -> married to

Linked Data: :Bill_Clinton -> :child -> :Chelsea_Clinton
Navigate

Query: Bill Clinton \rightarrow daughter \rightarrow married to

Linked Data: :Bill_Clinton \rightarrow :child \rightarrow :Chelsea_Clinton

(PIVOT ENTITY)
Predicate Search

Query:

Bill Clinton -> daughter -> married to

Linked Data:

:Bill_Clinton :child :Chelsea_Clinton :spouse :Mark_Mezvinsky

(PIVOT ENTITY)
"Who is the daughter of Bill Clinton married to?"

Answer

- Chelsea Clinton spouse Marc Mezvinsky
- Bill Clinton child Chelsea Clinton
- Bill Clinton children Chelsea Clinton
- William Jefferson Blythe, Jr. child Bill Clinton
- Virginia Clinton Kelley child Bill Clinton
- Virginia Clinton Kelley children Bill Clinton
"Give me all actors starring in Batman Begins?"

Answer:

- Batman Begins’s starring is Michael Caine
- Batman Begins’s starring is Liam Neeson
- Batman Begins’s starring is Katie Holmes
- Batman Begins’s starring is Gary Oldman
- Batman Begins’s starring is Cillian Murphy
- Batman Begins’s starring is Morgan Freeman
- Batman Begins’s starring is Morgan Freeman
- Batman Begins’s starring is Cillian Murphy
- Batman Begins’s starring is Gary Oldman
- Batman Begins’s starring is Michael Caine
- Batman Begins’s starring is Christian Bale
- Christian Bale’s typo is English-Child Actor.
Random Projection is one solution:

- Estimate a VSM by a random projection matrix consists a set of randomly created vectors.
  - i.e. based on the Johnson-Lindenstrauss lemma
  - verified by the results reported in (Hecht-Nielsen, 1994)

* The above figure is copyrighted by Alex Clemmer (http://nullspace.io/)
• **Application:** Extraction of Technology Terms (term classification)

• **Random Projection**
  - **Data Size:** 10,000 publications
  - **Contexts:** words and their position in the neighbourhood of terms

• **Original Dimension:**
  - approximately 5 million

• **Reducing the dimension to 2000 using Random Projection**

Behrang’s research evolves around classification and finding the optimal contexts in random vector spaces for the extraction of technology terms and their relation. If you are interested please email him at behrangatoffice@gmail.com
Conclusions

- **Distributional Relational Networks (DRNs) provide:**
  - A Knowledge Representation model.
  - Built-in semantic approximation.
  - Associated large-scale commonsense knowledge base (KB) with no manual construction effort.

- **The compositional-distributional model supports a schema-agnostic QA system over a database:**
  - Better recall and query coverage compared to baseline systems.
References

Querying

Reasoning

KR
