Comparison of Hyper-dimensional LSA Spaces for Semantic Differences

John C. Martin
Dissertation Defense
20 May 2016
Overview

• Review LSA model of learning
  – What is meaning?
• Measures
• Experiments
• Semantic Measurement Model
• Q & A
The LSA Model of Learning

- Orthogonal Axes
- Dimensionality Reduction
- Mapping System

Meaning
\[ A_{k} = U_{k} \Sigma_{k} V^{T}_{k} \]
Compositionality Constraint

The meaning of a document is the sum of the meaning of its words

$$D = \frac{q^T U_k}{\sum_k}$$
Compositionality Constraint Corollary

The meaning of a word is defined by the documents in which it appears (and does not appear)
Meaning

The Mapping system consists of:

- Term Vector Dictionary
- Singular Values
Motivation

Documents to Document Centroid

Driving Documents
Objective

Find a measure or set of measures that can quantify the difference between two spaces
Measures

- Direct Comparison
- Projected Content Comparison
- Rotated Item Comparison
Direct Comparison Measures
Individual Space Measures

• Document Count
• Term Count
• Non-zeroes
Distribution Analysis

Document Centroid Cosine Distribution: mm-300-A

Document Centroid Cosine Distribution: mm-1600-A
Projected Content Comparisons

Matched items projected into each space
Projected Item Distribution

Projection Centroid Cosine Distribution:

Projection Item to Item Cosine Distribution:

Project centroids into RTRC-GCat

Project centroids into RTRC-GCat-mod
Three-Tuple Comparisons

\{(A, B, C) \mid A = p_i, B = p_j, C = p_k, \text{where } i \neq j \neq k, \forall p \in P\}
Three-Tuple Relationship Changes

Three-Tuple Changes
Compared to GL 3-A Space Using NICHD04 Projections

- Max
- Min
- Both
- Total Change

GL 6-A
GL 9-A
GL 12-A
GL Col-A

0.00%
5.00%
10.00%
15.00%
20.00%
25.00%
30.00%
35.00%
40.00%
45.00%
50.00%
Rotations and Transform Comparisons
The Transform

\[ A_1 = \text{Project}(A, S_1) \]
\[ A_2 = \text{Project}(A, S_2) \]
\[ A_1^T A_2 = U \Sigma V^T \]
\[ Q = U V^T \]

\[ \| A_1 Q - A_2 \|_F \]
Comparative Space Centroid Analysis
Overlapping Term Vector Norm

\[ \| T_1 Q - T_2 \|_F = \| \hat{T} \|_F = \sqrt{\sum_{i=1}^{|\hat{T}|} \sum_{j=1}^{k} |\hat{t}_{i,j}|^2} \]
<table>
<thead>
<tr>
<th>RTRC GCat</th>
<th>Grade Level Series</th>
<th>GL Non-Overlapping</th>
<th>GL Frozen Vocabulary</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>A</td>
<td>B</td>
<td>C</td>
</tr>
<tr>
<td>Mod</td>
<td>3rd</td>
<td>3rd</td>
<td>3rd</td>
</tr>
<tr>
<td>150k</td>
<td>6th</td>
<td>6th</td>
<td>6th</td>
</tr>
<tr>
<td>150k B</td>
<td>9th</td>
<td>9th</td>
<td>9th</td>
</tr>
<tr>
<td>1k</td>
<td>12th</td>
<td>12th</td>
<td>12th</td>
</tr>
</tbody>
</table>

- **Large 6th**
- **Large 9th**

- **Base**
  - A  | 6th | 6th |
  - B  | 12th | 12th |
## Projection/Anchor Sets

<table>
<thead>
<tr>
<th>Set</th>
<th>Documents</th>
<th>Unique Terms</th>
<th>Term Instances</th>
</tr>
</thead>
<tbody>
<tr>
<td>NICHD04</td>
<td>1,060</td>
<td>5,912</td>
<td>70,063</td>
</tr>
<tr>
<td>T-500</td>
<td>500</td>
<td>16,317</td>
<td>123,668</td>
</tr>
<tr>
<td>T-1000</td>
<td>1,000</td>
<td>24,319</td>
<td>252,372</td>
</tr>
<tr>
<td>T-5000</td>
<td>5,000</td>
<td>49,995</td>
<td>1,281,749</td>
</tr>
</tbody>
</table>
Control Experiment

Three-Tuple Changes vs. OTV-Norm

GCat compared to GCat-Mod

Projection Set

OTV-Norm
General Experiment

OTV-Norm vs. Total Change% and PC Standard Deviation

GCat-150k compared to GCat

- Blue: Sum PC StdDev
- Red: Total Chg%
- Green: OTV-Norm

Projection Set:
- NICHD04
- T-500
- T-1000
- T-5000
General Experiment
Grade Level Series Experiment

Three-Tuple Changes

Compared to GL 3-A Space Using NICHD04 Projections
Grade Level Series OTV-Norm

Comparative OTV-norm in Series
Using NICHD04 Anchors

Comparison Space
Large Volume Experiment

Large Space Comparisons

TC% Between Spaces

Grade Level Compared

- L6 - Series A
- L6 - Series B
- L9 - Series A
- L9 - Series B
Large Volume Experiment

Large Space Comparisons

OTV-Norm Between Spaces

Grade Level Compared

GL 3 | GL 6 | GL 9 | GL 12 | Col

L6 - Series A
L6 - Series B
L9 - Series A
L9 - Series B
Non-overlapping Series Experiment
Non-Overlapping Series OTV-Norm

Non-Overlapping OTV-Norm
Cross Series Comparison - NICHD04

Comparison Space

GL 3-B  GL 6-B  GL 9-B  GL 12-B  GL Col-B
Frozen Vocabulary Experiment

Frozen Vocabulary

TC% vs OTV-Norm

Comparison

Base/6A  Base/6B  Base/12B  6A/12A  6A/12B  12A/12B  6B/12B

Base/6A  Base/6B  Base/12B  6A/12A  6A/12B  12A/12B  6B/12A

TC%  OTV-Norm
OTV-Norm

All Experiment Comparisons

TC% vs OTV-Norm
Semantic Measurement Model

\[ TC\% \approx -0.207882 + 0.0507194 \cdot (OTVNorm) + -0.339339 \cdot (TOR) \]
Linear Regression of TOR and OTV-Norm Across All Comparisons
Linear Regression of TOR and OTV-Norm Across All Comparisons
Summary of Contributions

• Semantic differences are observable
  – Measurable
  – Quality based
• Similarity not dependent on overlapping content
• OTV-Norm & Semantic Measurement Model
  – Whole-space measurement
Further Research

• Refine the model
  – Anchor set selection/influence
  – Account for non-overlapping terms
  – Investigate non-linear model

• Other questions raised
Leverage for Answering Other Questions

• Is it possible to identify key documents that affect the meaning of a space?
• Do additional items added to a space have any impact?
• Is there a point at which adding any items to a space makes no difference?
• Is it possible to identify necessary knowledge that would align two spaces?
Q&A
Backup Slides
Projection of New Content

Text Sources → LSA Space → Mapping Information → Projection
Data

• 42 Spaces
• 592 Comparisons
• 4 Projection Sets
• 4 Anchor Sets
• 26 Measures

61,568 Data Items Collected
Distribution Analysis

Space Distribution Measures

Grade Level Spaces

Cosine

GL-3-A  GL-no-3-A  GL-6-B  GL-no-6-B  GL-9-C  GL-12-A  GL-no-12-A  GL-Col-C

Avg DC Cosine  Avg TC Cosine  Avg DD Cosine