Improving & Better Understanding
Word Vector Representations

Manaal Faruqui
Language Technologies Institute
Carnegie Mellon University
Distributional Hypothesis

“You shall know a word by the company it keeps.”

— Harris (1954); Firth (1957)

I will take what is mine with fire and blood.
...the end battle would be between fire and ice.
My dragons are large and can breathe fire now,
...flame is the visible portion of a fire...
...take place whereby fires can sustain their own heat.
Word Vector Representations

Singular Value Decomposition (Turney & Pantel, 2010)

\[ M = U \times \Sigma \times V^T \]

- Task independent
- Features in NLP models

word2vec (Mikolov et al, 2013)

Neural Language Model (Bengio et al, 2003)
Motivation
Motivation

What resources other than distributional context?
Motivation

What resources other than distributional context?

Can vector dimensions be similar to traditional features?
Motivation

What resources other than distributional context?

Can vector dimensions be similar to traditional features?

Is distributional context necessary to produce word vectors?
Outline

1. Retrofitting word vectors to lexicons
2. Sparsity & Overcompleteness
3. Non-distributional word vectors
Word Vectors Vs. Semantic Lexicons
Word Vectors Vs. Semantic Lexicons

Word vectors!
Word Vectors Vs. Semantic Lexicons

Word vectors!  Word lexicons!
Word Vectors Vs. Semantic Lexicons

Word vectors!

Word lexicons!
Word Vectors + Semantic Lexicons

Word vectors!

Word lexicons!

BOOM!
Word Vectors + Semantic Lexicons

Word vectors!

Word lexicons!
Word Vectors + Semantic Lexicons

Word vectors! → Word lexicons!
Word Vectors + Semantic Lexicons

Word vectors!  →  Retrofitting  ←  Word lexicons!

Word vectors!  ❤  Word lexicons!
Retrofitting

• Incorporates information from lexicons in word vectors
• Post-processing approach
• Applicable to any vector model
• Applicable to any lexicon

https://github.com/mfaruqui/retrofitting
Semantic Lexicons

WordNet (Miller, 1995)
Retrofitting

Distributional vectors

Retrofitted vectors
Retrofitting

- flawed
- incorrect
- untrue
- wrong
- false
- boy
- car

Distributional vectors

Retrofitted vectors
Retrofitting

Distributional vectors

Retrofitted vectors
Retrofitting

boy
car

flawed -> wrong -> untrue
false -> wrong
incorrect -> untrue
flawed
false
incorrect

Distributional vectors
Retrofitted vectors
Retrofitting

Distributional vectors

Retrofitted vectors
Retrofitting

\[ \alpha_i \| q_i - \hat{q}_i \|^2 \]

\[ \beta_{ij} \| q_i - q_j \|^2 \]
Retrofitting

\[ \Psi(Q) = \sum_{i=1}^{n} \left[ \alpha_i \| q_i - r_i \|^2 + \sum_{(i,j) \in E} \beta_{ij} \| q_i - q_j \|^2 \right] \]
Retrofitting

\[ \Psi(Q) = \sum_{i=1}^{n} \left[ \alpha_i \|q_i - r_i\|^2 + \sum_{(i,j) \in E} \beta_{ij} \|q_i - q_j\|^2 \right] \]

Iterative Updates:

\[ q_i = \frac{\sum_{j:(i,j) \in E} \beta_{ij} q_j + \alpha_i r_i}{\sum_{j:(i,j) \in E} \beta_{ij} + \alpha_i} \]
Alternative: Lexicons during Learning

\[ p(Q) \propto \exp \left( -\gamma \sum_{i=1}^{n} \sum_{j: (i,j) \in E} \beta_{ij} \|q_i - q_j\|^2 \right) \]

- Gaussian prior
- Easily differentiable

(Yu & Dredze, ’14; Bian et al, ’14; Xu et al, ’14; Fried & Duh, ‘14)
Retrofitting

- Post-processing operation
- No training of vectors
- No change in objective function
- Applicable to any vector model

(Faruqui et al, NAACL 2015)

Previous Work

- While training
- Requires training of vectors
- Change in objective function
- Redesigning for every vector model

(Yu & Dredze, ’14; Bian et al, ’14; Xu et al, ’14; Fried & Duh, ’14; Rastogi et al 2015)
Graph Construction

<table>
<thead>
<tr>
<th>Lexicons</th>
<th>Nodes</th>
<th>Edges</th>
</tr>
</thead>
<tbody>
<tr>
<td>PPDB</td>
<td>102,902</td>
<td>374,555</td>
</tr>
<tr>
<td>WN_{syn}</td>
<td>148,730</td>
<td>304,856</td>
</tr>
<tr>
<td>WN_{all}</td>
<td>148,730</td>
<td>934,705</td>
</tr>
<tr>
<td>FN</td>
<td>10,822</td>
<td>417,456</td>
</tr>
</tbody>
</table>

PPDB (Ganitkevitch et al, 2013)
Paraphrases

WordNet (Miller, 1995)
Synonyms
Hypernyms, Hyponyms

FrameNet (Baker et al, 1998)
Words evoking frames
# Pre-trained Word Vectors

<table>
<thead>
<tr>
<th>Method</th>
<th>Reference</th>
<th>Corpus size</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glove</td>
<td>Pennington et al, 2014</td>
<td>6 billion</td>
<td>300</td>
</tr>
<tr>
<td>Skip-Gram</td>
<td>Mikolov et al, 2013</td>
<td>100 billion</td>
<td>300</td>
</tr>
<tr>
<td>Global Context</td>
<td>Huang et al, 2012</td>
<td>1 billion</td>
<td>50</td>
</tr>
<tr>
<td>Multi</td>
<td>Faruqui &amp; Dyer, 2014</td>
<td>360 million</td>
<td>512</td>
</tr>
</tbody>
</table>
Evaluation Tasks

Word Similarity Tasks

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>tiger</td>
<td>tiger</td>
<td>10</td>
</tr>
<tr>
<td>king</td>
<td>cabbage</td>
<td>0.2</td>
</tr>
<tr>
<td>love</td>
<td>sex</td>
<td>6.7</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>sugar</td>
<td>approach</td>
<td>0.8</td>
</tr>
</tbody>
</table>

Humans

WS-353 (Finkelstein et al, 2002)
MEN-3k (Bruni et al, 2012),
RG-65 (Rubenstein & Goodenough, 1965)

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>tiger</td>
<td>tiger</td>
<td>1</td>
</tr>
<tr>
<td>king</td>
<td>cabbage</td>
<td>0.3</td>
</tr>
<tr>
<td>love</td>
<td>sex</td>
<td>-0.2</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>sugar</td>
<td>approach</td>
<td>0.4</td>
</tr>
</tbody>
</table>

Vectors

Spearman’s correlation
Evaluation Tasks

Sentiment analysis

The movie sucked big time

I really enjoyed the acting

Dataset: Movie Reviews (Socher et al, 2013)

Features: Average of word vectors of words in the sentence

Model: Logistic regression model with $l_2$ regularization
Word Similarity

Spearman's rank correlation (%)

Skip-Gram

Skip-Gram + PPDB

MEN-3k
Word Similarity

- **Spearman's rank correlation (%)**
  - Skip-Gram
  - Skip-Gram + PPDB

### MEN-3k
- Skip-Gram: 65%
- Skip-Gram + PPDB: 75%

### WS-353
- Skip-Gram: 65%
- Skip-Gram + PPDB: 75%

### RG-65
- Skip-Gram: 65%
- Skip-Gram + PPDB: 80%
Word Similarity

Spearman's rank correlation (%)

<table>
<thead>
<tr>
<th>MEN-3k</th>
<th>SG</th>
<th>SG + WordNet (syn)</th>
</tr>
</thead>
<tbody>
<tr>
<td>60</td>
<td></td>
<td></td>
</tr>
<tr>
<td>65</td>
<td></td>
<td></td>
</tr>
<tr>
<td>70</td>
<td></td>
<td></td>
</tr>
<tr>
<td>75</td>
<td></td>
<td></td>
</tr>
<tr>
<td>80</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>SG</th>
<th>SG + WordNet (all)</th>
</tr>
</thead>
<tbody>
<tr>
<td>60</td>
<td></td>
<td></td>
</tr>
<tr>
<td>65</td>
<td></td>
<td></td>
</tr>
<tr>
<td>70</td>
<td></td>
<td></td>
</tr>
<tr>
<td>75</td>
<td></td>
<td></td>
</tr>
<tr>
<td>80</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

clab
Sentiment Analysis

- Glove
- Glove + PPDB

Accuracy (%)

<table>
<thead>
<tr>
<th>75</th>
<th>76</th>
<th>77</th>
<th>78</th>
<th>79</th>
<th>80</th>
<th>81</th>
<th>82</th>
<th>83</th>
</tr>
</thead>
</table>
Sentiment Analysis

- **Accuracy (%)**
  - **Glove**
  - **Glove + PPDB**
  - **Glove + WordNet (syn)**
  - **Glove + WordNet (all)**

- **Clab**
- **Carnegie Mellon**

Accuracy (%) range:
- 75 to 83
Retrofitting for Multiple Languages

Spearman’s correlation (%)

<table>
<thead>
<tr>
<th>Language</th>
<th>SG</th>
<th>SG + WordNet</th>
</tr>
</thead>
<tbody>
<tr>
<td>German</td>
<td>69</td>
<td>80</td>
</tr>
<tr>
<td>French</td>
<td>58</td>
<td>72</td>
</tr>
<tr>
<td>Spanish</td>
<td>47</td>
<td>58</td>
</tr>
</tbody>
</table>

**Similarity Dataset:**
- German: RG-65 (Gurevych, ’05)
- French: RG-65 (Joubarne & Inkpen, ’11)
- Spanish: MC-30 (Hassan & Mihalcea, ‘09)

**Lexicon:**
- Global WordNet (deMelo & Weikum, ’09)
Retrofitting vs Vector Length

Spearman's rank correlation (%)

Vector length

MEN-3k
Outline

1. Retrofitting word vectors to lexicons
2. Sparsity & Overcompleteness
3. Non-distributional word vectors
Motivation

Can vector dimensions be similar to traditional features?

Cambridge is a beautiful town.
Motivation

Can vector dimensions be similar to traditional features?

Natural language features

High-dimensional

Categorical

Very sparse

Interpretable
Sparse Overcomplete Word Vectors

Dense Vectors

(Faruqui et al, ACL 2015)
Sparse Overcomplete Word Vectors

Dense Vectors

+ Sparsity

(Faruqui et al, ACL 2015)
Sparse Overcomplete Word Vectors

(Daruqui et al, ACL 2015)
Sparse Overcomplete Word Vectors

- + Sparsity
- + Interpretability
- + Better features

(Faruqui et al, ACL 2015)
Sparse Overcomplete Word Vectors

Dense Vectors

- + Sparsity
- + Interpretability
- + Better features
- + Categorical

(Faruqui et al, ACL 2015)
Sparse Overcomplete Word Vectors

\[ \begin{align*}
V & \quad X \\
\downarrow L & \quad \text{Sparse coding} & \quad K >> L \\
\downarrow K & \quad D \\
\downarrow X & \quad A \\
\uparrow K & \quad V \\
\uparrow \text{Loss: } \| X - DA \|_2^2 & & \text{(Lee et al, 2006)}
\end{align*} \]

0: white, +: blue, -: red
Sparse Overcomplete Word Vectors

\[
X = D \cdot A
\]

Sparse coding

Loss: \( \| X - DA \|_2^2 \)

0: white, +: blue, -: red

\( K \gg L \)

\( \| D \|_2^2 \)

\( \| A \|_1 \)

(Lee et al, 2006)
Sparse Non-negative Word Vectors

\[ X = VD \]

\[ \forall i, j \ a_{ij} \geq 0 \]

Loss: \( \| X - DA \|_2^2 \)

Non-negative sparse coding (Hoyer, 2002)
Non-negative sparse vectors (Murphy et al, 2012)
Sparse Binary (Categorical) Word Vectors

\[ V \]

\[ X \]

\[ = \]

\[ K \]

\[ D \]

\[ X \]

\[ V \]

\[ A \]

Loss: \[ \| X - DA \|_2^2 \quad \forall \ i,j \ a_{ij} \in \{0,1\} \]

Mixed-integer bilinear programming
Sparse Binary (Categorical) Word Vectors

\[
X = DV \quad \forall \ i,j \ a_{ij} \in \{0,1\}
\]

Loss: \[ || X - DA ||_2^2 \]

Mixed-integer bilinear programming

\[ V = 100,000, \ K = 1000 \]

#parameters > 100 million
Sparse Binary (Categorical) Word Vectors

\[
V \times \mathbf{X} = V \times D \times \mathbf{X}
\]

\[
\text{Loss: } \| X - DA \|_2^2 \quad \forall \ i,j \ a_{ij} \in \{0,1\}
\]

Mixed-integer bilinear programming

\[
V = 100,000, \ K = 1000
\]

#parameters > 100 million
Sparse Binary (Categorical) Word Vectors

\[ \forall i,j \ a_{ij} \geq 0 \]

\[ \forall i,j \ a_{ij} \in \{0,1\} \]
## Sparse Overcomplete Word Vectors

Hyperparameters tuned on WS-353 task

<table>
<thead>
<tr>
<th>Method</th>
<th>L</th>
<th>K</th>
<th>$l_1$</th>
<th>$l_2$</th>
<th>%Sparse</th>
<th>Non-zero</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glove</td>
<td>300</td>
<td>3000</td>
<td>1</td>
<td>$10^{-5}$</td>
<td>91</td>
<td>270</td>
</tr>
<tr>
<td>Skip-Gram</td>
<td>300</td>
<td>3000</td>
<td>0.5</td>
<td>$10^{-5}$</td>
<td>92</td>
<td>240</td>
</tr>
<tr>
<td>GC</td>
<td>50</td>
<td>500</td>
<td>1</td>
<td>$10^{-5}$</td>
<td>98</td>
<td>10</td>
</tr>
<tr>
<td>Multi</td>
<td>48</td>
<td>960</td>
<td>0.1</td>
<td>$10^{-5}$</td>
<td>93</td>
<td>67</td>
</tr>
</tbody>
</table>
Sentiment Analysis

Glove

Dense
Sparse
Binary

(Faruqui et al, ACL 2015)
Sentiment Analysis

Glove

Skip-Gram

Multi

(Faruqui et al, ACL 2015)
Average Performance

Accuracy (%)

Dense
Sparse
Binary

Glove

(Faruqui et al, ACL 2015)
Average Performance

<table>
<thead>
<tr>
<th>Method</th>
<th>Dense</th>
<th>Sparse</th>
<th>Binary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glove</td>
<td>68</td>
<td>77</td>
<td>80</td>
</tr>
<tr>
<td>Skip-Gram</td>
<td>74</td>
<td>77</td>
<td>80</td>
</tr>
<tr>
<td>Multi</td>
<td>68</td>
<td>68</td>
<td></td>
</tr>
</tbody>
</table>

(Faruqui et al, ACL 2015)
Experiments

Performance vs Overcomplete Vector Length

Glove vectors, $L = 300$ (Faruqui et al, ACL 2015)
Interpreting Sparse Word Vectors

Word Intrusion Detection (Chang et al, 2009; Murphy et al, 2012)

<table>
<thead>
<tr>
<th>hospital</th>
<th>medicine</th>
<th>garden</th>
<th>grass</th>
<th>doctor</th>
<th>fracture</th>
<th>plant</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.5</td>
<td>2.8</td>
<td>0.3</td>
<td>0.01</td>
<td>3.7</td>
<td>9.3</td>
<td>0.1</td>
</tr>
<tr>
<td>0.4</td>
<td>0.05</td>
<td>4.8</td>
<td>7.8</td>
<td>1</td>
<td>0.2</td>
<td>8.0</td>
</tr>
</tbody>
</table>

(Faruqui et al, ACL 2015)
Interpreting Sparse Word Vectors

Word Intrusion Detection (Chang et al, 2009; Murphy et al, 2012)

<table>
<thead>
<tr>
<th>hospital</th>
<th>medicine</th>
<th>garden</th>
<th>grass</th>
<th>doctor</th>
<th>fracture</th>
<th>plant</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.5</td>
<td>2.8</td>
<td>0.3</td>
<td>0.01</td>
<td>3.7</td>
<td>9.3</td>
<td>0.1</td>
</tr>
<tr>
<td>0.4</td>
<td>0.05</td>
<td>4.8</td>
<td>7.8</td>
<td>1</td>
<td>0.2</td>
<td>8.0</td>
</tr>
</tbody>
</table>

Question: \{hospital, medicine, grass, doctor, fracture\}

(Faruqui et al, ACL 2015)
# Interpreting Sparse Word Vectors

<table>
<thead>
<tr>
<th></th>
<th>Acc (%)</th>
<th>IAA</th>
<th>kappa</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dense</td>
<td>57</td>
<td>70</td>
<td>0.40</td>
</tr>
<tr>
<td>Sparse</td>
<td>71</td>
<td>77</td>
<td>0.45</td>
</tr>
</tbody>
</table>

(Faruqui et al, ACL 2015)
Interpreting Sparse Word Vectors

<table>
<thead>
<tr>
<th></th>
<th>Acc (%)</th>
<th>IAA</th>
<th>kappa</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dense</td>
<td>57</td>
<td>70</td>
<td>0.40</td>
</tr>
<tr>
<td>Sparse</td>
<td>71</td>
<td>77</td>
<td>0.45</td>
</tr>
</tbody>
</table>

(Faruqui et al, ACL 2015)
Sparse Word Vectors

(Faruqui et al, ACL 2015)
Sparse Overcomplete Word Vectors

Dense Vectors

- + Sparsity
- + Interpretability
- + Better features
- + Categorical
Outline

1. Retrofitting word vectors to lexicons
2. Sparsity & Overcompleteness
3. Non-distributional word vectors
Non-distributional Word Vectors

WordNet

- flawed
- false
- untrue
- wrong
- correct

Synonym (green)
Antonym (red)

FrameNet

- false
- genuine
- bogus
- Artificiality
Non-distributional Word Representations

<table>
<thead>
<tr>
<th>Features</th>
<th>push</th>
<th>raise</th>
<th>incorrect</th>
<th>correct</th>
<th>fish</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frame: Change_position</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Frame: Artificilaity</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Synonym of: true</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Antonym of: true</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Supersense: ANIMAL</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

(Faruqui & Dyer, ACL 2015)
Non-distributional Word Representations

<table>
<thead>
<tr>
<th>Lexicon</th>
<th>Vocabulary</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>WordNet</td>
<td>10,794</td>
<td>92,117</td>
</tr>
<tr>
<td>Supersense</td>
<td>71,836</td>
<td>54</td>
</tr>
<tr>
<td>FrameNet</td>
<td>9,462</td>
<td>4,221</td>
</tr>
<tr>
<td>Emotion</td>
<td>6,468</td>
<td>10</td>
</tr>
<tr>
<td>Connotation</td>
<td>76,134</td>
<td>12</td>
</tr>
<tr>
<td>Color</td>
<td>14,182</td>
<td>12</td>
</tr>
<tr>
<td>Part of Speech</td>
<td>35,606</td>
<td>20</td>
</tr>
<tr>
<td>Syn. &amp; Ant.</td>
<td>35,693</td>
<td>75,972</td>
</tr>
<tr>
<td>Union</td>
<td>119,257</td>
<td>172,418</td>
</tr>
</tbody>
</table>

(Faruqui & Dyer, ACL 2015)
Non-distributional Word Representations

Accuracy (%)

Glove  |  Skip-Gram  |  Non-distributional

SimLex

Sentiment

(Faruqui & Dyer, ACL 2015)
Non-distributional Word Representations

Competitive

No training!

99.9% sparse

100% interpretable

(Faruqui & Dyer, ACL 2015)
Improvements & Better Understanding
Word Vector Representations

+ Sparsity
+ Interpretability
+ Better features

+ Categorical
+ Semantic lexicons
+ Non-distributional

(Faruqui et al, NAACL 2015; Faruqui et al, ACL 2015; Faruqui & Dyer, ACL 2015)
Data & Tools

http://www.cs.cmu.edu/~mfaruqui/soft.html

Retrofitting:
https://github.com/mfaruqui/retrofitting

Sparse-coding:
https://github.com/mfaruqui/sparse-coding

Non-distributitional:
https://github.com/mfaruqui/non-distributional
Thank you!

Jesse Dodge  Sujay Jauhar  Yulia Tsvetkov  Dani Yogatama

Chris Dyer  Noah A. Smith  Eduard Hovy