Sparse Overcomplete Word Vector Representations

Manaal Faruqui, Yulia Tsvetkov, Dani Yogatama
Chris Dyer & Noah Smith
Outline

1. Motivation
2. Framework
3. Evaluation
4. Interpretation
5. Summary
Motivation

- Task independent
- Features in NLP models

Singular Value Decomposition (Turney & Pantel, 2010)

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

word2vec (Mikolov et al, 2013)

Neural Language Model (Bengio et al, 2003)
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
- Very sparse
- Categorical
- Interpretable
Sparse Overcomplete Word Vectors
Sparse Overcomplete Word Vectors

Dense Vectors

+ Sparsity
Sparse Overcomplete Word Vectors

Dense Vectors

+ Sparsity

+ Interpretability
Sparse Overcomplete Word Vectors

Dense Vectors

+ Sparsity

+ Interpretability

+ Better features
Sparse Overcomplete Word Vectors

Dense Vectors

+ Sparsity
+ Interpretability
+ Better features
+ Categorical (binary)
Outline

1. Motivation ✓
2. Framework ←
3. Evaluation
4. Interpretation
5. Summary
Overcomplete Space

$d$ dimensions

+ Stability

+ Higher order features

+ Sparsity

$k$ dimensions

$k \gg d$

(Olhausen & Field, 1997; Lewicki and Sejnowski, 2000; Dono et al, 2006)
Sparse Overcomplete Word Vectors

Sparse coding

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

0: white, +: blue, -:red

( Lee et al, 2006 )
Sparse Overcomplete Word Vectors

Sparse coding

Loss: $|| X - DA ||_2^2$

0: white, +: blue, -: red

(K >> L)

(LEE ET AL, 2006)
Sparse Non-negative Word Vectors

\[ X = D \times V \]

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

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

\[ V \times X = V \times D \times X \]

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

Mixed-integer bilinear programming

\( V = 100,000, \ K = 1000 \)

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

\[ \begin{align*}
V & \quad X \quad = \quad D \quad X \\
V & \quad = \quad 100,000, \quad K = 1000 \\
\#parameters & \quad > \quad 100 \text{ million}
\end{align*} \]

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

Mixed-integer bilinear programming

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

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

∀ i,j a_{ij} ≥ 0

∀ i,j a_{ij} ∈ \{0,1\}
Previous Work

\[
\begin{align*}
V & \xrightarrow{\text{Threshold}} X \\
V & \xrightarrow{x_{i,j} \geq M^+ \atop x_{i,j} \leq M^-} \text{Threshold} \quad \text{(Guo et al, 2014)}
\end{align*}
\]
Outline

1. Motivation ✅
2. Framework ✅
3. Evaluation ➡️
4. Interpretation
5. Summary
## 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>
### Sparse Overcomplete Word Vectors

<table>
<thead>
<tr>
<th>Method</th>
<th>dense (K)</th>
<th>sparse (L)</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>

Hyperparameters tuned on WS-353 task
# Evaluation Tasks

<table>
<thead>
<tr>
<th>Task</th>
<th>Dataset/Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word Similarity</td>
<td>SimLex-999 (Hill et al, 2014)</td>
</tr>
<tr>
<td>Sentiment Analysis</td>
<td>Movie reviews (Socher et al, 2013)</td>
</tr>
<tr>
<td>Question Classification</td>
<td>TREC (Li &amp; Roth, 2002)</td>
</tr>
<tr>
<td>Text Classification</td>
<td>20-Newsgroups</td>
</tr>
<tr>
<td>NP-bracketing</td>
<td>(Lazaridou, et al 2013)</td>
</tr>
</tbody>
</table>
Evaluation Tasks

**Text classification**

The movie sucked big time

I really enjoyed the acting

• Sentiment Analysis
• Question Classification
• Text Classification

**NP-bracketing**

[[blood pressure] machine]

[local [phone company]]

Classifier

sentence vector

Labels: right, left
Experimental Space

Dense / Sparse / Binary

Evaluation Tasks

Word Vectors
Sentiment Analysis

Accuracy (%)

- Dense
- Sparse
- Binary

Glove
Sentiment Analysis

- **Glove**
  - Dense: 75%
  - Sparse: 83%
  - Binary: 75%

- **Skip-Gram**
  - Dense: 83%
  - Sparse: 83%
  - Binary: 83%

- **Multi**
  - Dense: 75%
  - Sparse: 81%
  - Binary: 81%
NP-bracketing

Accuracy (%)

- Dense
- Sparse
- Binary

Accuracy (%)

Glove

Clab

Carnegie Mellon
NP-bracketing

Accuracy (%)

- Dense
- Sparse
- Binary

Dense: 75, 76, 77, 78, 79, 80, 81, 82, 83
Sparse: 75, 76, 77, 78, 79, 80, 81, 82, 83
Binary: 75, 76, 77, 78, 79, 80, 81, 82, 83

Glove
Skip-Gram
Multi
Average Performance

- **Dense**
- **Sparse**
- **Binary**

Glove
Average Performance

Glove

Skip-Gram

Multi

Dense

Sparse

Binary
Performance vs. Overcomplete Length

Glove vectors, L = 300
## Comparison to Baselines

<table>
<thead>
<tr>
<th>Method</th>
<th>Reference</th>
<th>Avg. Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dense</td>
<td>Pennington et al 2014</td>
<td>76.2</td>
</tr>
<tr>
<td>Naive Thresh</td>
<td>—</td>
<td>75.7</td>
</tr>
<tr>
<td>Average Thresh</td>
<td>Guo et al, 2014</td>
<td>75.8</td>
</tr>
<tr>
<td>Sparse OC</td>
<td>Faruqui et al, 2015</td>
<td><strong>79.4</strong></td>
</tr>
</tbody>
</table>
Outline

1. Motivation ✓
2. Framework ✓
3. Evaluation ✓
4. Interpretation
5. Summary
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>
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\}
## 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>

#annotators = 3
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>

#annotators = 3

<table>
<thead>
<tr>
<th>fracture</th>
<th>wound</th>
<th>tissue</th>
<th>relief</th>
</tr>
</thead>
<tbody>
<tr>
<td>naval</td>
<td>industrial</td>
<td>technological</td>
<td>marine</td>
</tr>
<tr>
<td>stadium</td>
<td>belt</td>
<td>championship</td>
<td>coach</td>
</tr>
<tr>
<td>relationships</td>
<td>connections</td>
<td>identity</td>
<td>relations</td>
</tr>
</tbody>
</table>
Sparse Word Vectors
Outline

1. Motivation
2. Framework
3. Evaluation
4. Interpretation
5. Summary
Sparse Overcomplete Word Vectors

Dense Vectors

- + Sparsity
- + Interpretability
- + Better features
- + Categorical (binary)
Thank you!

Come to my poster:

**Non-distributional** Word Vector Representations
Tuesday Session P2.06 - Plenary Hall B

https://github.com/mfaruqui/sparse-coding
https://github.com/mfaruqui/non-distributional
Baselines for Sparse Vectors

Word co-occurrence matrix

Sparse coding with sparsity

(Murphy et al, 2012)
## Comparison to Baselines

<table>
<thead>
<tr>
<th>Reference</th>
<th>Avg. Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dense</td>
<td>76.2</td>
</tr>
<tr>
<td>Dense + Sparse</td>
<td>72.3</td>
</tr>
<tr>
<td>Naive Thresh</td>
<td>75.7</td>
</tr>
<tr>
<td>Average Thresh</td>
<td>75.8</td>
</tr>
<tr>
<td>Sparse OC</td>
<td>79.4</td>
</tr>
</tbody>
</table>

Pennington et al 2014
Murphy et al, 2012
—
Guo et al, 2014
Faruqui et al, 2015
Evaluation Tasks

Word Similarity Tasks

<table>
<thead>
<tr>
<th>Humans</th>
<th>Vectors</th>
</tr>
</thead>
<tbody>
<tr>
<td>tiger</td>
<td>tiger</td>
</tr>
<tr>
<td>king</td>
<td>cabbage</td>
</tr>
<tr>
<td>love</td>
<td>sex</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>sugar</td>
<td>approach</td>
</tr>
</tbody>
</table>

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