A Simple but Tough-to-beat Baseline for **Sentence Embeddings**

Sanjeev Arora, Yingyu Liang, Tengyu Ma
Princeton University
In submission to ICLR 2017

**Presenter:** Danqi Chen
Word $\rightarrow$ Sentence?

\[
\text{linguistics} = \begin{pmatrix}
0.286 \\
0.792 \\
-0.177 \\
-0.107 \\
0.109 \\
-0.542 \\
0.349 \\
0.271
\end{pmatrix}
\]
Word $\rightarrow$ Sentence

linguistics =

\[
\begin{pmatrix}
0.286 \\
0.792 \\
-0.177 \\
-0.107 \\
0.109 \\
-0.542 \\
0.349 \\
0.271
\end{pmatrix}
\]

Natural language processing is fun. =

\[
\begin{pmatrix}
-0.132 \\
1.129 \\
0.827 \\
0.110 \\
-0.527 \\
0.156 \\
0.349 \\
-0.286
\end{pmatrix}
\]
Sentence embedding

• Compute **sentence similarity** using the inner product:

| S1: Mexico wishes to guarantee citizen’s safety. | Score: 4 (/5) |
| S2: Mexico wishes to avoid more violence. | |

| S1: Iranians Vote in Presidential Election. | Score: 0.4 (/5) |
| S2: Keita Wins Mali Presidential Election. | |
Sentence embedding

• Use as features for **sentence classification** (e.g., sentiment analysis):

  \[
  \text{Natural language processing is fun.} = \begin{pmatrix}
  -0.132 \\
  1.129 \\
  0.827 \\
  0.110 \\
  -0.527 \\
  0.156 \\
  0.349 \\
  -0.286
  \end{pmatrix}
  \]
From Bag-of-words to Complex Models…

- Bag-of-words (BoW)
  
  \[ v(\text{”natural language processing”}) = \frac{1}{3} (v(\text{”natural”}) + v(\text{”language”}) + v(\text{”processing”})) \]
From Bag-of-words to Complex Models…

• Bag-of-words (BoW)

\[ v(\text{"natural language processing"}) = \frac{1}{3} (v(\text{"natural"}) + v(\text{"language"}) + v(\text{"processing"})) \]

• Recurrent neural networks, recursive neural networks, convolutional neural networks..
This paper

- A VERY SIMPLE unsupervised method
- weighted Bag-of-words + remove some special direction
This paper

• A VERY SIMPLE unsupervised method
  • weighted Bag-of-words + remove some special direction

• Step 1:

\[
\text{for all sentence } s \text{ in } S \text{ do}
\]

\[
\upsilon_s \leftarrow \frac{1}{|s|} \sum_{w \in s} \frac{a}{a + p(w)} \upsilon_w
\]

end for
This paper

- A VERY SIMPLE unsupervised method
- weighted Bag-of-words + remove some special direction

Step 1:

\[
\text{for all sentence } s \text{ in } S \text{ do} \\
\quad v_s \leftarrow \frac{1}{|s|} \sum_{w \in s} \frac{a}{a + p(w)} v_w \\
\text{end for}
\]

Step 2:

Compute the first principal component \(u\) of \(\{v_s : s \in S\}\)

\[
\text{for all sentence } s \text{ in } S \text{ do} \\
\quad v_s \leftarrow v_s - uu^\top v_s \\
\text{end for}
\]
A Probabilistic Interpretation

\[ Pr[w_t \mid c_s] = \alpha p(w_t) + (1 - \alpha) \frac{\exp(\nu_{w_t}, b_s)}{Z_{\tilde{c}_s}} \]

Smoothing term:
- \( w_t \) is emitted from background probability (irrelevant to the vector \( c_s \))
- \( w_t \) is emitted according to correlation with the shifted context vector

\[ b_s = \beta c_0 + (1 - \beta) c_s \]

common discourse, often related to syntax
## Results

### sentence similarity

<table>
<thead>
<tr>
<th>Supervised or not</th>
<th>Su.</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>Un.</th>
<th>Se.</th>
<th>Our approach</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>avg-GloVe</td>
<td>tfidf-GloVe</td>
<td>avg-PSL</td>
</tr>
<tr>
<td>Tasks</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>STS’12</td>
<td>58.7</td>
<td>60.0</td>
<td>56.0</td>
<td>48.1</td>
<td>58.4</td>
<td>51.0</td>
<td>46.4</td>
<td>52.5</td>
<td>58.7</td>
</tr>
<tr>
<td>STS’13</td>
<td>55.8</td>
<td>56.8</td>
<td>54.2</td>
<td>44.7</td>
<td>56.7</td>
<td>45.2</td>
<td>41.5</td>
<td>42.3</td>
<td>52.1</td>
</tr>
<tr>
<td>STS’14</td>
<td>70.9</td>
<td>71.3</td>
<td>69.5</td>
<td>57.7</td>
<td>70.9</td>
<td>59.8</td>
<td>51.5</td>
<td>54.2</td>
<td>63.8</td>
</tr>
<tr>
<td>STS’15</td>
<td>75.8</td>
<td>74.8</td>
<td>72.7</td>
<td>57.2</td>
<td>75.6</td>
<td>63.9</td>
<td>56.0</td>
<td>52.7</td>
<td>60.6</td>
</tr>
<tr>
<td>SICK’14</td>
<td>71.6</td>
<td>71.6</td>
<td>70.7</td>
<td>61.2</td>
<td>71.2</td>
<td>63.9</td>
<td>59.0</td>
<td>65.9</td>
<td>69.4</td>
</tr>
<tr>
<td>Twitter’15</td>
<td>52.9</td>
<td>52.8</td>
<td>53.7</td>
<td>45.1</td>
<td>52.9</td>
<td>47.6</td>
<td>36.1</td>
<td>30.3</td>
<td>33.8</td>
</tr>
</tbody>
</table>

### sentence classification

<table>
<thead>
<tr>
<th></th>
<th>PP</th>
<th>DAN</th>
<th>RNN</th>
<th>LSTM (no)</th>
<th>LSTM (o.g.)</th>
<th>skip-thought</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td>similarity (SICK)</td>
<td>84.9</td>
<td>85.96</td>
<td>73.13</td>
<td>85.45</td>
<td>83.41</td>
<td>85.8</td>
<td><strong>86.03</strong></td>
</tr>
<tr>
<td>entailment (SICK)</td>
<td>83.1</td>
<td>84.5</td>
<td>76.4</td>
<td>83.2</td>
<td>82.0</td>
<td>-</td>
<td><strong>84.6</strong></td>
</tr>
<tr>
<td>sentiment (SST)</td>
<td>79.4</td>
<td>83.4</td>
<td>86.5</td>
<td>86.6</td>
<td><strong>89.2</strong></td>
<td>-</td>
<td>82.2</td>
</tr>
</tbody>
</table>
**Results**

### sentence similarity

<table>
<thead>
<tr>
<th>Tasks</th>
<th>PP</th>
<th>PP-proj.</th>
<th>DAN</th>
<th>RNN</th>
<th>iRNN</th>
<th>LSTM (no)</th>
<th>LSTM (o.g.)</th>
<th>ST</th>
<th>Un.</th>
<th>Se.</th>
<th>Un.</th>
<th>Se.</th>
</tr>
</thead>
<tbody>
<tr>
<td>STS’12</td>
<td>58.7</td>
<td>60.0</td>
<td>56.0</td>
<td>48.1</td>
<td>58.4</td>
<td>51.0</td>
<td>46.4</td>
<td>30.8</td>
<td>52.5</td>
<td>58.7</td>
<td>52.8</td>
<td>56.2</td>
</tr>
<tr>
<td>STS’13</td>
<td>55.8</td>
<td>56.8</td>
<td>54.2</td>
<td>44.7</td>
<td>56.7</td>
<td>45.2</td>
<td>41.5</td>
<td>24.8</td>
<td>42.3</td>
<td>52.1</td>
<td>46.4</td>
<td>56.6</td>
</tr>
<tr>
<td>STS’14</td>
<td>70.9</td>
<td>71.3</td>
<td>69.5</td>
<td>57.7</td>
<td>70.9</td>
<td>59.8</td>
<td>51.5</td>
<td>31.4</td>
<td>54.2</td>
<td>63.8</td>
<td>59.5</td>
<td>68.5</td>
</tr>
<tr>
<td>STS’15</td>
<td>75.8</td>
<td>74.8</td>
<td>72.7</td>
<td>57.2</td>
<td>75.6</td>
<td>63.9</td>
<td>56.0</td>
<td>31.0</td>
<td>52.7</td>
<td>60.6</td>
<td>60.0</td>
<td>71.7</td>
</tr>
<tr>
<td>SICK’14</td>
<td>71.6</td>
<td>71.6</td>
<td>70.7</td>
<td>61.2</td>
<td>71.2</td>
<td>63.9</td>
<td>59.0</td>
<td>49.8</td>
<td>65.9</td>
<td>69.4</td>
<td>66.4</td>
<td>72.2</td>
</tr>
<tr>
<td>Twitter’15</td>
<td>52.9</td>
<td>52.8</td>
<td>53.7</td>
<td>45.1</td>
<td>52.9</td>
<td>47.6</td>
<td>36.1</td>
<td>24.7</td>
<td>30.3</td>
<td>33.8</td>
<td>36.3</td>
<td>48.0</td>
</tr>
</tbody>
</table>

### sentence classification

<table>
<thead>
<tr>
<th></th>
<th>PP</th>
<th>DAN</th>
<th>RNN</th>
<th>LSTM (no)</th>
<th>LSTM (o.g.)</th>
<th>skip-thought</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td>similarity (SICK)</td>
<td>84.9</td>
<td>85.96</td>
<td>73.13</td>
<td>85.45</td>
<td>83.41</td>
<td>85.8</td>
<td><strong>86.03</strong></td>
</tr>
<tr>
<td>entailment (SICK)</td>
<td>83.1</td>
<td>84.5</td>
<td>76.4</td>
<td>83.2</td>
<td>82.0</td>
<td>-</td>
<td><strong>84.6</strong></td>
</tr>
<tr>
<td>sentiment (SST)</td>
<td>79.4</td>
<td>83.4</td>
<td>86.5</td>
<td>86.6</td>
<td><strong>89.2</strong></td>
<td>-</td>
<td><strong>82.2</strong></td>
</tr>
</tbody>
</table>

Thanks!