Advanced Recursive Neural Networks

Dr. Kira Radinsky
CTO SalesPredict
Visiting Professor/Scientist Technion

Slides were adapted from lectures by Richard Socher
Recursive Neural Networks

• Focused on compositional representation learning of
  • Hierarchical structure, features and predictions
• Different combinations of:

1. Training Objective

2. Composition Function

3. Tree Structure
Overview

Last lecture: Recursive Neural Networks

This lecture: Different RNN composition functions and NLP tasks

1. Standard RNNs: Paraphrase detection
2. Matrix--Vector RNNs: Relation classification
3. Recursive Neural Tensor Networks: Sentiment Analysis
4. Tree LSTMs: Phrase Similarity
Applications and Models

- Note: All models can be applied to all tasks

- More powerful models are needed for harder tasks

- Models get increasingly more expressive and powerful:
  1. Standard RNNs: Paraphrase detection
  2. Matrix-Vector RNNs: Relation classification
  3. Recursive Neural Tensor Networks: Sentiment Analysis
  4. Tree LSTMs: Phrase Similarity
Paraphrase Detection

Pollack said the plaintiffs failed to show that Merrill and Blodget directly caused their losses.

Basically, the plaintiffs did not show that omissions in Merrill’s research caused the claimed losses.

The initial report was made to Modesto Police December 28.

It stems from a Modesto police report.
How to compare the meaning of two sentences?
RNNs for Paraphrase Detection

Unsupervised RNNs and a pair–wise sentence comparison of nodes in parsed trees (Socher et al., NIPS 2011)
## RNNs for Paraphrase Detection

Experiments on Microsoft Research Paraphrase Corpus (Dolan et al. 2004)

<table>
<thead>
<tr>
<th>Method</th>
<th>Acc.</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rus et al. (2008)</td>
<td>70.6</td>
<td>80.5</td>
</tr>
<tr>
<td>Mihalcea et al. (2006)</td>
<td>70.3</td>
<td>81.3</td>
</tr>
<tr>
<td>Islam et al. (2007)</td>
<td>72.6</td>
<td>81.3</td>
</tr>
<tr>
<td>Qiu et al. (2006)</td>
<td>72.0</td>
<td>81.6</td>
</tr>
<tr>
<td>Fernando et al. (2008)</td>
<td>74.1</td>
<td>82.4</td>
</tr>
<tr>
<td>Wan et al. (2006)</td>
<td>75.6</td>
<td>83.0</td>
</tr>
<tr>
<td>Das and Smith (2009)</td>
<td>73.9</td>
<td>82.3</td>
</tr>
<tr>
<td>Das and Smith (2009) + 18 Surface Features</td>
<td>76.1</td>
<td>82.7</td>
</tr>
<tr>
<td>F. Bu et al. (ACL 2012): String Re--writing Kernel</td>
<td>76.3</td>
<td>—</td>
</tr>
<tr>
<td>Unfolding Recursive Autoencoder (NIPS 2011)</td>
<td><strong>76.8</strong></td>
<td><strong>83.6</strong></td>
</tr>
</tbody>
</table>

Dataset is problematic, a better evaluation is introduced later.
## RNNs for Paraphrase Detection

<table>
<thead>
<tr>
<th>L</th>
<th>Pr</th>
<th>Sentences</th>
<th>Sim. Mat.</th>
</tr>
</thead>
</table>
| P | 0.95 | (1) LLEYTON Hewitt yesterday traded his tennis racquet for his first sporting passion - Australian football - as the world champion relaxed before his Wimbledon title defence  
(2) LLEYTON Hewitt yesterday traded his tennis racquet for his first sporting passion - Australian rules football - as the world champion relaxed ahead of his Wimbledon defence |          |
| P | 0.82 | (1) The lies and deceptions from Saddam have been well documented over 12 years  
(2) It has been well documented over 12 years of lies and deception from Saddam                                                  |          |
| P | 0.67 | (1) Pollack said the plaintiffs failed to show that Merrill and Blodget directly caused their losses  
(2) Basically, the plaintiffs did not show that omissions in Merrill’s research caused the claimed losses                     |          |
| N | 0.49 | (1) Prof Sally Baldwin, 63, from York, fell into a cavity which opened up when the structure collapsed at Tiburtina station. Italian railway officials said  
(2) Sally Baldwin, from York, was killed instantly when a walkway collapsed and she fell into the machinery at Tiburtina station |          |
| N | 0.44 | (1) Bremer, 61, is a onetime assistant to former Secretaries of State William P. Rogers and Henry Kissinger and was ambassador-at-large for counterterrorism from 1986 to 1989  
(2) Bremer, 61, is a former assistant to former Secretaries of State William P. Rogers and Henry Kissinger |          |
| N | 0.11 | (1) The initial report was made to Modesto Police December 28  
(2) It stems from a Modesto police report                                                                                                           |          |
<table>
<thead>
<tr>
<th>L</th>
<th>Pr</th>
<th>Sentences</th>
<th>Sim.Mat.</th>
</tr>
</thead>
</table>
| P | 0.95 | (1) LLEYTON Hewitt yesterday traded his tennis racquet for his first sporting passion - Australian football - as the world champion relaxed before his Wimbledon title defence  
(2) LLEYTON Hewitt yesterday traded his tennis racquet for his first sporting passion - Australian rules football - as the world champion relaxed ahead of his Wimbledon defence |         |
| P | 0.82 | (1) The lies and deceptions from Saddam have been well documented over 12 years  
(2) It has been well documented over 12 years of lies and deception from Saddam  

**A shifted structure** |
| P | 0.67 | (1) Pollack said the plaintiffs failed to show that Merrill and Blodget directly caused their losses  
(2) Basically, the plaintiffs did not show that omissions in Merrill’s research caused the claimed losses |         |
| N | 0.49 | (1) Prof Sally Baldwin, 63, from York, fell into a cavity which opened up when the structure collapsed at Tiburtina station. Italian railway officials said  
(2) Sally Baldwin, from York, was killed instantly when a walkway collapsed and she fell into the machinery at Tiburtina station |         |
| N | 0.44 | (1) Bremer, 61, is a onetime assistant to former Secretaries of State William P. Rogers and **Henry Kissinger** and was ambassador-at-large for countterterrorism from 1986 to 1989  
(2) Bremer, 61, is a former assistant to former Secretaries of State William P. Rogers and Henry Kissinger |         |
| N | 0.11 | (1) The initial report was made to Modesto Police December 28  
(2) It stems from a Modesto police report |         |
Recursive Deep Learning

1. Standard RNNs:  
   Paraphrase Detection

2. Matrix-Vector RNNs:  
   Relation classification

3. Recursive Neural Tensor Networks:  
   Sentiment Analysis

4. Tree LSTMs:  
   Phrase Similarity
Compositionality Through Recursive Matrix---Vector Spaces

\[ p = \tanh(Wc_1 + b) \]

One way to make the composition function more powerful was by untying the weights \( W \).

But what if words act mostly as an operator, e.g. “very” in \textit{very good}.

Proposal: A new composition function
Compositionality Through Recursive Matrix-Vector Recursive Neural Networks

\[ p = \tanh(W \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} + b) \]

\[ p = \tanh(W \begin{pmatrix} c_2 & c_1 \\ c_1 & c_2 \end{pmatrix} + b) \]

Standard NN but the composition depends on the inputs themselves
Predicting Sentiment Distributions

Good example for non-linearity in language
**MV-RNN for Relationship Classification**

### Relationship | Sentence with labeled nouns for which to predict relationships
---|---
**Cause—Effect** (e2, e1) | Avian [influenza]$_{e1}$ is an infectious disease caused by type a strains of the influenza [virus]$_{e2}$.

**Entity—Origin** (e1, e2) | The [mother]$_{e1}$ left her native [land]$_{e2}$ about the same time and they were married in that city.

**Message—Topic** (e2, e1) | Roadside [attractions]$_{e1}$ are frequently advertised with [billboards]$_{e2}$ to attract tourists.

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Feature Sets</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>POS, stemming, syntactic patterns</td>
<td>60.1</td>
</tr>
<tr>
<td>SVM</td>
<td>word pair, words in between</td>
<td>72.5</td>
</tr>
<tr>
<td>SVM</td>
<td>POS, WordNet, stemming, syntactic patterns</td>
<td>74.8</td>
</tr>
<tr>
<td>SVM</td>
<td>POS, WordNet, morphological features, thesaurus, Google n-grams</td>
<td>77.6</td>
</tr>
<tr>
<td>MaxEnt</td>
<td>POS, WordNet, morphological features, noun compound system, thesaurus, Google n-grams</td>
<td>77.6</td>
</tr>
<tr>
<td>SVM</td>
<td>POS, WordNet, prefixes and other morphological features, POS, dependency parse features, Levin classes, PropBank, FrameNet, NomLex-Plus, Google n-grams, paraphrases, TextRunner</td>
<td>82.2</td>
</tr>
<tr>
<td>RNN</td>
<td>-</td>
<td>74.8</td>
</tr>
<tr>
<td>Lin.MVR</td>
<td>-</td>
<td>73.0</td>
</tr>
<tr>
<td>MV-RNN</td>
<td>-</td>
<td>79.1</td>
</tr>
</tbody>
</table>

Not enough training data – helped to add features
Sentiment detection is crucial to business intelligence, stock trading, ...
Sentiment Detection and Bag-of-Words Models

Most methods start with a bag of words + linguistic features/processing/lexica

But such methods (including tf-idf) can’t distinguish:

+ white blood cells destroying an infection
– an infection destroying white blood cells
Sentiment Detection and Bag-of-Words Models

- Sentiment is that sentiment is “easy”
- Detection accuracy for longer documents ~90%
- Lots of easy cases (... horrible ... or ... awesome ...)

- For dataset of single sentence movie reviews (Pang and Lee, 2005) accuracy never reached above 80% for >7 years

- Harder cases require actual understanding of negation and its scope + other semantic effects
Stealing Harvard doesn’t care about cleverness, wit or any other kind of intelligent humor.

There are slow and repetitive parts but it has just enough spice to keep it interesting.
MV- RNN Mistakes

1. Negated Positives

When we say something positive but one word turns it negative, the model cannot weigh that one word strong enough to flip the sentiment of the entire sentence.
The MV-RNN can not recognize that the word “not” lessens the sentiment from negative to neutral.
MV- RNN Mistakes

3. X but Y conjunction

Here the X might be negative BUT if the Y is positive then the model’s sentiment output for the sentence should be positive! MV-RNNs struggle with this.
Two missing pieces for improving sentiment

1. Compositional Training Data

2. Better Compositional model
1. New Sentiment Treebank (using Mturk)
1. New Sentiment Treebank

- Parse trees of 11,855 sentences
- 215,154 phrases with labels
- Allows training and evaluating with compositional information
BeVer Dataset Helped All Models

- Positive/negative full sentence classification

- But hard negation cases are still mostly incorrect
- We also need a more powerful model!
Better Dataset Helped

• This improved performance for full sentence positive/negative classification by 2 – 3 %

• Yay!

• But a more in depth analysis shows: hard negation cases are still mostly incorrect

• We also need a more powerful model!
2. New Compositional Model

• Recursive Neural Tensor Network
• More expressive than previous RNNs
• Idea: Allow more interactions of vectors
2. New Compositional Model

- Recursive Neural Tensor Network

\[ p_2 = g(a, p_1) \]

\[ p_1 = g(b, c) \]

\[ V \in \mathbb{R}^{2d \times 2d \times d} \]

V is a 3rd-order tensor
2. New Compositional Model

- Recursive Neural Tensor Network
Recursive Neural Tensor Network

Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank
Socher et al. 2013

\[ h^{(1)} = \tanh(x^TVx + Wx) \]

---

<table>
<thead>
<tr>
<th>Neural Tensor Layer</th>
<th>Slices of Tensor Layer</th>
<th>Standard Layer</th>
</tr>
</thead>
<tbody>
<tr>
<td>( p = f )</td>
<td>( \begin{bmatrix} b^T \end{bmatrix} V^{[1:2]} b ) + ( W \begin{bmatrix} b \end{bmatrix} )</td>
<td></td>
</tr>
</tbody>
</table>

Diagram:
- \( p_2 = g(a, p_1) \)
- \( p_1 = g(b, c) \)
- Not, Very, Good...
Details: Tensor Backpropagation Training

- Main new matrix derivative needed for a tensor:

\[
\frac{\partial a^T X a}{\partial X} = \frac{\partial a^T X^T a}{\partial X} = aa^T
\]
Details: Tensor Backpropagation Training

- Minimizing cross entropy error:
  \[ E(\theta) = \sum_i \sum_j t_j^i \log y_j^i + \lambda \|\theta\|^2 \]

- Standard softmax error message:
  \[ \delta_{i,s} = (W_s^T (y_i^s - t_i^s)) \otimes f'(x_i^s) \]

- For each slice, we have update:

- Main backprop rule to pass error down from parent:
  \[ \delta_{p_2,\text{down}} = (W^T \delta_{p_2,\text{com}} + S) \otimes f'(\begin{bmatrix} a \\ p_1 \end{bmatrix}) \]
  \[ S = \sum_{k=1}^d \delta_{p_2,\text{com}}^k \left( V[k] + (V[k]^T) \right) \begin{bmatrix} a \\ p_1 \end{bmatrix} \]

- Finally, add errors from parent and current softmax:
  \[ \delta_{p_1,\text{com}} = \delta_{p_1,s} + \delta_{p_2,\text{down}}[d + 1 : 2d] \]
Positive/Negative Results on Treebank

Classifying Sentences: Accuracy improves to 85.4
### Fine Grained Results on Treebank

![Graph showing accuracy over different models](image)

<table>
<thead>
<tr>
<th>Model</th>
<th>Fine-grained</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>All</td>
<td>Root</td>
<td></td>
</tr>
<tr>
<td>NB</td>
<td>67.2</td>
<td>41.0</td>
<td></td>
</tr>
<tr>
<td>SVM</td>
<td>64.3</td>
<td>40.7</td>
<td></td>
</tr>
<tr>
<td>BiNB</td>
<td>71.0</td>
<td>41.9</td>
<td></td>
</tr>
<tr>
<td>VecAvg</td>
<td>73.3</td>
<td>32.7</td>
<td></td>
</tr>
<tr>
<td>RNN</td>
<td>79.0</td>
<td>43.2</td>
<td></td>
</tr>
<tr>
<td>MV-RNN</td>
<td>78.7</td>
<td>44.4</td>
<td></td>
</tr>
<tr>
<td>RNTN</td>
<td><strong>80.7</strong></td>
<td><strong>45.6</strong></td>
<td></td>
</tr>
</tbody>
</table>
Negation Results
**Negation Results**

- Most methods capture that negation often makes things more negative (See Potts, 2010)
- Analysis on negation dataset
- Accuracy:

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>biNB</td>
<td>19.0</td>
</tr>
<tr>
<td>RNN</td>
<td>33.3</td>
</tr>
<tr>
<td>MV-RNN</td>
<td>52.4</td>
</tr>
<tr>
<td>RNTN</td>
<td>71.4</td>
</tr>
</tbody>
</table>

**Negated Positive Sentences: Change in Activation**

- biNB: -0.16
- RNN: -0.34
- MV-RNN: -0.5
- RNTN: -0.57
Results on Negating Negatives

- But how about negating negatives?
- No flips, but positive activation should increase!
Results on Negating Negatives

- Evaluation: Positive activation should increase

<table>
<thead>
<tr>
<th>Model</th>
<th>Negated Positive</th>
<th>Negated Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>biNB</td>
<td>19.0</td>
<td>27.3</td>
</tr>
<tr>
<td>RNN</td>
<td>33.3</td>
<td>45.5</td>
</tr>
<tr>
<td>MV-RNN</td>
<td>52.4</td>
<td>54.6</td>
</tr>
<tr>
<td>RNTN</td>
<td>71.4</td>
<td>81.8</td>
</tr>
<tr>
<td>$n$</td>
<td>Most positive $n$-grams</td>
<td>Most negative $n$-grams</td>
</tr>
<tr>
<td>-----</td>
<td>---------------------------------------------------------------------------------------</td>
<td>---------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>1</td>
<td>engaging; best; powerful; love; beautiful; entertaining; clever; terrific; excellent; great;</td>
<td>bad; dull; boring; fails; worst; stupid; painfully; cheap; forgettable; disaster;</td>
</tr>
<tr>
<td>2</td>
<td>excellent performances; amazing performance; terrific performances; A masterpiece; masterful film; wonderful film; terrific performance; masterful piece; wonderful movie; marvelous performances;</td>
<td>worst movie; bad movie; very bad; shapeless mess; worst thing; tepid waste; instantly forgettable; bad film; extremely bad; complete failure;</td>
</tr>
<tr>
<td>3</td>
<td>an amazing performance; a terrific performance; a wonderful film; wonderful all-ages triumph; A masterful film; a wonderful movie; a tremendous performance; drawn excellent performances; most visually stunning; A stunning piece;</td>
<td>for worst movie; A lousy movie; most joyless movie; a complete failure; another bad movie; fairly terrible movie; a bad movie; extremely unfunny film; most painfully marginal; very bad sign;</td>
</tr>
<tr>
<td>5</td>
<td>nicely acted and beautifully shot; gorgeous imagery; effective performances; the best of the year; a terrific American sports movie; very solid; very watchable; a fine documentary does best; refreshingly honest and ultimately touching;</td>
<td>silliest and most incoherent movie; completely crass and forgettable movie; just another bad movie; ...; A cumbersome and cliche-ridden movie; a humorless, disjointed mess;</td>
</tr>
<tr>
<td>8</td>
<td>one of the best films of the year; simply the best family film of the year; the best film of the year so far; A love for films shines through each frame; created a masterful piece of artistry right here; A masterful film from a master filmmaker; is easily his finest American film ... comes;</td>
<td>A trashy, exploitative, thoroughly unpleasant experience; this sloppy drama is an empty vessel; ...; a meandering, inarticulate and ultimately disappointing film; an unimaginative, nasty, glibly disappointing piece; bad, he's really bad, and; quickly drops on becoming boring and predictable; be the worst special-effects creation of the year;</td>
</tr>
</tbody>
</table>
Visualizing Deep Learning: Word Embeddings
LSTMs

- Remember LSTMs?
- Historically only over temporal sequences

\[
\begin{align*}
    y_1, y_2, y_3, y_4 \\
    x_1, x_2, x_3, x_4 \\
    \tilde{c}_t
\end{align*}
\]

We used

\[
\begin{align*}
    i_t &= \sigma \left( W^{(i)} x_t + U^{(i)} h_{t-1} + b^{(i)} \right), \\
    f_t &= \sigma \left( W^{(f)} x_t + U^{(f)} h_{t-1} + b^{(f)} \right), \\
    o_t &= \sigma \left( W^{(o)} x_t + U^{(o)} h_{t-1} + b^{(o)} \right), \\
    u_t &= \tanh \left( W^{(u)} x_t + U^{(u)} h_{t-1} + b^{(u)} \right), \\
    c_t &= i_t \odot u_t + f_t \odot c_{t-1}, \\
    h_t &= o_t \odot \tanh(c_t),
\end{align*}
\]
Tree LSTMs

- We can use those ideas in grammatical tree structures!
- Paper: Tai et al. 2015: Improved Semantic Representations From Tree---Structured Long Short---Term Memory Networks

- Idea: Sum the child vectors in a tree structure
- Each child has its own forget gate
- Same softmax on h

\[ \tilde{h}_j = \sum_{k \in C(j)} \tilde{h}_k, \]
\[ i_j = \sigma \left( W^{(i)} x_j + U^{(i)} \tilde{h}_j + b^{(i)} \right), \]
\[ f_{jk} = \sigma \left( W^{(f)} x_j + U^{(f)} h_k + b^{(f)} \right), \]
\[ o_j = \sigma \left( W^{(o)} x_j + U^{(o)} \tilde{h}_j + b^{(o)} \right), \]
\[ u_j = \tanh \left( W^{(u)} x_j + U^{(u)} \tilde{h}_j + b^{(u)} \right), \]
\[ c_j = i_j \odot u_j + \sum_{k \in C(j)} f_{jk} \odot c_k, \]
\[ h_j = o_j \odot \tanh(c_j), \]
## Results on Stanford Sentiment Treebank

<table>
<thead>
<tr>
<th>Method</th>
<th>Fine-grained</th>
<th>Binary</th>
</tr>
</thead>
<tbody>
<tr>
<td>RAE (Socher et al., 2013)</td>
<td>43.2</td>
<td>82.4</td>
</tr>
<tr>
<td>MV-RNN (Socher et al., 2013)</td>
<td>44.4</td>
<td>82.9</td>
</tr>
<tr>
<td>RNTN (Socher et al., 2013)</td>
<td>45.7</td>
<td>85.4</td>
</tr>
<tr>
<td>DCNN (Blunsom et al., 2014)</td>
<td>48.5</td>
<td>86.8</td>
</tr>
<tr>
<td>Paragraph-Vec (Le and Mikolov, 2014)</td>
<td>48.7</td>
<td>87.8</td>
</tr>
<tr>
<td>CNN-non-static (Kim, 2014)</td>
<td>48.0</td>
<td>87.2</td>
</tr>
<tr>
<td>CNN-multichannel (Kim, 2014)</td>
<td>47.4</td>
<td>88.1</td>
</tr>
<tr>
<td>DRNN (Irsoy and Cardie, 2014)</td>
<td>49.8</td>
<td>86.6</td>
</tr>
<tr>
<td>LSTM</td>
<td>45.8</td>
<td>86.7</td>
</tr>
<tr>
<td>Bidirectional LSTM</td>
<td>49.1</td>
<td>86.8</td>
</tr>
<tr>
<td>2-layer LSTM</td>
<td>47.5</td>
<td>85.5</td>
</tr>
<tr>
<td>2-layer Bidirectional LSTM</td>
<td>46.2</td>
<td>84.8</td>
</tr>
<tr>
<td>Constituency Tree LSTM (no tuning)</td>
<td>46.7</td>
<td>86.6</td>
</tr>
<tr>
<td>Constituency Tree LSTM</td>
<td><strong>50.6</strong></td>
<td>86.9</td>
</tr>
</tbody>
</table>

No tuning of word vectors
**Semantic Similarity**

- Better than binary paraphrase classification!
- Dataset from a competition:
  SemEval-2014 Task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness [and textual entailment]

<table>
<thead>
<tr>
<th>Relatedness score</th>
<th>Example</th>
</tr>
</thead>
</table>
| 1.6               | A: “A man is jumping into an empty pool”  
|                   | B: “There is no biker jumping in the air”   |
| 2.9               | A: “Two children are lying in the snow and are making snow angels”  
|                   | B: “Two angels are making snow on the lying children”   |
| 3.6               | A: “The young boys are playing outdoors and the man is smiling nearby”  
|                   | B: “There is no boy playing outdoors and there is no man smiling”   |
| 4.9               | A: “A person in a black jacket is doing tricks on a motorbike”  
|                   | B: “A man in a black jacket is doing tricks on a motorbike”   |
## Semantic Similarity Results (correlation and MSE)

<table>
<thead>
<tr>
<th>Method</th>
<th>$r$</th>
<th>$\rho$</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean vectors</td>
<td>0.8046</td>
<td>0.7294</td>
<td>0.3595</td>
</tr>
<tr>
<td>DT-RNN (Socher et al., 2014)</td>
<td>0.7863</td>
<td>0.7305</td>
<td>0.3983</td>
</tr>
<tr>
<td>SDT-RNN (Socher et al., 2014)</td>
<td>0.7886</td>
<td>0.7280</td>
<td>0.3859</td>
</tr>
<tr>
<td>Illinois-LH (Lai and Hockenmaier, 2014)</td>
<td>0.7993</td>
<td>0.7538</td>
<td>0.3692</td>
</tr>
<tr>
<td>UNAL-NLP (Jimenez et al., 2014)</td>
<td>0.8070</td>
<td>0.7489</td>
<td>0.3550</td>
</tr>
<tr>
<td>Meaning Factory (Bjerva et al., 2014)</td>
<td>0.8268</td>
<td>0.7721</td>
<td>0.3224</td>
</tr>
<tr>
<td>ECNU (Zhao et al., 2014)</td>
<td>0.8414</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>LSTM</td>
<td>0.8477</td>
<td>0.7921</td>
<td>0.2949</td>
</tr>
<tr>
<td>Bidirectional LSTM</td>
<td>0.8522</td>
<td>0.7952</td>
<td>0.2850</td>
</tr>
<tr>
<td>2-layer LSTM</td>
<td>0.8411</td>
<td>0.7849</td>
<td>0.2980</td>
</tr>
<tr>
<td>2-layer Bidirectional LSTM</td>
<td>0.8488</td>
<td>0.7926</td>
<td>0.2893</td>
</tr>
<tr>
<td>Constituency Tree LSTM</td>
<td>0.8491</td>
<td>0.7873</td>
<td>0.2852</td>
</tr>
<tr>
<td>Dependency Tree LSTM</td>
<td><strong>0.8627</strong></td>
<td><strong>0.8032</strong></td>
<td><strong>0.2635</strong></td>
</tr>
</tbody>
</table>
Semantic Similarity Results, Pearson Correlation

![Graph showing semantic similarity results with Pearson correlation for different models: DepTree-LSTM, LSTM, Bi-LSTM, and ConstTree-LSTM. The x-axis represents the mean sentence length, while the y-axis represents the correlation coefficient (r). The graph compares the performance of these models across various sentence lengths, with DepTree-LSTM generally showing higher correlation values.]