Compositionality in Semantic Vector Spaces

CS224U: Natural Language Understanding

Feb. 28, 2012

Richard Socher

Joint work with Chris Manning, Andrew Ng
Jeffrey Pennington, Eric Huang and Cliff Lin

More information and code at www.socher.org
Word Vector Space Models

Each word is associated with an n-dimensional vector.

the country of my birth
the place where I was born

But how can we represent the meaning of longer phrases?

By mapping them into the same vector space!
How should we map phrases into a vector space?

Use the principle of compositionality!

The meaning (vector) of a sentence is determined by (1) the meanings of its words and (2) the rules that combine them.

Algorithm jointly learns compositional vector representations (and tree structure).
Outline

Goal: Algorithms that recover and learn semantic vector representations based on recursive structure for multiple language tasks.

1. Introduction

2. Word Vectors and Recursive Neural Networks

3. Recursive Autoencoders for Sentiment Analysis

4. Paraphrase Detection
Distributional Word Representations

France

Monday

In

Germany

France

Monday

Tuesday
There are many well known algorithms that use cooccurrence statistics to compute a distributional representation for words

• (Brown et al., 1992; Turney et al., 2003 and many others).
• LSA (Landauer & Dumais, 1997).
• Latent Dirichlet Allocation (LDA; Blei et al., 2003)

Recent development: “Neural Language models.”
• Bengio et al., (2003) introduced a language model to predict words given previous words which also learns vector representations.
• Collobert & Weston (2008), Maas et al. (2011) from last lecture
Distributional Word Representations

Recent development: “Neural language models”
Collobert & Weston, 2008, Turian et al, 2010
The movie was not really exciting.
The movie was not really exciting.
Recursive Neural Networks for Structure Prediction

Basic computational unit: Recursive Neural Network

Inputs: two candidate children’s representations

Outputs:
1. The semantic representation if the two nodes are merged.
2. Label that carries some information about this node
Recursive Neural Network Definition

\[ p = \text{sigmoid}(W \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} + b), \]

where sigmoid:

\[ label_p = \text{softmax}(W^{label} p) \]
gives a distribution over a set of labels:
Recursive Neural Network Definition

Related Work:

• Previous RNN work (Goller & Küchler (1996), Costa et al. (2003))
  • assumed fixed tree structure and used one hot vectors.
  • No softmax classifiers

• Jordan Pollack (1990): Recursive auto-associative memories (RAAMs)

• Hinton 1990 and Bottou (2011): Related ideas about recursive models.
The movie was not really exciting.
Predicting Sentiment with RNNs

The movie was not really exciting.
Predicting Sentiment with RNNs

\[ p = \text{sigmoid}(W_{c_1} + W_{c_2} + b) \]

\[ \text{label}_p = \text{softmax}(W_{\text{label}_p}) \]

The movie was not really exciting.

\[ p = \frac{1}{1 + e^{-x}} \]
The movie was not really exciting.
The movie was not really exciting.
The movie was not really exciting.
Goal: Algorithms that recover and learn semantic vector representations based on recursive structure for multiple language tasks.

1. Introduction

2. Word Vectors and Recursive Neural Networks

3. Recursive Autoencoders for Sentiment Analysis [Socher et al., EMNLP 2011]

4. Paraphrase Detection
Sentiment detection is crucial to business intelligence, stock trading, …
• Sentiment detection is crucial to business intelligence, stock trading, ...

• 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
Stealing Harvard doesn't care about cleverness, wit or any other kind of intelligent humor.

A film of ideas and wry comic mayhem.
Recursive Autoencoders

Main Idea: A phrase vector is good, if it keeps as much information as possible about its children.
Recursive Autoencoders

- Similar to RNN but with 2 differences: (1) Reconstruction error to keep as much information as possible

\[ p = \text{sigmoid}(W \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} + b) \]
Recursive Autoencoders

- Reconstruction error details

\[
\begin{bmatrix} c'_1; c'_2 \end{bmatrix} = W^{(2)}p + b^{(2)}.
\]

\[
E_{rec}([c_1; c_2]) = \frac{1}{2} \| [c_1; c_2] - [c'_1; c'_2] \|^2
\]
Recursive Autoencoders

- Reconstruction error at every node
- Important detail: normalization

\[ E_{rec}([c_1; c_2]) = \frac{1}{2} \| [c_1; c_2] - [c'_1; c'_2] \|^2 \]

\[ p_2 = f(W[x_1;p_1] + b) \]

\[ p_1 = f(W[x_2;x_3] + b) \]

\[ p = \frac{p}{\|p\|} \]
Recursive Autoencoders

- Similar to RNN but with 2 differences:
  - Tree structure is determined by reconstruction error:
    - does not require a parser
    - get task dependent trees

The movie was not really exciting.
Recursive Autoencoders

The movie was not really exciting.
Recursive Autoencoders

The movie was not really exciting.
Recursive Autoencoders

\[ RAE_\theta(x) = \arg \min_{y \in A(x)} \sum_{s \in T(y)} E_{rec}([c_1; c_2]_s) \]

The movie was not really exciting.
RAE Training

• Lower error over entire sentence $x$ and its label $t$ (+ regularization)

$$J = \frac{1}{N} \sum_{(x,t)} E(x, t; \theta) + \frac{\lambda}{2} \|\theta\|^2.$$

• Error of a sentence is the error at all nodes in its tree:

$$E(x, t; \theta) = \sum_{s \in T(RAE_\theta(x))} E([c_1; c_2]_s, p_s, t, \theta).$$
RAE Training

- Error at each node is a weighted combination of reconstruction error and cross-entropy (distribution likelihood) from softmax classifier

$$\alpha E_{rec}(c_1; c_2; \theta) + (1 - \alpha) E_{CE}(p_s, t; \theta)$$

Reconstruction error

Cross-entropy error

$W^{(1)}$

$W^{(2)}$

$W^{(\text{label})}$
Details for Training RNNs

• Minimizing error by taking gradient steps computed from matrix derivatives

• More efficient implementation via the backpropagation algorithm

• Since we compute derivatives in a tree structure we can, we call it backpropagation *through structure* (Goller et al. 1996)
Accuracy of Positive/Negative Sentiment Classification

- Results on movie reviews (MR) and opinions (MPQA).
- All other methods use hand-designed polarity shifting rules or sentiment lexica.
- RAE: no hand-designed features, learns vector representations for n-grams

<table>
<thead>
<tr>
<th>Method</th>
<th>MR</th>
<th>MPQA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phrase voting with lexicons</td>
<td>63.1</td>
<td>81.7</td>
</tr>
<tr>
<td>Bag of features with lexicons</td>
<td>76.4</td>
<td>84.1</td>
</tr>
<tr>
<td>Tree-CRF (Nakagawa et al. 2010)</td>
<td>77.3</td>
<td>86.1</td>
</tr>
<tr>
<td>RAE (this work)</td>
<td>77.7</td>
<td>86.4</td>
</tr>
</tbody>
</table>
### Sorted Negative and Positive N-grams

<table>
<thead>
<tr>
<th>Most Negative N-grams</th>
<th>Most Positive N-grams</th>
</tr>
</thead>
<tbody>
<tr>
<td>bad; boring; dull; flat; pointless</td>
<td>touching; enjoyable; powerful</td>
</tr>
<tr>
<td>that bad; abysmally pathetic</td>
<td>the beautiful; with dazzling</td>
</tr>
<tr>
<td>is more boring; manipulative and contrived</td>
<td>funny and touching; a small gem</td>
</tr>
<tr>
<td>boring than anything else.; a major waste ... generic</td>
<td>cute, funny, heartwarming; with wry humor and genuine</td>
</tr>
<tr>
<td>loud, silly, stupid and pointless. ; dull, dumb and derivative horror film.</td>
<td>, deeply absorbing piece that works as a; ... one of the most ingenious and entertaining;</td>
</tr>
</tbody>
</table>
Learning Compositionality from Movie Reviews

- Probability of being positive of several n-grams

<table>
<thead>
<tr>
<th>n-gram</th>
<th>P(positive</th>
<th>n-gram)</th>
</tr>
</thead>
<tbody>
<tr>
<td>good</td>
<td>0.45</td>
<td></td>
</tr>
<tr>
<td>not good</td>
<td>0.20</td>
<td></td>
</tr>
<tr>
<td>very good</td>
<td>0.61</td>
<td></td>
</tr>
<tr>
<td>not very good</td>
<td>0.15</td>
<td></td>
</tr>
<tr>
<td>not</td>
<td>0.03</td>
<td></td>
</tr>
<tr>
<td>very</td>
<td>0.23</td>
<td></td>
</tr>
</tbody>
</table>
Vector representations when training only for sentiment

For pdf, see http://www.socher.org/index.php/Main/Semi-SupervisedRecursiveAutoencodersForPredictingSentimentDistributions
Sentiment Distribution Experiments

• Learn distributions over multiple complex sentiments → New dataset and task

• Experience Project
  – http://www.experienceproject.com
  – “I walked into a parked car”
  – Sorry, Hugs; You rock; Tee-hee ; I understand; Wow just wow
  – Over 31,000 entries with 113 words on average
### Sentiment distributions

- **Sorry, Hugs; You rock; Tee-hee ; I understand; Wow just wow**

<table>
<thead>
<tr>
<th>Predicted and Gold Distribution</th>
<th>Anonymous Confession</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="Graph" /></td>
<td>i am a very succesfull business man. i make good money but i have been addicted to crack for 13 years. i moved 1 hour away from my dealers 10 years ago to stop using now i dont use daily but ...</td>
</tr>
<tr>
<td><img src="image2" alt="Graph" /></td>
<td>well i think hairy women are attractive</td>
</tr>
<tr>
<td><img src="image3" alt="Graph" /></td>
<td>Dear Love, I just want to say that I am looking for you. Tonight I felt the urge to write, and I am becoming more and more frustrated that I have not found you yet. I’m also tired of spending so much heart on an old dream. ...</td>
</tr>
</tbody>
</table>
### Sentiment distributions

- **Sorry, Hugs; You rock; Tee-hee ; I understand; Wow just wow**

<table>
<thead>
<tr>
<th>Predicted and Gold Distribution</th>
<th>Anonymous Confession</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image" alt="Graph" /></td>
<td>I loved her but I screwed it up. Now she’s moved on. I’ll never have her again. I don’t know if I’ll ever stop thinking about her.</td>
</tr>
<tr>
<td><img src="image" alt="Graph" /></td>
<td>Could be kissing you right now. I should be wrapped in your arms in the dark, but instead I’ve ruined everything. I’ve piled bricks to make a wall where there never should have been one. I feel an ache that I shouldn’t feel because…</td>
</tr>
<tr>
<td><img src="image" alt="Graph" /></td>
<td>My paper is due in less than 24 hours and I’m still dancing round my room!</td>
</tr>
</tbody>
</table>
Experience Project most votes results

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>20</td>
</tr>
<tr>
<td>Most frequent class</td>
<td>38</td>
</tr>
<tr>
<td>Bag of words; MaxEnt classifier</td>
<td>46</td>
</tr>
<tr>
<td>Spellchecker, sentiment lexica, SVM</td>
<td>47</td>
</tr>
<tr>
<td>SVM on neural net word features</td>
<td>46</td>
</tr>
<tr>
<td>RAE (this work)</td>
<td>50</td>
</tr>
</tbody>
</table>
Experience Project most votes results

Average KL between gold and predicted label distributions:

$$KL(g||p) = \sum_i g_i \log \left( \frac{g_i}{p_i} \right)$$

![Bar chart showing average KL values]

- Avg.Distr: 0.83
- BoW: 0.81
- Features Word Vec: 0.72
- RAE: 0.70
Goal: Algorithms that recover and learn semantic vector representations based on recursive structure for multiple language tasks.

1. Introduction

2. Word Vectors and Recursive Neural Networks

3. Recursive Autoencoders for Sentiment Analysis

4. Paraphrase Detection
   [Socher et al., NIPS 2011]
• 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?
Unsupervised unfolding RAE

\[ E_{rec}(y(i,j)) = \| [x_i; \ldots; x_j] - [x'_i; \ldots; x'_j] \|^2 \]
## Nearest Neighbors of the Unfolding RAE

- More semantic vector representations

<table>
<thead>
<tr>
<th>Center Phrase</th>
<th>RAE</th>
<th>Unfolding RAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>the U.S.</td>
<td>the Swiss</td>
<td>the former U.S.</td>
</tr>
<tr>
<td>suffering low morale</td>
<td>suffering due to no fault of my own</td>
<td>suffering heavy casualties</td>
</tr>
<tr>
<td>advance to the next round</td>
<td>advance to the final of the UNK 1.1 million Kremlin Cup</td>
<td>advance to the semis</td>
</tr>
<tr>
<td>a prominent political figure</td>
<td>the second high-profile opposition figure</td>
<td>a powerful business figure</td>
</tr>
<tr>
<td>conditions of his release</td>
<td>conditions of peace, social stability and political harmony</td>
<td>negotiations for their release</td>
</tr>
</tbody>
</table>
How much can the vectors capture?

<table>
<thead>
<tr>
<th>Encoding Input</th>
<th>Generated Text from Unfolded Reconstruction</th>
</tr>
</thead>
<tbody>
<tr>
<td>a December summit</td>
<td>a December summit</td>
</tr>
<tr>
<td>the first qualifying session</td>
<td>the first qualifying session</td>
</tr>
<tr>
<td>English premier division club</td>
<td>Irish presidency division club</td>
</tr>
<tr>
<td>the safety of a flight</td>
<td>the safety of a flight</td>
</tr>
<tr>
<td>the signing of the accord</td>
<td>the signing of the accord</td>
</tr>
<tr>
<td>the U.S. House of Representatives</td>
<td>the U.S. House of Representatives</td>
</tr>
<tr>
<td>enforcement of the economic embargo</td>
<td>enforcement of the national embargo</td>
</tr>
<tr>
<td>visit and discuss investment possibilities</td>
<td>visit and postpone financial possibilities</td>
</tr>
<tr>
<td>the agreement it made with Malaysia</td>
<td>the agreement it made with Malaysia</td>
</tr>
<tr>
<td>the full bloom of their young lives</td>
<td>the lower bloom of their democratic lives</td>
</tr>
<tr>
<td>the organization for which the men work</td>
<td>the organization for Romania the reform work</td>
</tr>
<tr>
<td>a pocket knife was found in his suitcase in the</td>
<td>a bomb corpse was found in the mission in the Irish</td>
</tr>
<tr>
<td>plane’s cargo hold</td>
<td>car language case</td>
</tr>
</tbody>
</table>
Recursive Autoencoders for Full Sentence Paraphrase Detection

- Unsupervised RAE and a pair-wise sentence comparison of nodes in parsed trees

<table>
<thead>
<tr>
<th>Recursive Autoencoder</th>
<th>Neural Network for Variable-Sized Input</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Recursive Autoencoder Diagram" /></td>
<td><img src="image2.png" alt="Neural Network Diagram" /></td>
</tr>
</tbody>
</table>

- Pairwise Classification Output
- Neural Network
- Variable-Sized Pooling Layer
- Similarity Matrix
Recursive Autoencoders for Full Sentence Paraphrase Detection

- Pooling Operation: Min-Pooling to find close match:
Recursive Autoencoders for Full Sentence 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>All Paraphrase Baseline</td>
<td>66.5</td>
<td>79.9</td>
</tr>
<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><strong>Unfolding Recursive Autoencoder (our method)</strong></td>
<td><strong>76.4</strong></td>
<td><strong>83.4</strong></td>
</tr>
</tbody>
</table>
Recursive Autoencoders for Full Sentence 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                                                                                                    |          |
Recursive Neural Networks for Compositional Vectors

• Questions?

• More information and code at www.socher.org

\[ p = \text{sigmoid}(W \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} + b), \]

\[ \text{label}_p = \text{softmax}(W^{\text{label}} p) \]