Notes on (Socher's) RNN-based models of semantic composition

Chris Potts, Ling 236/Psych 236c: Representations of meaning, Spring 2013

May 23

Overview

These notes attempt to make concrete the models and experiments of Socher et al. 2012 (the MV-RNN model). I also draw on Socher et al. 2011 (the basic RNN model).

1 Relationships between models

Based on slides from Richard[1] and Socher et al. 2012:§2.2:

(1) MV-RNN (Socher et al. 2012):

$$p = f\left(W \begin{bmatrix} Ba \\ Ab \end{bmatrix}\right)$$

- f is sigmoid ($\mathbb{R} \mapsto [0,1]$) or tanh ($\mathbb{R} \mapsto [-1,1]$), applied element-wise.
- W, A, and B all do compositional work.

(2) RNN (Socher et al. 2011):

$$p = f\left(W \begin{bmatrix} I_{n\times n}a \\ I_{n\times n}b \end{bmatrix}\right)$$

- f is sigmoid or tanh.
- A and B are trivialized; W does all of the compositional work.

(3) Mitchell & Lapata 2010:

$$p = Ba + Ab = identity\left([I_{n\times n}I_{n\times n}] \begin{bmatrix} Ba \\ Ab \end{bmatrix}\right)$$

- No transformation to the parent vector.
- No global composition matrix (the identities trivialize it).
- Each daughter contributes a lexical matrix.

(4) Baroni & Zamparelli 2010: A is an adjective matrix, b is a noun vector:

$$p = Ab = identity\left([0_{n\times n}I_{n\times n}] \begin{bmatrix} Ba \\ Ab \end{bmatrix}\right)$$

- No transformation to the parent vector.
- The global composition matrix cancels any contribution made by Ba and is identity on the contribution of Ab.
- The contribution of the adjective is solely its matrix.

[1] `http://www.stanford.edu/class/cs224u/slides/2013-01-21-224U-RichardSocher.pdf`

2 Propositional logic (§3.2)

Socher et al. (2012) present this in terms of the MV-RNN model, but I think the RNN suffices. (To present this model as an MV-RNN, assume that all of the matrices are identities.) I found W by just playing around with equations until I hit a viable solution. Socher et al. used L-BFGS.

(5) $\mathrm{T} = 1$

(6) $\mathrm{F} = 0$

(7) $\neg = 1$

(8) $\wedge = 1$

(9) $W = [1, -1]$

(10) $g(x) = \max(\min(x, 1), 0)$

(11) Negation:

$$g\left(W\begin{bmatrix}1\\1\end{bmatrix}\right) = g(0) = 0$$

$$\neg = 1 \quad \mathrm{T} = 1$$

$$g\left(W\begin{bmatrix}1\\0\end{bmatrix}\right) = g(1) = 1$$

$$\neg = 1 \quad \mathrm{F} = 0$$

(12) Conjunction:

$$g\left(W\begin{bmatrix}1\\0\end{bmatrix}\right) = g(1) = 1$$

$$\mathrm{T} = 1 \qquad g\left(W\begin{bmatrix}1\\1\end{bmatrix}\right) = g(0) = 0$$

$$\wedge = 1 \quad \mathrm{T} = 1$$

$$g\left(W\begin{bmatrix}1\\1\end{bmatrix}\right) = g(0) = 0$$

$$\mathrm{T} = 1 \qquad g\left(W\begin{bmatrix}1\\0\end{bmatrix}\right) = g(1) = 1$$

$$\wedge = 1 \quad \mathrm{F} = 0$$

$$g\left(W\begin{bmatrix}0\\1\end{bmatrix}\right) = g(0) = 0$$

$$\mathrm{F} = 0 \qquad g\left(W\begin{bmatrix}1\\1\end{bmatrix}\right) = g(0)$$

$$\wedge = 1 \quad \mathrm{T} = 1$$

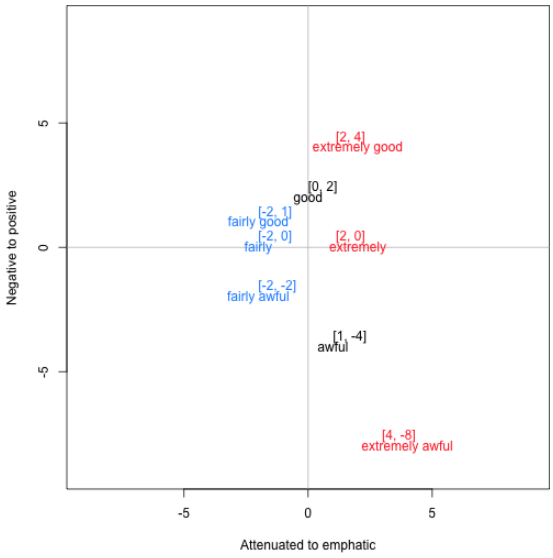
$$g\left(W\begin{bmatrix}1\\1\end{bmatrix}\right) = g(0)$$

$$\mathrm{F} = 0 \qquad g\left(W\begin{bmatrix}1\\0\end{bmatrix}\right) = g(1) = 1$$

$$\wedge = 1 \quad \mathrm{F} = 0$$

Question Could this be done without a global matrix W? If the value of any two nodes of dimension n is a vector of dimension n, then it seems impossible to simulate the two-place nature of binary connectives.

3 Adverb–adjective combinations



$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} [\![extremely]\!]^T + \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} [\![good]\!]^T = [4, -8]$$

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} [\![extremely]\!]^T + \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} [\![awful]\!]^T = [1, -4]$$

$$[\![extremely]\!] = \begin{bmatrix} 8 & 1 \\ 0 & 2 \end{bmatrix}$$

$$[\![extremely]\!]([\![good]\!]^T) = [2, 4]$$

$$[\![extremely]\!]([\![awful]\!]^T) = [4, -8]$$

Figure 1: A toy example to illustrate the effects of matrix multiplication.

Socher et al.'s (2012) adverb–adjective experiments show off the power of the MV-RNN model.[2]

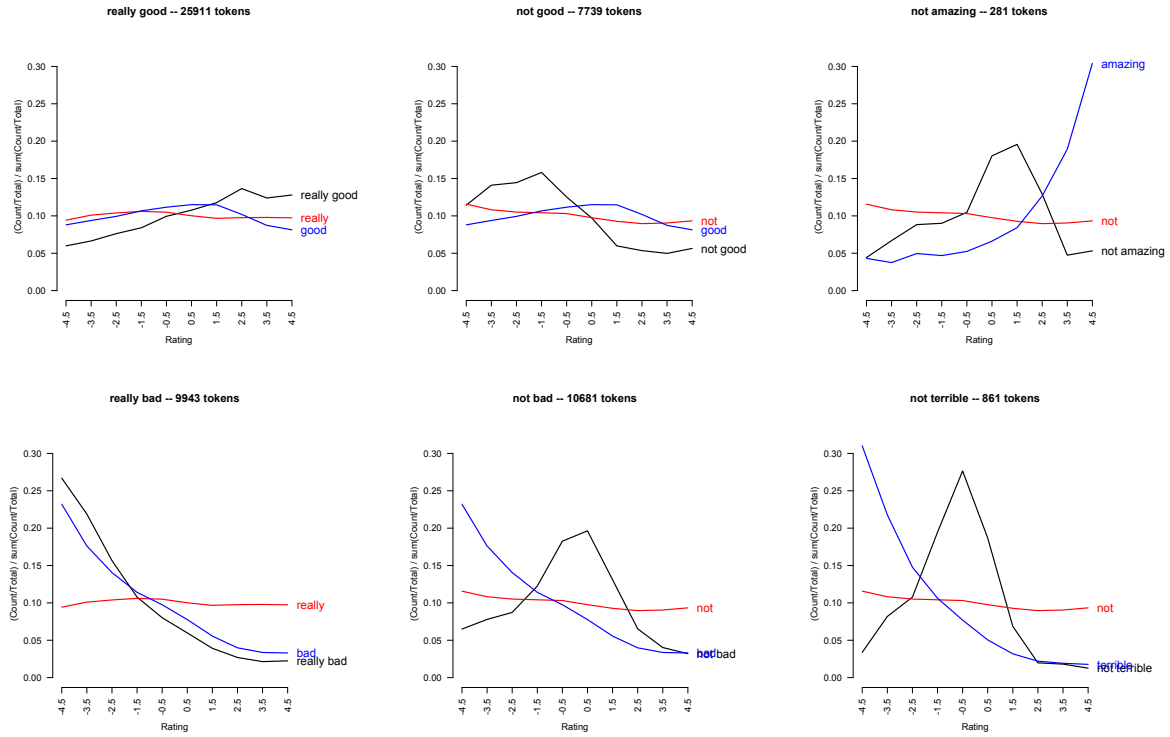


Figure 2: Adverb–adjective distributions.

[2]Data and code: `http://nasslli2012.christopherpotts.net/composition.html`, `http://www.socher.org/index.php/Main/SemanticCompositionalityThroughRecursiveMatrix-VectorSpaces`

Softmax layer To predict the distribution over star-ratings, Socher et al. (2012) train a logistic classifier on the parent vectors, which maps them to $[0, 1]$, a normalized version of the ratings $1 \ldots 10$. This is the technique I described on p. 18 of the handout 'Distributional approaches to word meanings', and it is basically what Baroni et al. (2012) do with their VSMs.

Method	Avg KL
Uniform	0.327
Mean train	0.193
$p = \frac{1}{2}(a + b)$	0.103
$p = a \otimes b$	0.103
$p = [a; b]$	0.101
$p = Ab$	0.103
RNN	0.093
Linear MVR	0.092
MV-RNN	**0.091**

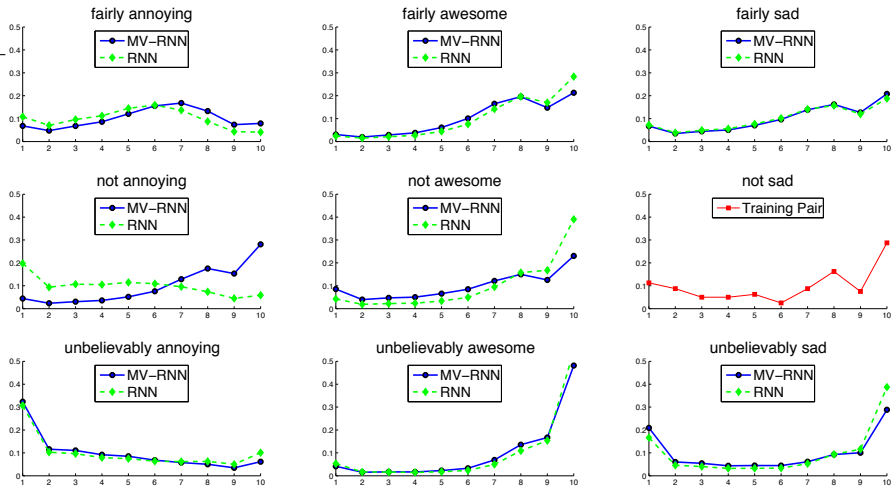


Figure 3: **Left**: Average KL-divergence for predicting sentiment distributions of unseen adverb-adjective pairs of the test set. See text for p descriptions. Lower is better. The main difference in the KL divergence comes from the few negation pairs in the test set. **Right**: Predicting sentiment distributions (over 1-10 stars on the x-axis) of adverb-adjective pairs. Each row has the same adverb and each column the same adjective. Many predictions are similar between the two models. The RNN and linear MVR are not able to modify the sentiment correctly: *not awesome* is more positive than *fairly awesome* and *not annoying* has a similar shape as *unbelievably annoying*. Predictions of the linear MVR model are almost identical to the standard RNN for these examples.

Figure 3: From Socher et al. (2012), p. 1206.

References

Baroni, Marco, Raffaella Bernardi, Ngoc-Quynh Do & Chung-chieh Shan. 2012. Entailment above the word level in distributional semantics. In *Proceedings of the 13th conference of the European chapter of the association for computational linguistics*, 23–32. Avignon, France: ACL.

Baroni, Marco & Roberto Zamparelli. 2010. Nouns are vectors, adjectives are matrices: Representing adjective–noun constructions in semantic space. In *Proceedings of the 2010 conference on empirical methods in natural language processing* EMNLP '10, 1183–1193. Stroudsburg, PA: Association for Computational Linguistics.

Mitchell, Jeff & Mirella Lapata. 2010. Composition in distributional models of semantics. *Cognitive Science* 34(8). 1388–1429.

Socher, Richard, Brody Huval, Christopher D. Manning & Andrew Y. Ng. 2012. Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of the 2012 conference on empirical methods in natural language processing*, 1201–1211. Stroudsburg, PA.

Socher, Richard, Jeffrey Pennington, Eric H. Huang, Andrew Y. Ng & Christopher D. Manning. 2011. Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Proceedings of the 2011 conference on empirical methods in natural language processing*, 151–161. Edinburgh, Scotland, UK.: ACL.