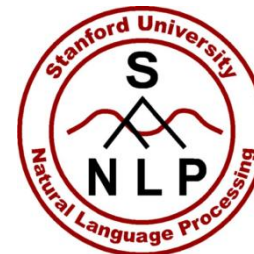# Holistic Compositionality in Semantic Vector Spaces

Semantic Representations for Textual Inference
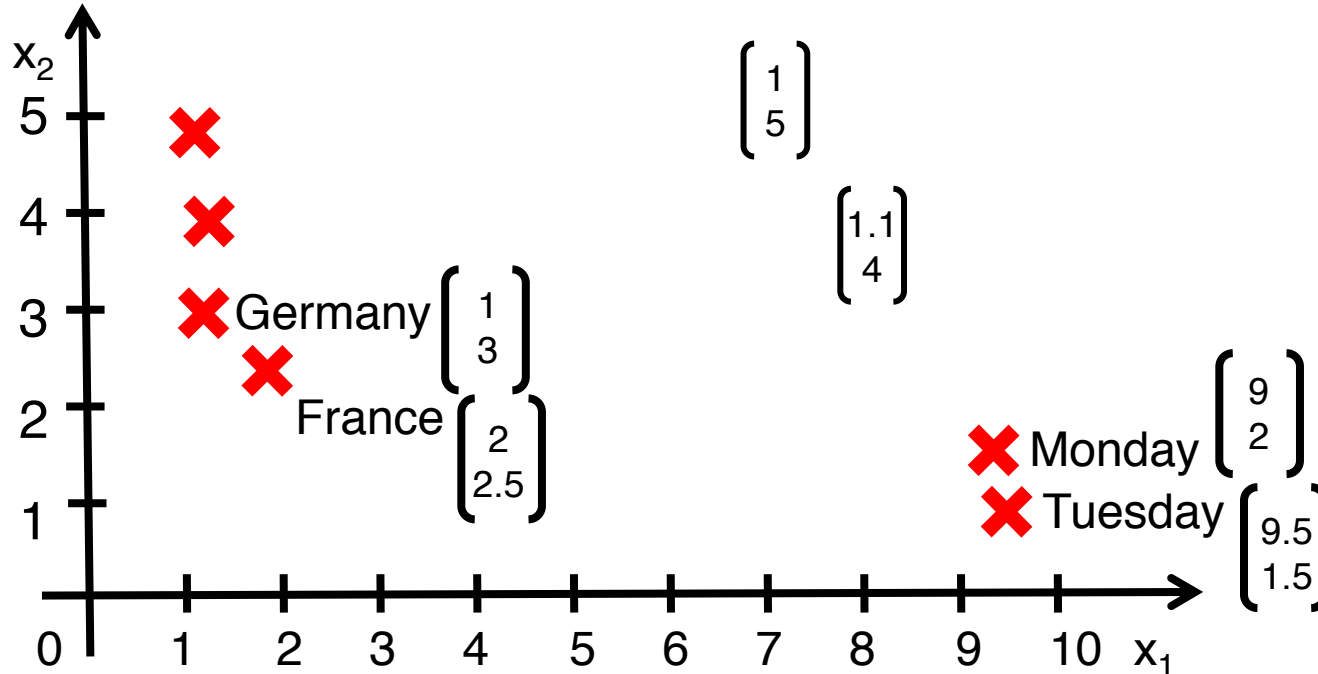## March 10, 2012

Richard Socher

Joint work with Andrew Ng and Chris Manning

Word Vector Space Models

Each word is associated with an n-dimensional vector.

the country of my birth
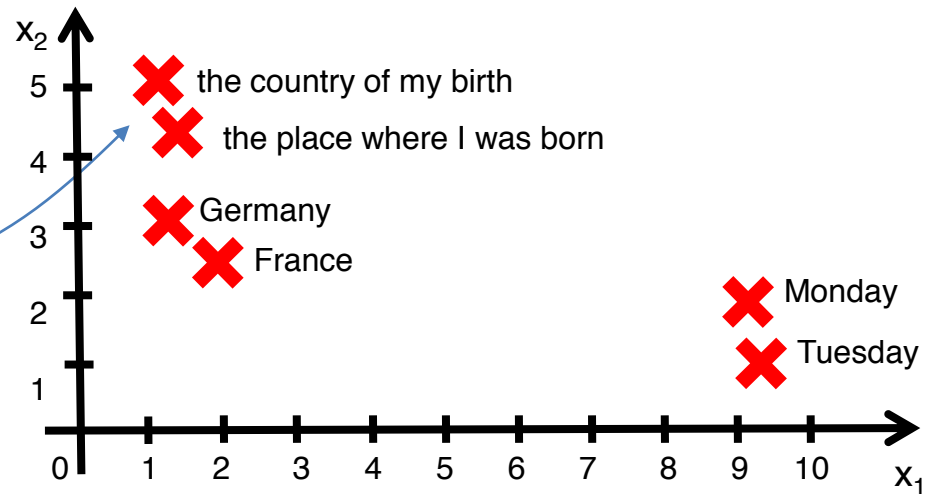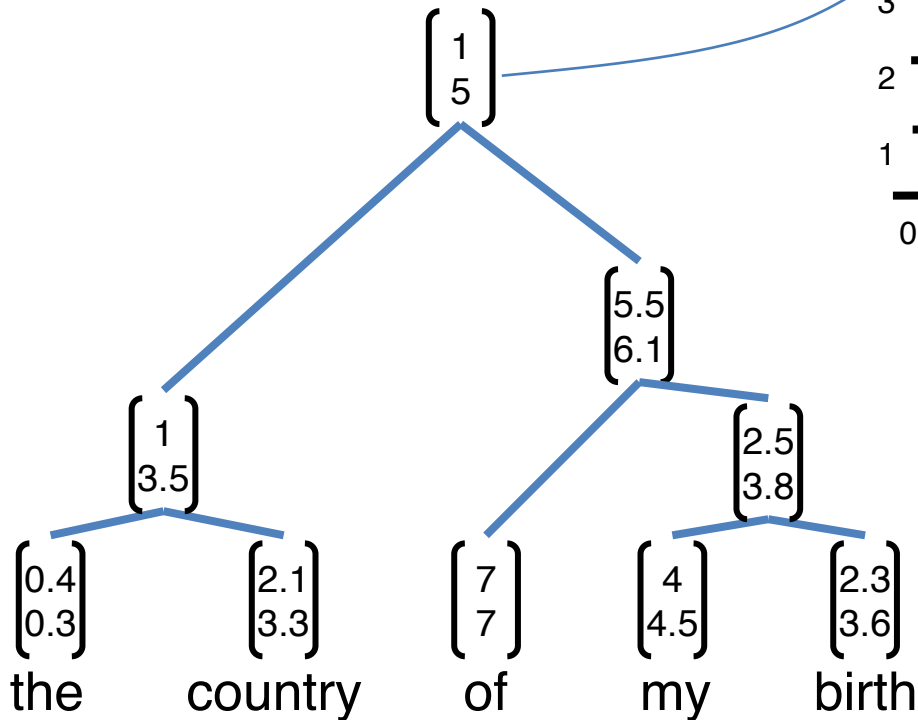 the place where I was born

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
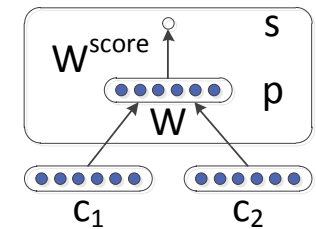(2) the rules that combine them.

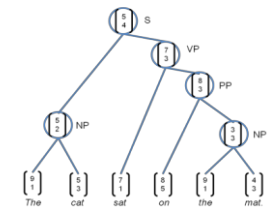Algorithm jointly learns compositional vector representations (and tree structure).

Goal: Algorithms that recover and learn semantic vector representations based on recursive structure for multiple language tasks.
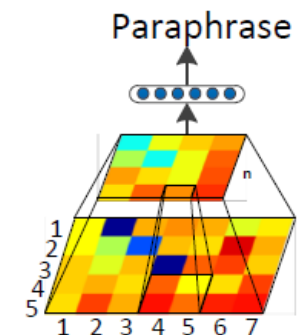
Distributional Word Representations
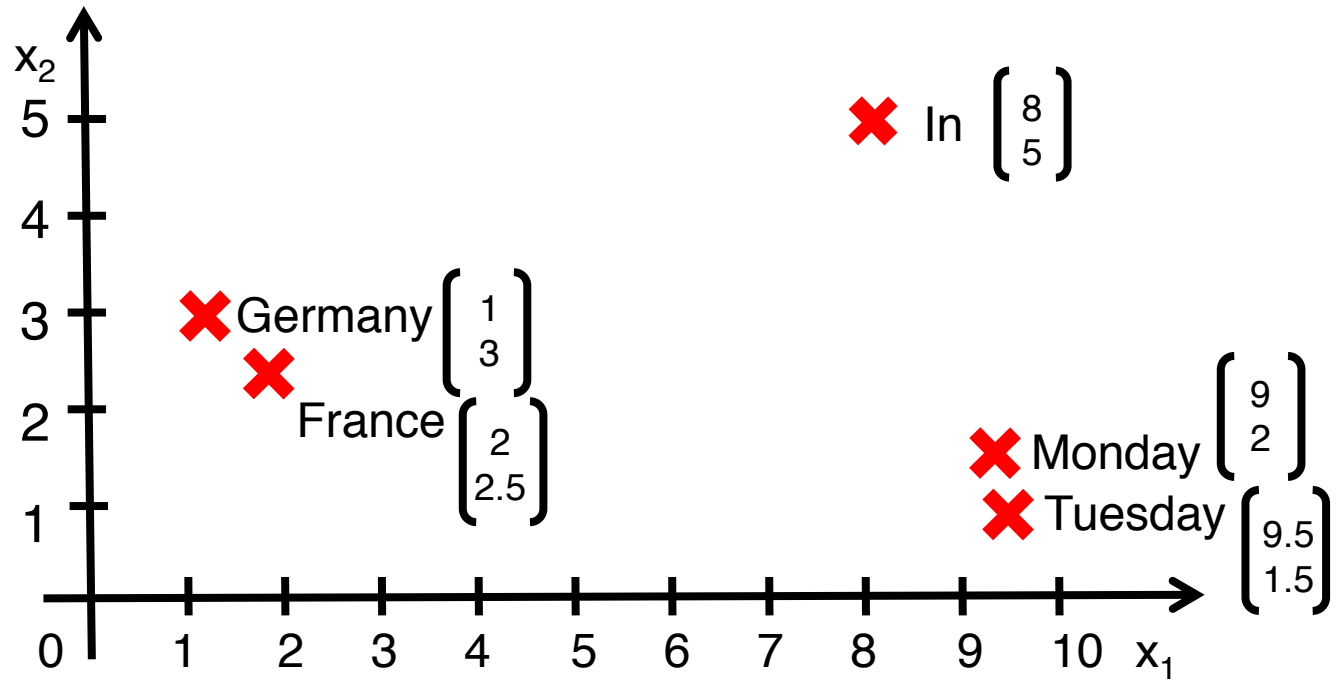
There are many well known algorithms that use cooccurrence statistics to compute a distributional representation for words
- (Brown et al., 1992; Turney et al., 2003 and many others).
- LSA (Landauer & Dumais, 1997).
- Latent Dirichlet Allocation (LDA; Blei et al., 2003)

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

Recent development: "Neural language models"
Collobert & Weston, 2008, Turian et al, 2010

Vectorial Sentence Meaning - Step 1: Parsing
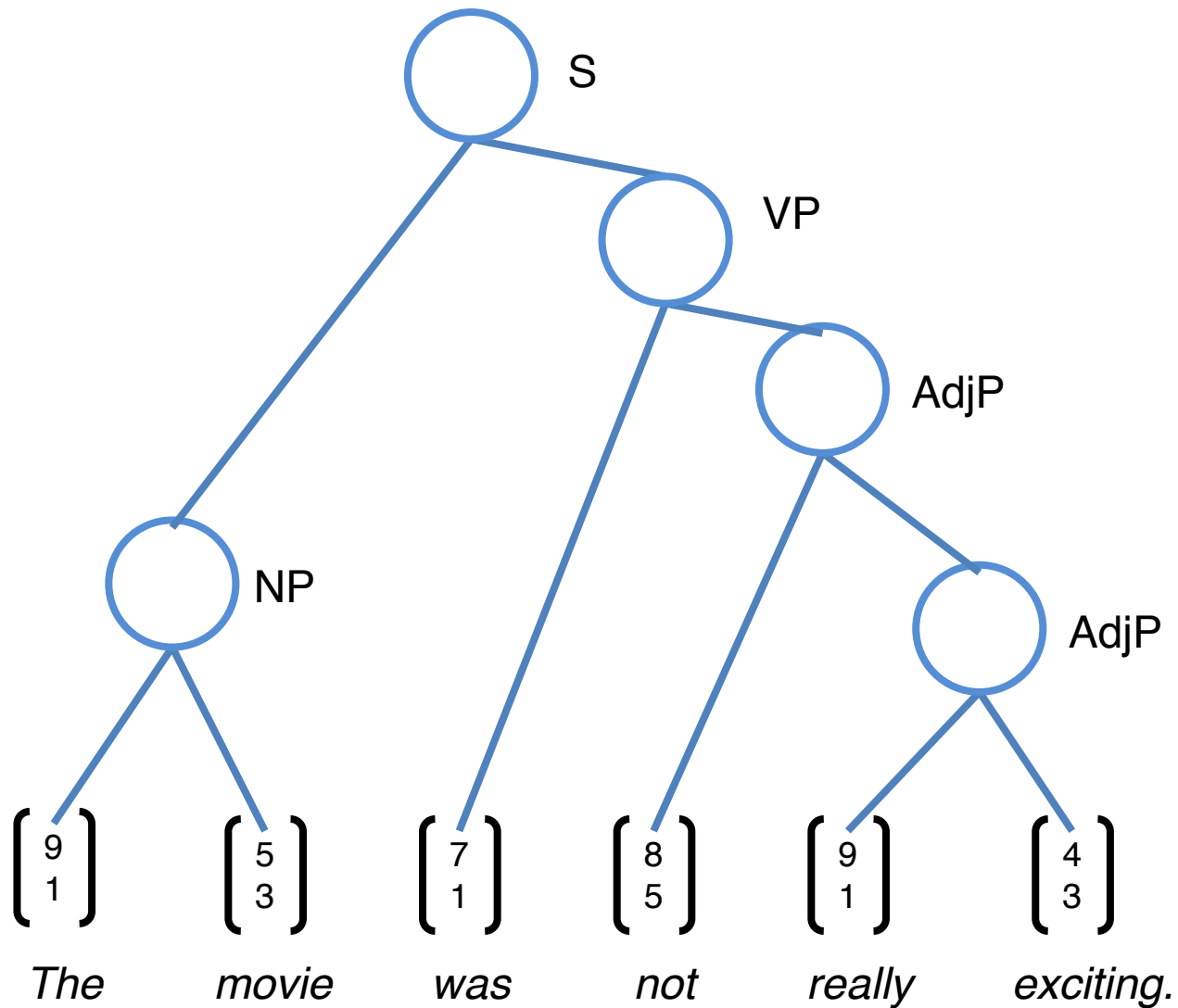
Vectorial Sentence Meaning - Step 2: Vectors at each node
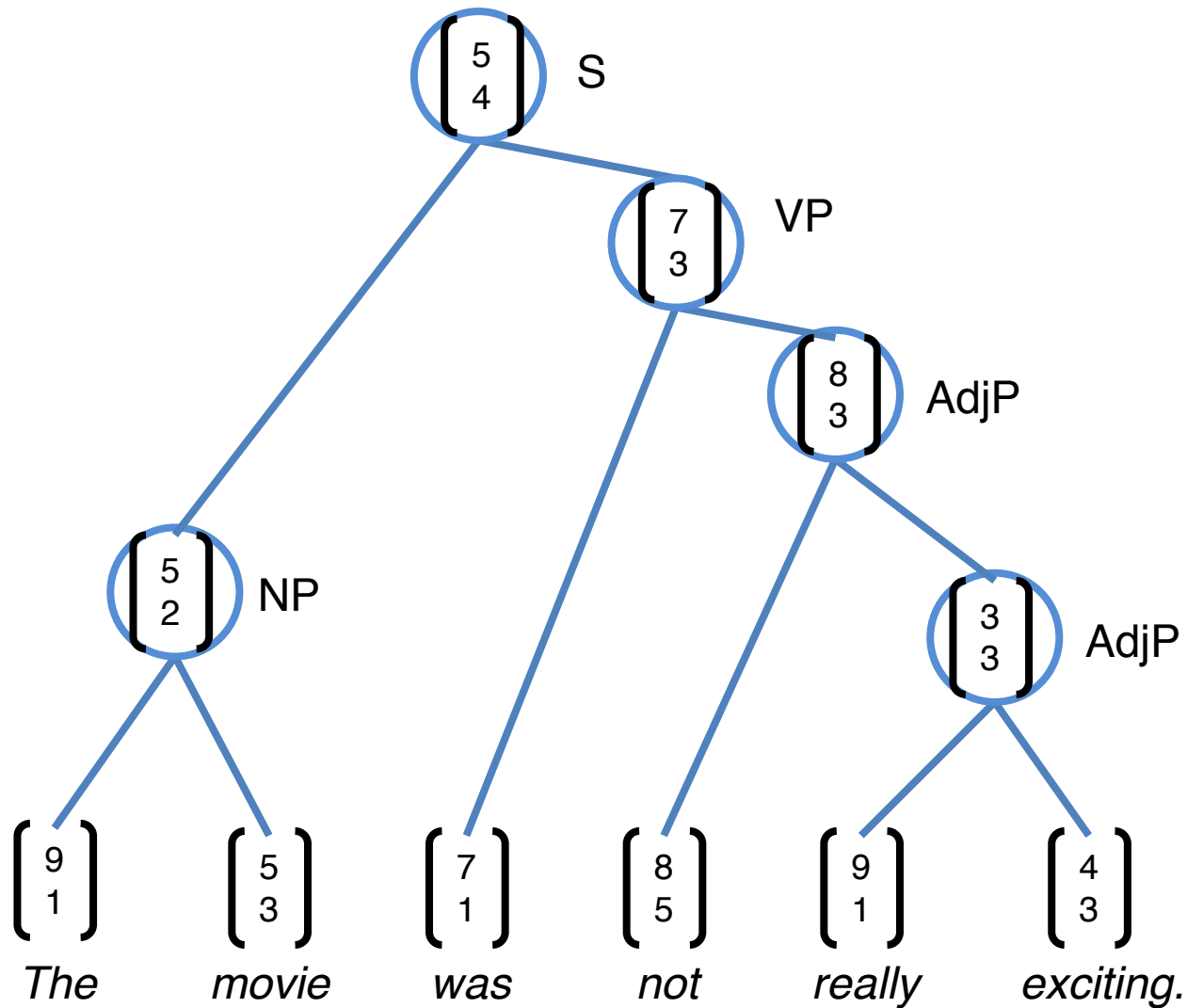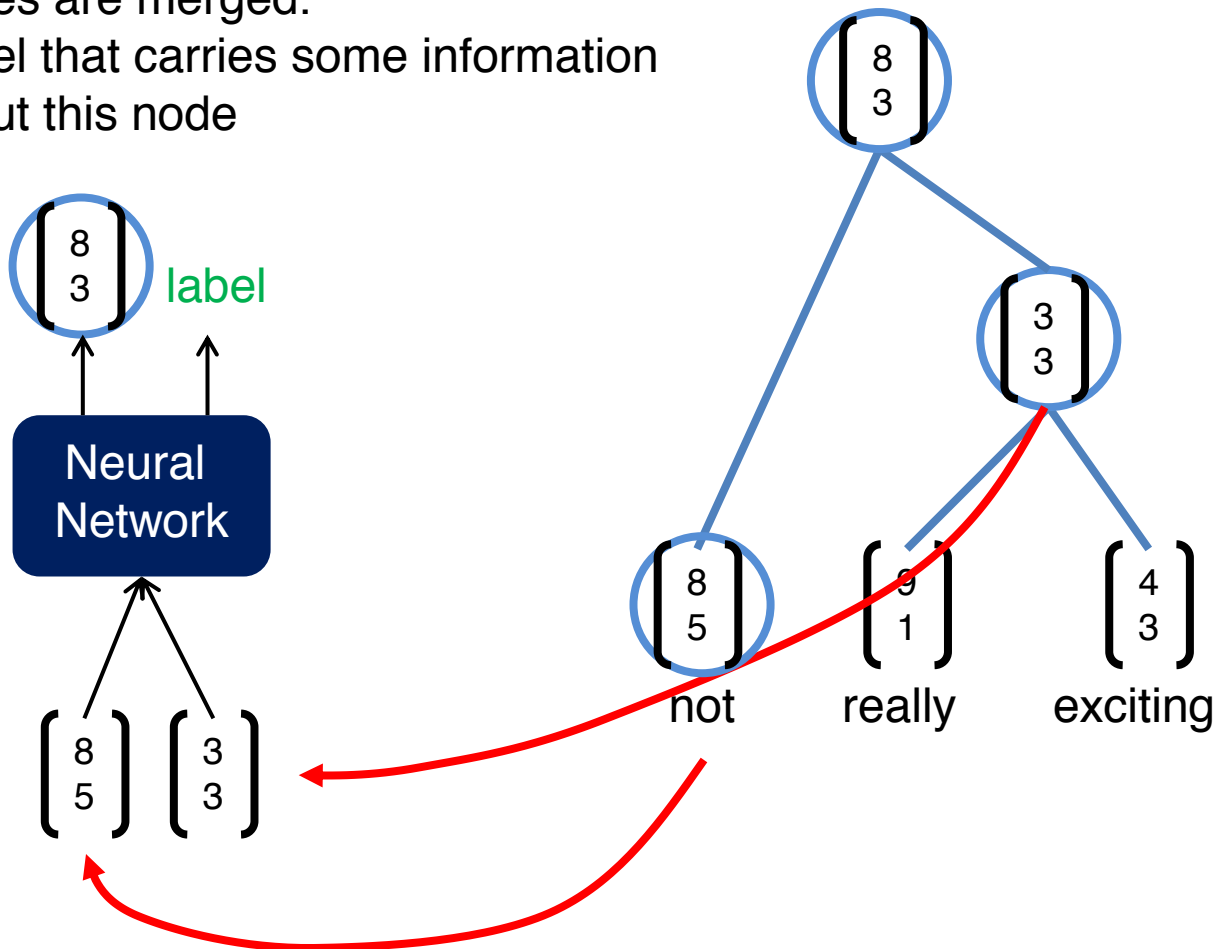
Recursive Neural Networks for Structure Prediction

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

label

Neural Network

not   really   exciting

Recursive Neural Network Definition
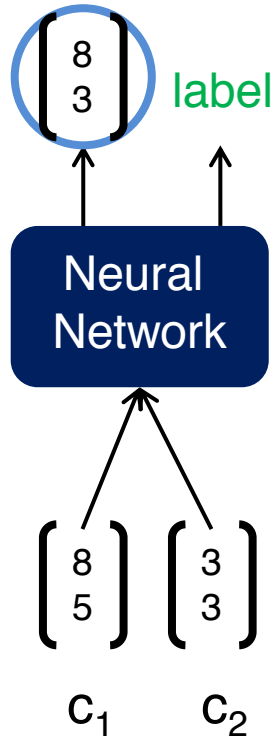
$$p = \text{sigmoid}\left(W \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} + b\right),$$

where sigmoid:
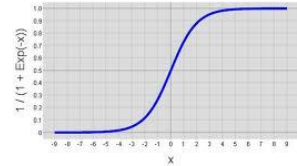
$$label_p = softmax(W^{label} p)$$

gives a distribution over a set of labels:

you rock (3)   teehee (0)   I understand (6)   sorry, hugs (1)   wow, just wow (0)
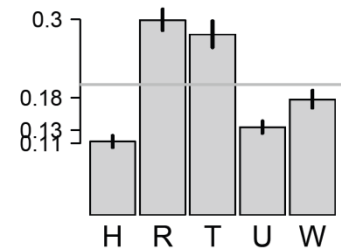
Recursive Neural Network Definition
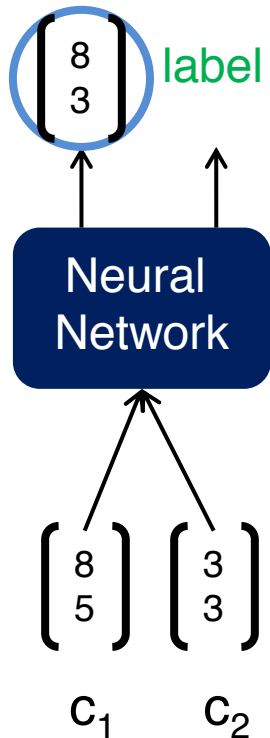
Related Work:

- 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.

Goal: Predict Pos/Neg Sentiment of Full Sentence
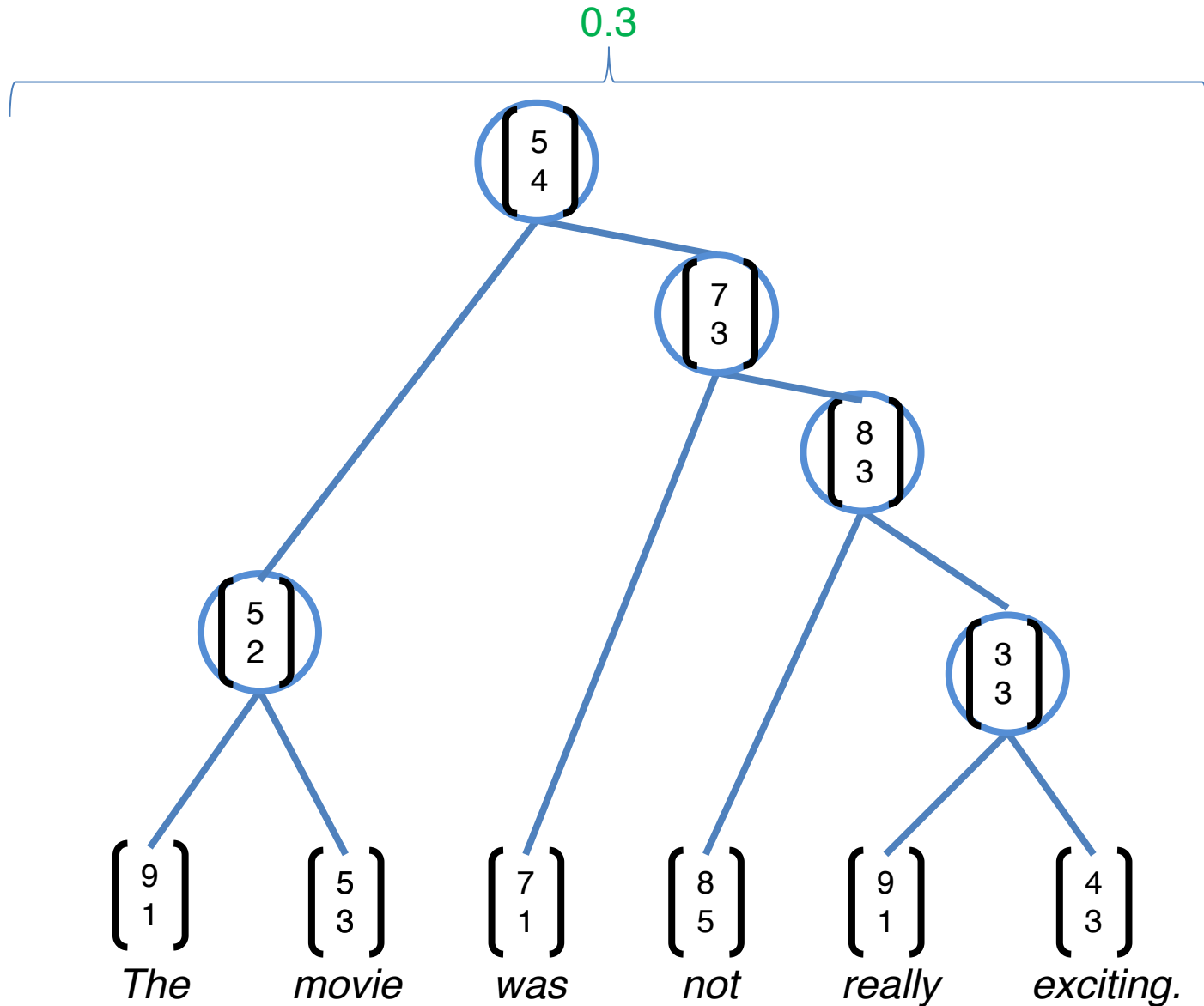
Predicting Sentiment with RNNs
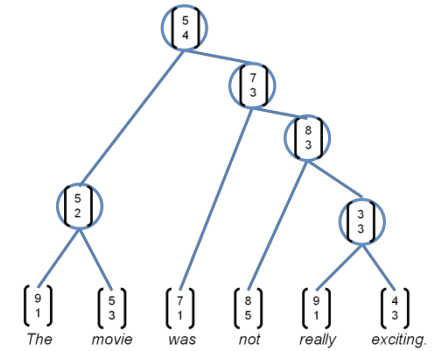
0.5	0.5	0.5	0.3	0.5	0.7
$\begin{pmatrix} 9 \\ 1 \end{pmatrix}$	$\begin{pmatrix} 5 \\ 3 \end{pmatrix}$	$\begin{pmatrix} 7 \\ 1 \end{pmatrix}$	$\begin{pmatrix} 8 \\ 5 \end{pmatrix}$	$\begin{pmatrix} 9 \\ 1 \end{pmatrix}$	$\begin{pmatrix} 4 \\ 3 \end{pmatrix}$
The	*movie*	*was*	*not*	*really*	*exciting.*

Predicting Sentiment with RNNs

$$p \;=\; sigmoid\left(W \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} + b\right)$$
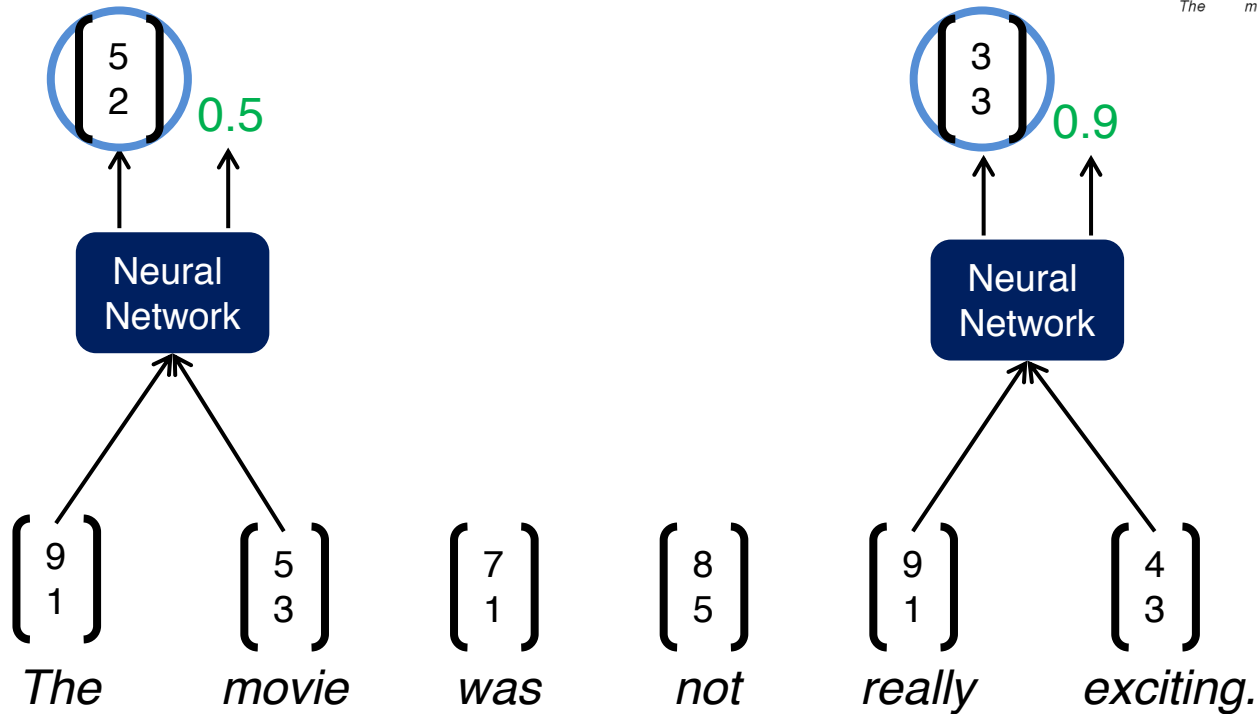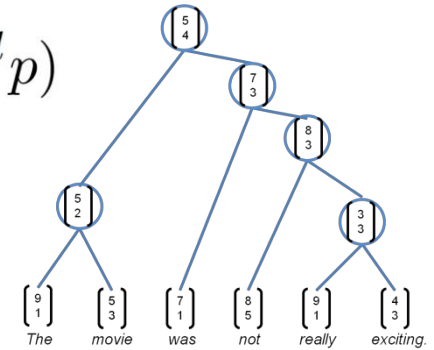
$$label_p = softmax(W^{label} p)$$

Predicting Sentiment with RNNs
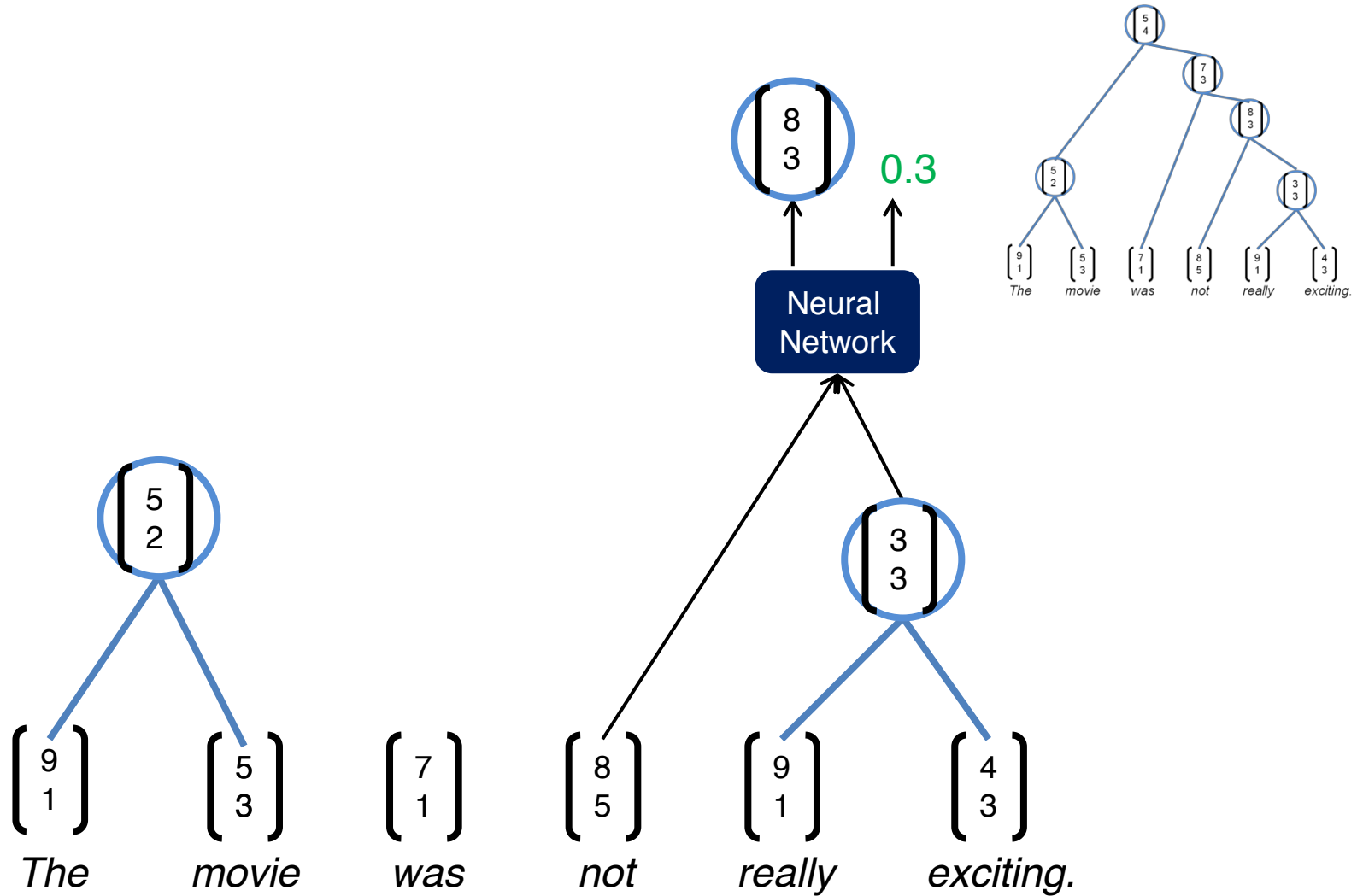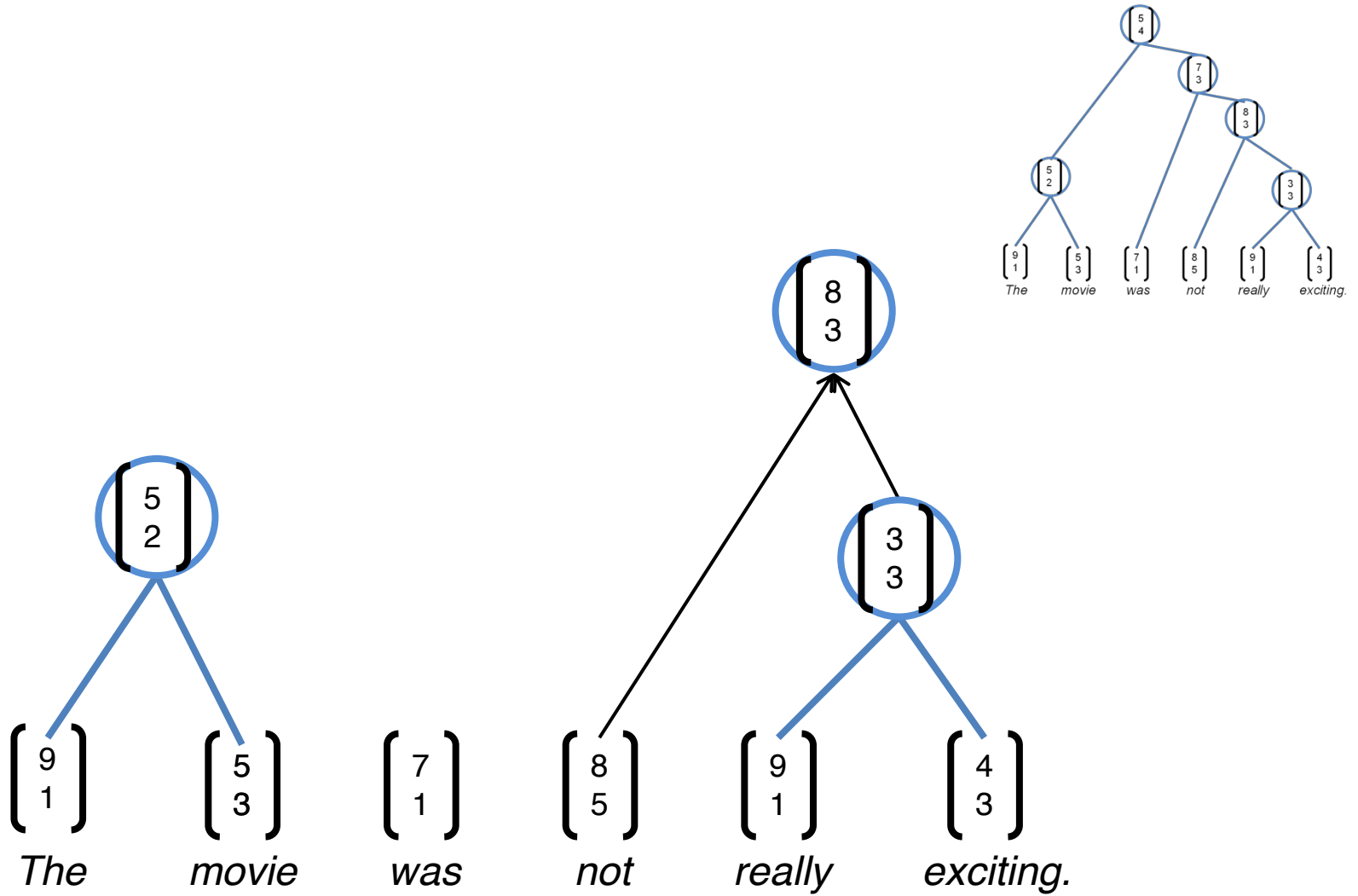
Predicting Sentiment with RNNs

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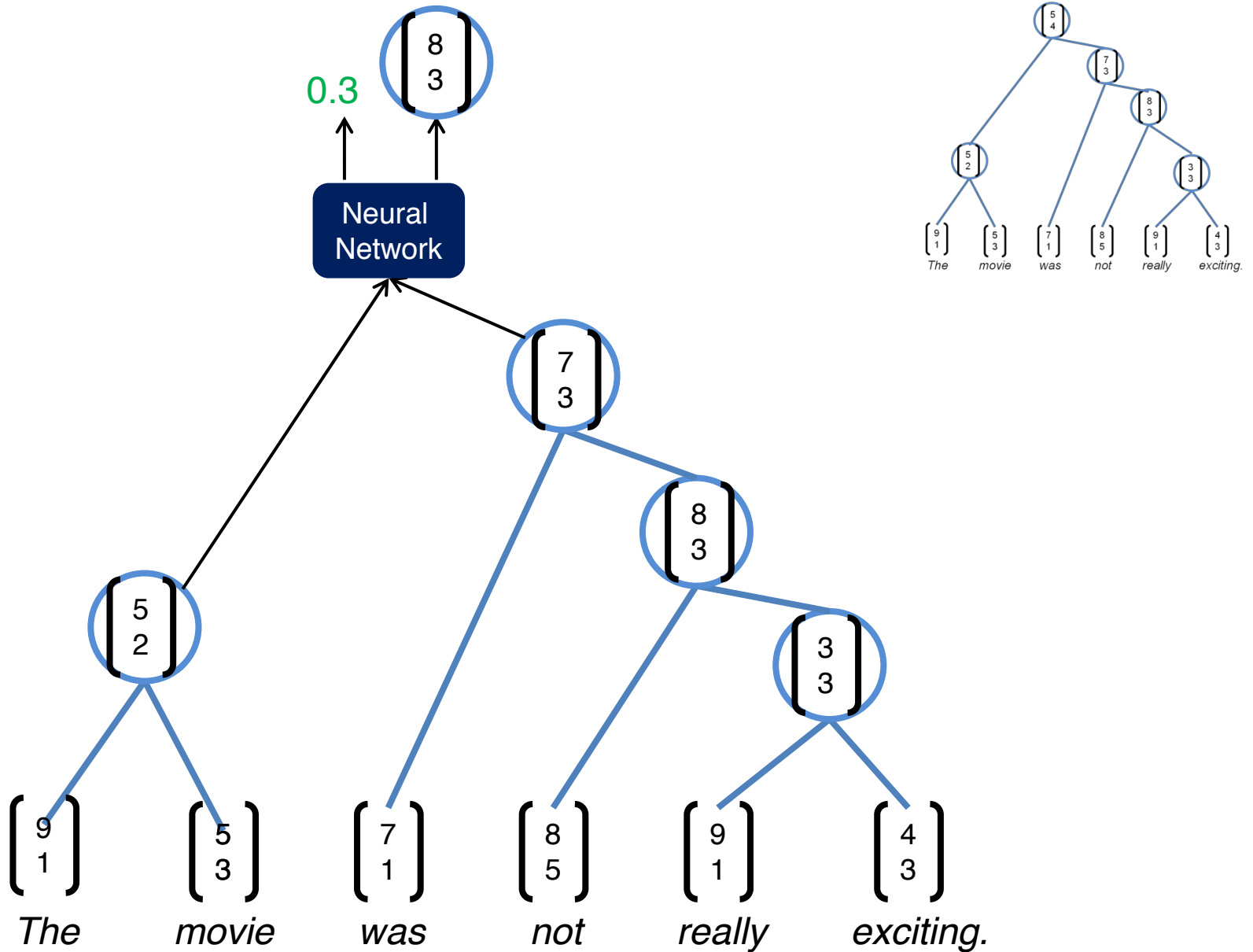
Goal: Algorithms that recover and learn semantic vector representations based on recursive structure for multiple language tasks.

1. Introduction

2. Word Vectors and Recursive Neural Networks

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

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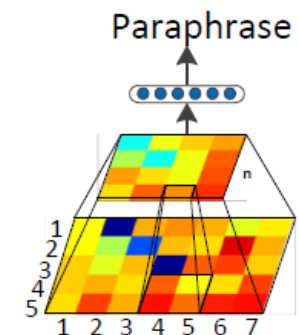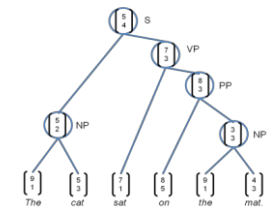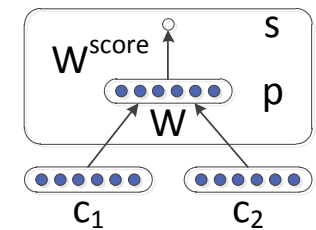
- Sentiment detection is crucial to business intelligence, stock trading, …

Maybe she'll change her name to Halliburton. Just to see.

3/18/11 at 4:00 PM | 17 Comments

Mentions of the Name 'Anne Hathaway' May Drive Berkshire Hathaway Stock

By Patrick Huguenin

The Huffington Post recently pointed out that whenever Anne Hathaway is in the news, the stock price for Warren Buffett's Berkshire Hathaway goes up. Really. When *Bride Wars* opened, the stock rose 2.61 percent. (*Rachel Getting Married* only kicked it up 0.44 percent, but, you know, that one was so light on plot compared to *Bride Wars*.)

- 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.
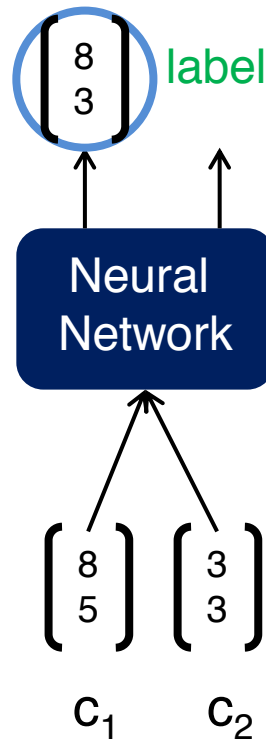
Recursive Autoencoders

- Similar to RNN but with additional reconstruction error to keep as much information as possible

Reconstruction error   Softmax Classifier

8
3   label

Neural
Network

$W^{(2)}$

$W^{(label)}$

$W^{(1)}$

$\begin{bmatrix} 8 \\ 5 \end{bmatrix}$ $\begin{bmatrix} 3 \\ 3 \end{bmatrix}$

c_1 c_2

$$p = \text{sigmoid}\left(W \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} + b\right)$$
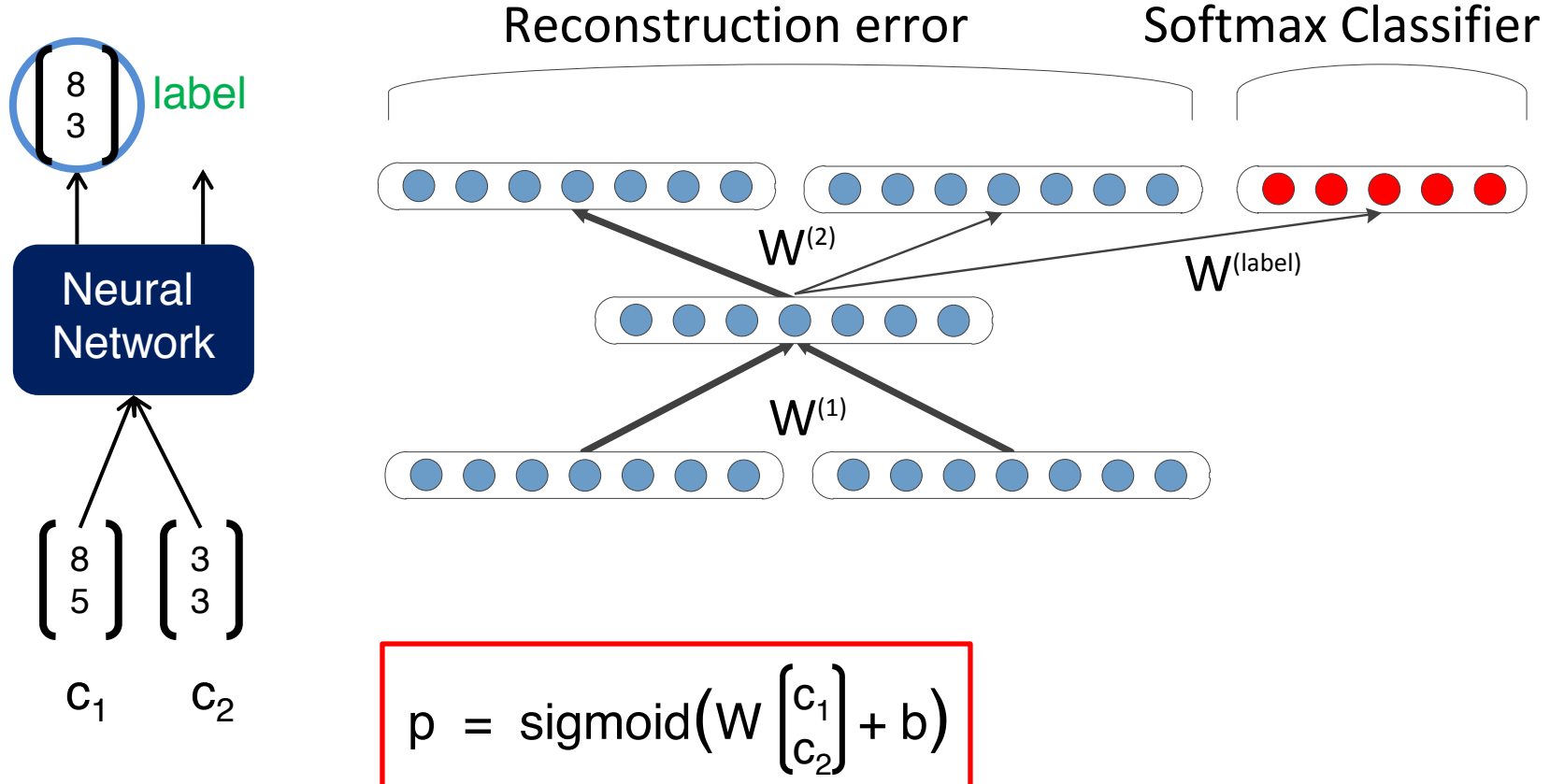
Recursive Autoencoders

- Reconstruction error details

$$[c_1'; c_2'] = W^{(2)}p + b^{(2)}.$$

Reconstruction error   Softmax Classifier

$W^{(2)}$

$W^{(label)}$

$W^{(1)}$
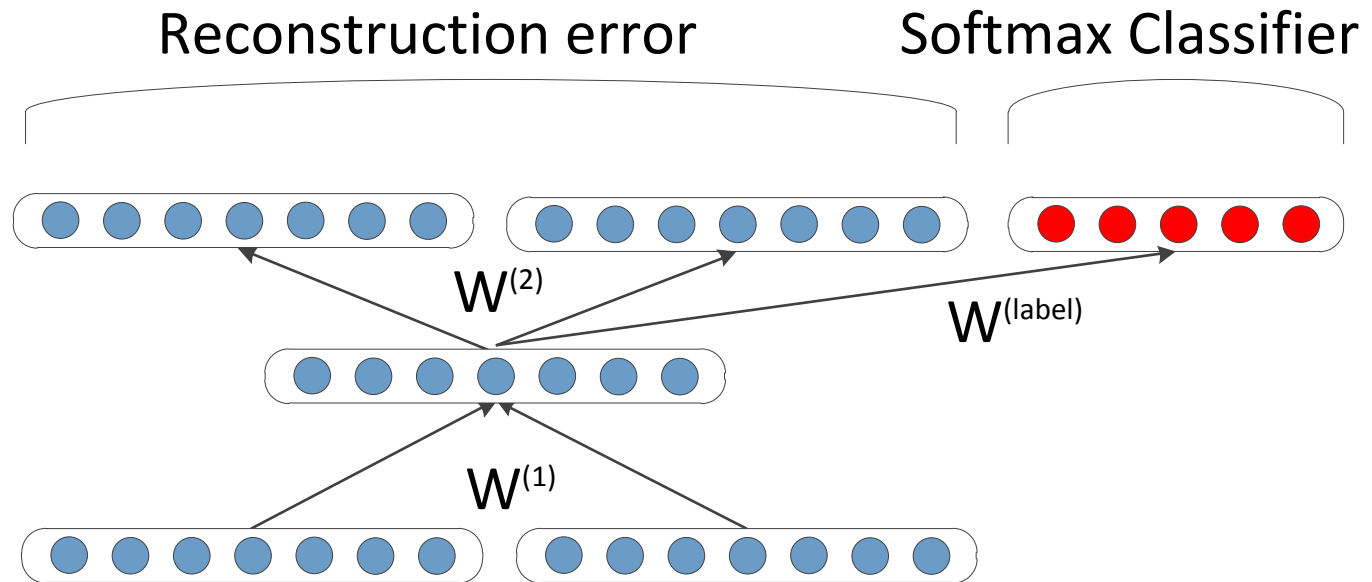
$$E_{rec}([c_1; c_2]) = \frac{1}{2}\left|\left|[c_1; c_2] - [c_1'; c_2']\right|\right|^2$$

Recursive Autoencoders

- Reconstruction error at every node
- Important detail: normalization

$$E_{rec}([c_1; c_2]) = \frac{1}{2} \left|\left| [c_1; c_2] - [c_1'; c_2'] \right|\right|^2$$

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

$p = \frac{p}{||p||}$

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

$p = \frac{p}{||p||}$

$x_1 \quad x_2 \quad x_3$
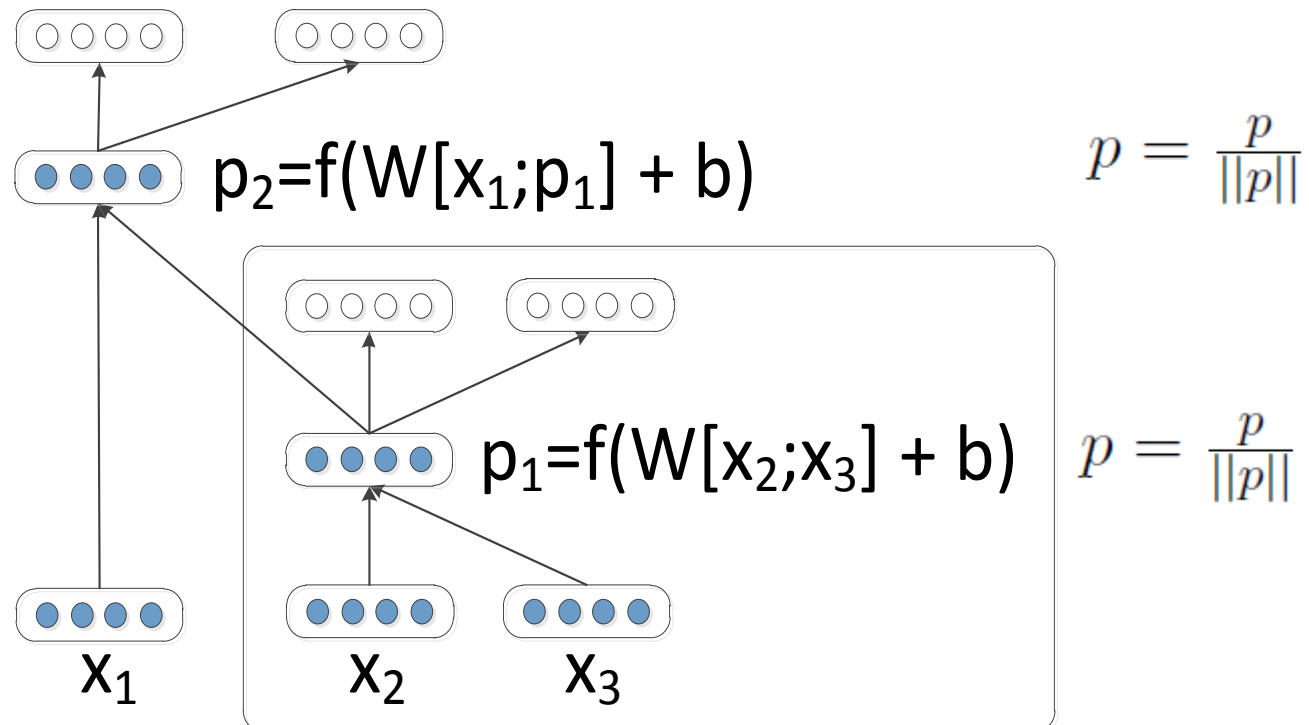
Accuracy of Positive/Negative Sentiment Classification

- 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

Method	MR	MPQA
Phrase voting with lexicons	63.1	81.7
Bag of features with lexicons	76.4	84.1
Tree-CRF (Nakagawa et al. 2010)	77.3	86.1
RAE (this work)	**77.7**	**86.4**

Sorted Negative and Positive N-grams

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

Learning Compositionality from Movie Reviews

- Probability of being positive of several n-grams

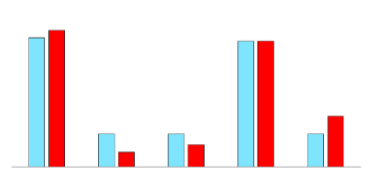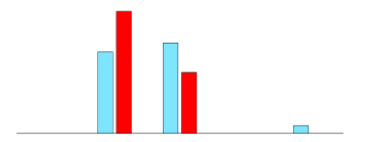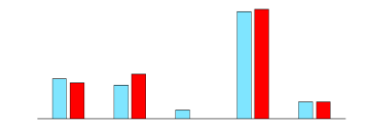| *n*-gram	P(positive | *n*-gram)
good	0.45
not good	0.20
very good	0.61
not very good	0.15
not	0.03
very	0.23

Vector representations when training only for sentiment
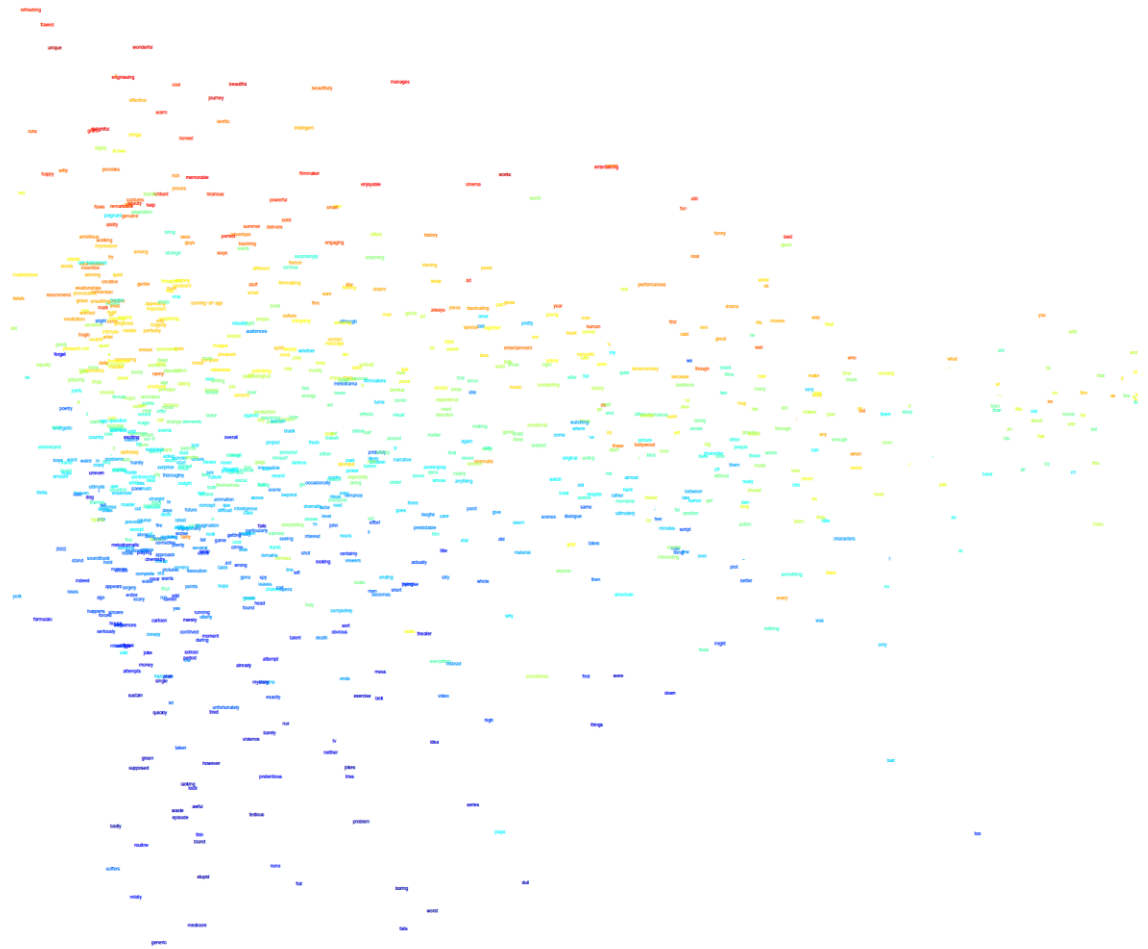
- Learn distributions over multiple complex sentiments → New dataset and task

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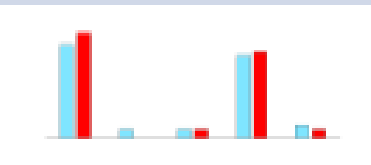
- Sorry, Hugs; You rock; Tee-hee ; I understand;
  Wow just wow

Predicted and Gold Distribution	Anonymous Confession
	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 …
	well i think hairy women are attractive
	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. …

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

Predicted and Gold Distribution	Anonymous Confession
	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.
	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…
	My paper is due in less than 24 hours and I'm still dancing round my room!

Experience Project most votes results

Method	Accuracy %
Random	20
Most frequent class	38
Bag of words; MaxEnt classifier	46
Spellchecker, sentiment lexica, SVM	47
SVM on neural net word features	46
RAE (this work)	**50**

Average KL between gold and predicted label distributions:

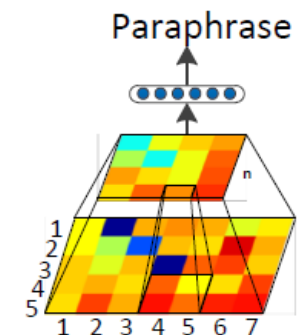$$KL(g||p) = \sum_i g_i \log(g_i/p_i)$$
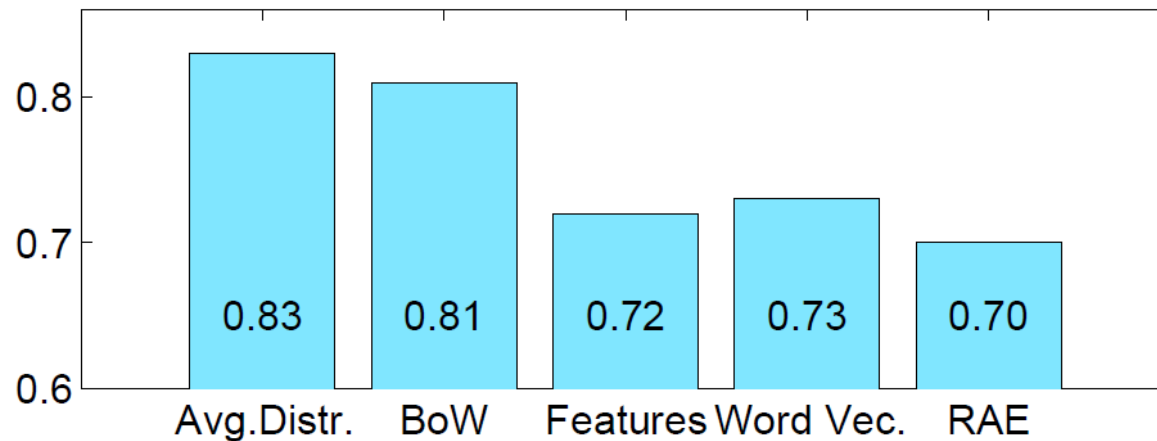
Goal: Algorithms that recover and learn semantic vector representations based on recursive structure for multiple language tasks.

1. Introduction

2. Word Vectors and Recursive Neural Networks

3. Recursive Autoencoders for Sentiment Analysis

4. Paraphrase Detection
   [Socher et al., NIPS 2011]
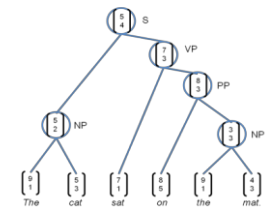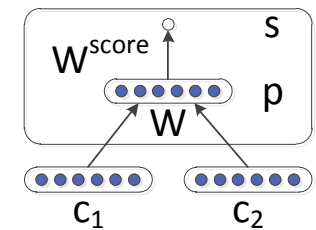
Paraphrase Detection

- 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?

Unsupervised unfolding RAE

$$E_{rec}(y_{(i,j)}) = \left|\left|[x_i; \ldots; x_j] - [x'_i; \ldots; x'_j]\right|\right|^2$$
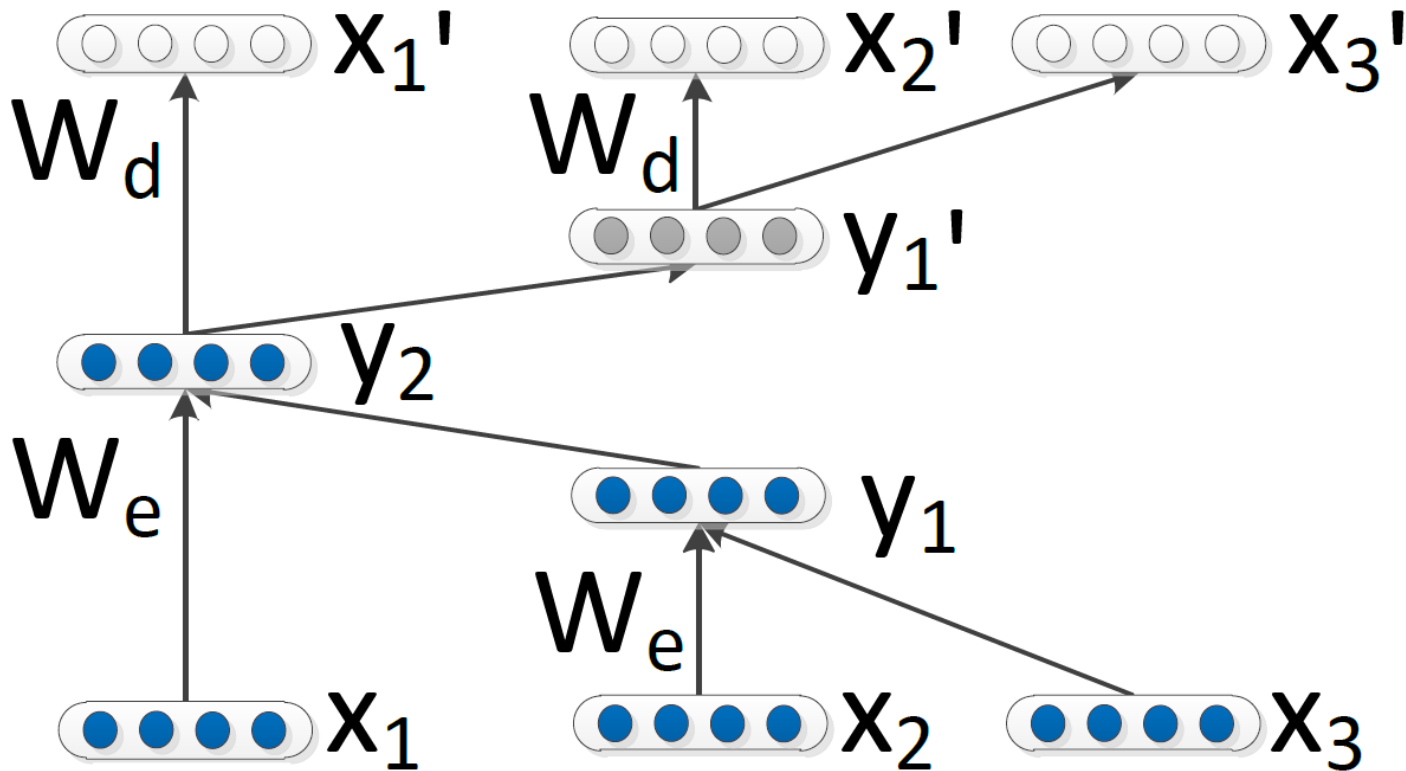
- ## More semantic vector representations

Center Phrase	RAE	Unfolding RAE
the U.S.	the Swiss	the former U.S.
suffering low morale	suffering due to no fault of my own	suffering heavy casualties
advance to the next round	advance to the final of the UNK 1.1 million Kremlin Cup	advance to the semis
a prominent political figure	the second high-profile opposition figure	a powerful business figure
conditions of his release	conditions of peace, social stability and political harmony	negotiations for their release

How much can the vectors capture?
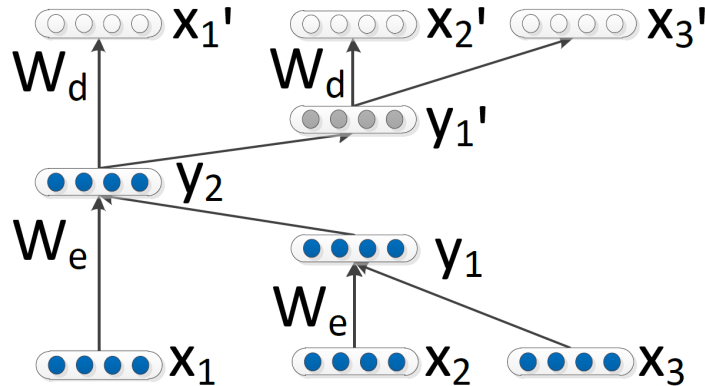
| Encoding Input	Generated Text from Unfolded Reconstruction
a December summit	a December summit
the first qualifying session	the first qualifying session
English premier division club	Irish presidency division club
the safety of a flight	the safety of a flight
the signing of the accord	the signing of the accord
the U.S. House of Representatives	the U.S. House of Representatives
enforcement of the economic embargo	enforcement of the national embargo
visit and discuss investment possibilities	visit and postpone financial possibilities
the agreement it made with Malaysia	the agreement it made with Malaysia
the full bloom of their young lives	the lower bloom of their democratic lives
the organization for which the men work	the organization for Romania the reform work
a pocket knife was found in his suitcase in the plane's cargo hold	a bomb corpse was found in the mission in the Irish car language case

- Unsupervised RAE and a pair-wise sentence comparison of nodes in parsed trees
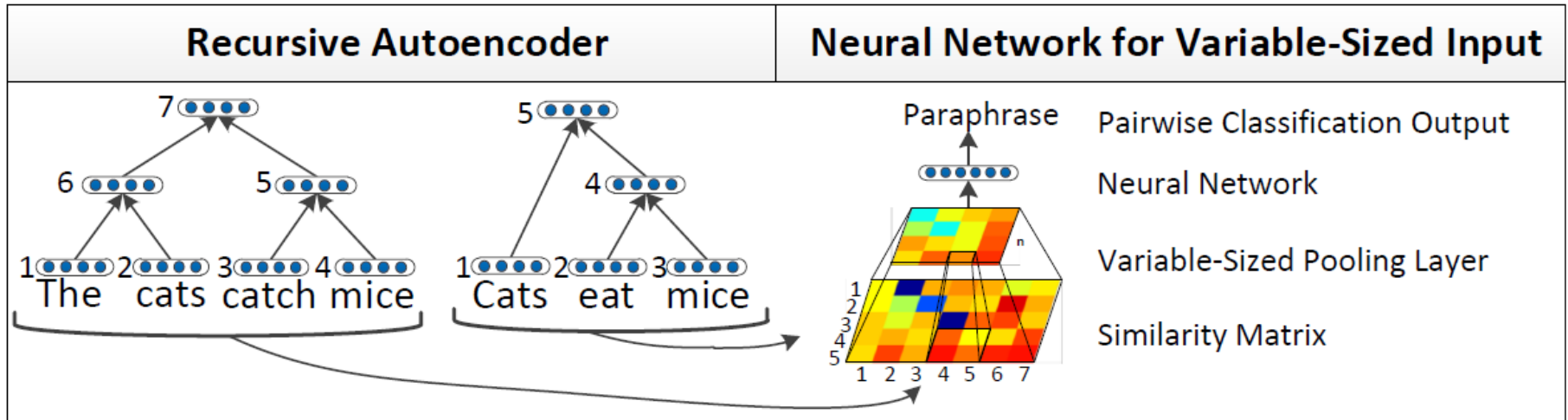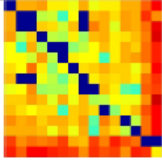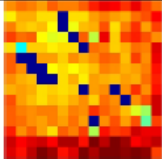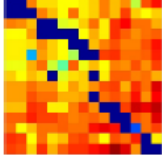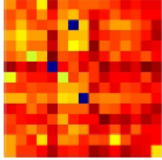
Recursive Autoencoders for Full Sentence Paraphrase Detection
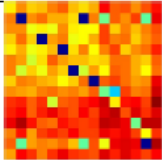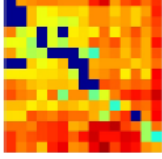
- Experiments on Microsoft Research Paraphrase Corpus (Dolan et al. (2004))

Method	Acc.	F1
All Paraphrase Baseline	66.5	79.9
Rus et al.(2008)	70.6	80.5
Mihalcea et al.(2006)	70.3	81.3
Islam et al.(2007)	72.6	81.3
Qiu et al.(2006)	72.0	81.6
Fernando et al.(2008)	74.1	82.4
Wan et al.(2006)	75.6	83.0
Das and Smith (2009)	73.9	82.3
Das and Smith (2009) + 18 Surface Features	76.1	82.7
Unfolding Recursive Autoencoder (our method)	**76.4**	**83.4**

Recursive Autoencoders for Full Sentence Paraphrase Detection

L	Pr	Sentences	Sim.Mat.
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	

- # Questions?

$$p = \text{sigmoid}\left(W \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} + b\right),$$

$$label_p = softmax(W^{label} p)$$