Deep Learning for NLP tasks

School of Information Studies
Syracuse University

NLP Word Level Classifiers

Uses word vector learning on inputs, but replaces single vector output with a classifier layer for the task, S

Diagram is similar to a conventional classifier, except that it includes learning for the feature input vectors from unsupervised data

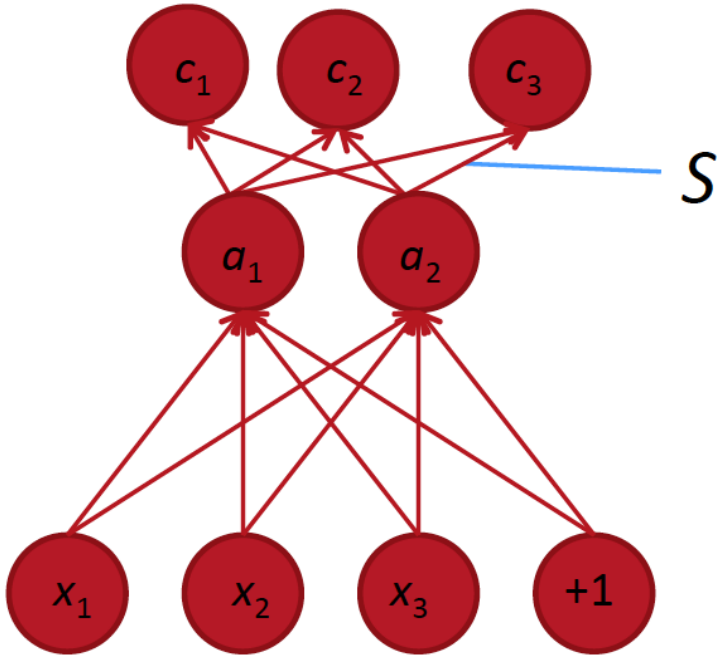


Diagram MS

Two NLP word level tasks

POS tagging and Named Entity Recognition (NER)

	POS WSJ (acc.)	NER CoNLL (F1)
State-of-the-art*	97.24	89.31
Supervised NN	**96.37**	**81.47**
Unsupervised pre-training followed by supervised NN**	**97.20**	**88.87**
+ hand-crafted features***	97.29	89.59

* Representative systems: POS: (Toutanova et al. 2003), NER: (Ando & Zhang 2005)

** 130,000-word embedding trained on Wikipedia and Reuters with 11 word window, 100 unit hidden layer – **for 7 weeks!** – then supervised task training

***Features are character suffixes for POS and a gazetteer for NER

Diagram MS

Architectures for NLP tasks with Structures

Would like a Deep Learning approach that can use good intermediate representations that can be shared across tasks

Many NLP language tasks share syntactic sentence structure

These structures are also recursive in nature => Recursive Deep Learning

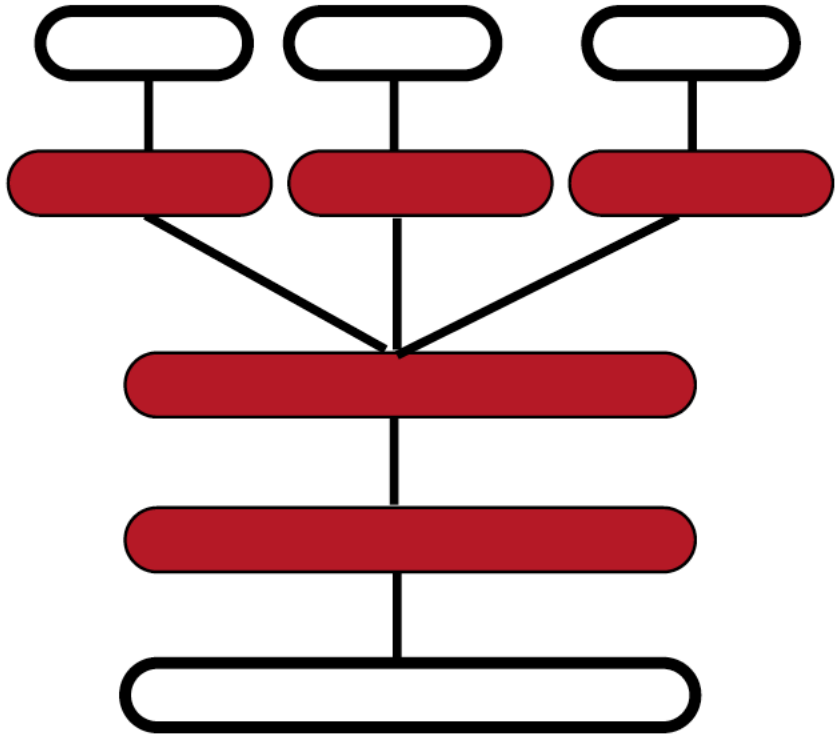


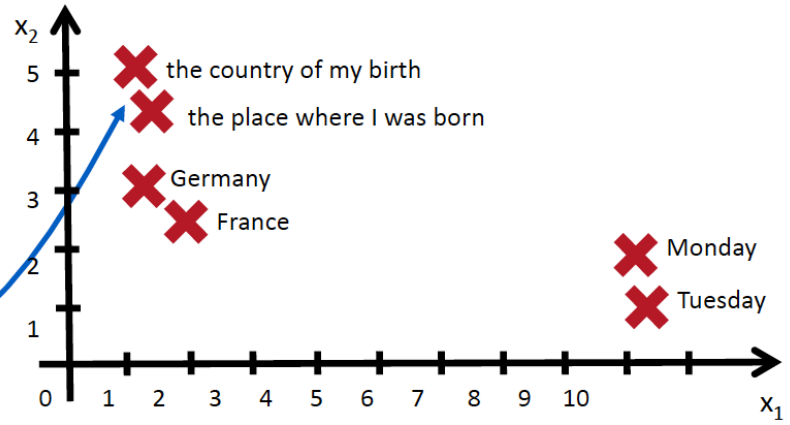
Diagram MS – one design of NN layer representations applied to several tasks

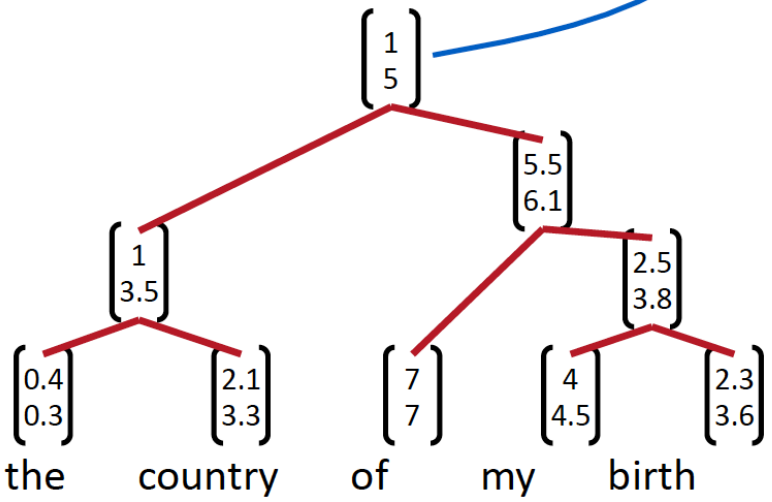
Phrase Level Vectors

Represent the meaning of longer phrases by mapping them into the same vector space

Use 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.





Models in this section can jointly learn parse trees and compositional vector representations

Diagram RS

Recursive Neural Networks

Instead of traditional combination of NN node output by summing weighted vectors, combine two semantic representations and score how plausible the result will be

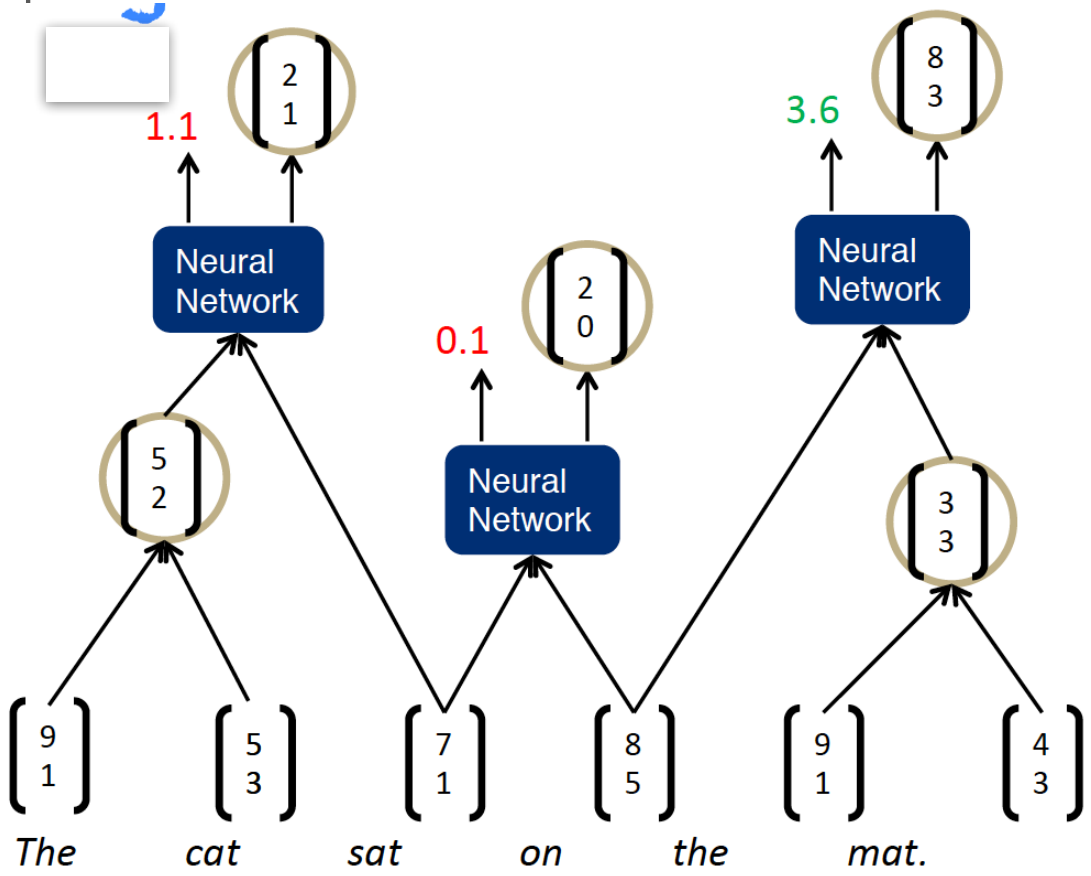


Diagram MS – In parsing "the cat sat on the mat", combining vectors representing phrases to get new phrase vector and score

Parsing Results

Tested on standard WSJ with F1 scores

- CVG is RNN combined with PCFG (probabilistic context free grammars) , SU is syntactically untied RNN (faster)

Parser	Test, All Sentences
Stanford PCFG, (Klein and Manning, 2003a)	85.5
Stanford Factored (Klein and Manning, 2003b)	86.6
Factored PCFGs (Hall and Klein, 2012)	89.4
Collins (Collins, 1997)	87.7
SSN (Henderson, 2004)	89.4
Berkeley Parser (Petrov and Klein, 2007)	90.1
CVG (RNN) (Socher et al., ACL 2013)	85.0
CVG (SU-RNN) (Socher et al., ACL 2013)	90.4
Charniak - Self Trained (McClosky et al. 2006)	91.0
Charniak - Self Trained-ReRanked (McClosky et al. 2006)	92.1

Advantage of Deep Learning

Current NLP systems, in general, are fragile because they depend on which lexical items are in the supervised training data

Add robustness with word representations from unlabeled data

- WSJ doesn't have many occurrences of foods, almost none of "bananas"
- But word vector for "bananas" is similar to "oranges", get a better parse

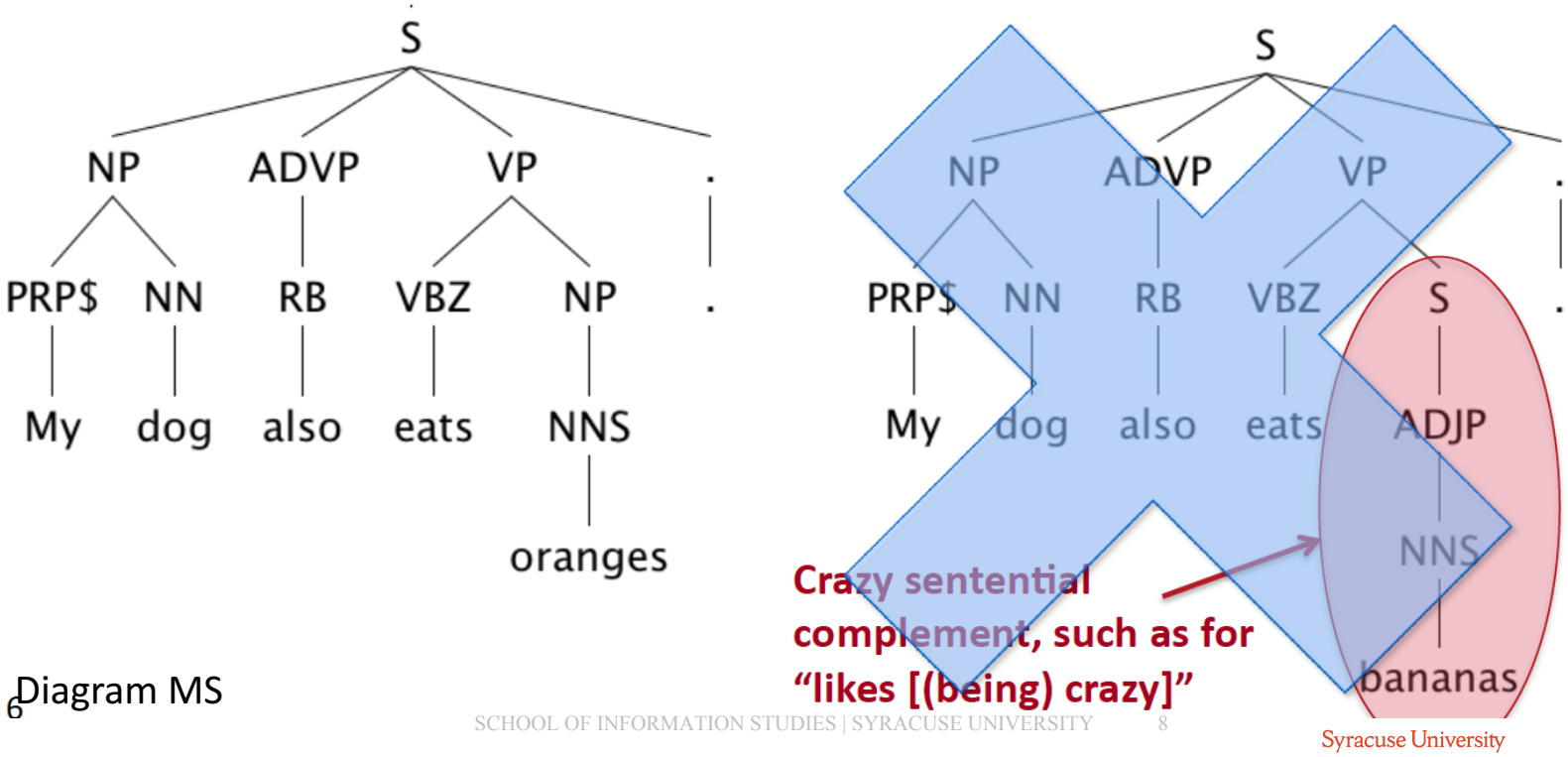


Diagram MS

Paraphrase Detection

Task is to compare sentences to see if they have the same semantics

Examples

- 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 in Decembr 28
- It stems from a Modesto police report

Solution is a RNN called Recursive Autoencoders to compare vector representations of sentences

Paraphrase Results

Experiments on Microsoft Paraphrase Corpus

Method	Acc.	F1
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
F. Bu et al. (ACL 2012): String Re-writing Kernel	76.3	--
Unfolding Recursive Autoencoder (NIPS 2011)	**76.8**	**83.6**

Sentiment Analysis on Movies

Label movie review sentences for positive or negative sentiment

Examples

- 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 going.

Solution:

- New sentiment phrase treebank to provide supervised means of combining sentiment phrases (available from Socher and Kaggle)
- Recursive Neural Tensor Network – most powerful method so far and provides more interaction of vectors in the composition of phrases (Socher et al 2013)

Sentiment Results

Evaluate results on sentences from movie reviews (Pang and Lee 2006)

Model	Fine-grained		Positive/Negative	
	All	Root	All	Root
NB	67.2	41.0	82.6	81.8
SVM	64.3	40.7	84.6	79.4
BiNB	71.0	41.9	82.7	83.1
VecAvg	73.3	32.7	85.1	80.1
RNN	79.0	43.2	86.1	82.4
MV-RNN	78.7	44.4	86.8	82.9
RNTN	**80.7**	**45.6**	**87.6**	**85.4**

Diagram RS

Visualization of Sentiment

RNTN for sentiment can capture “X but Y” (shown here) and also various negation constructs

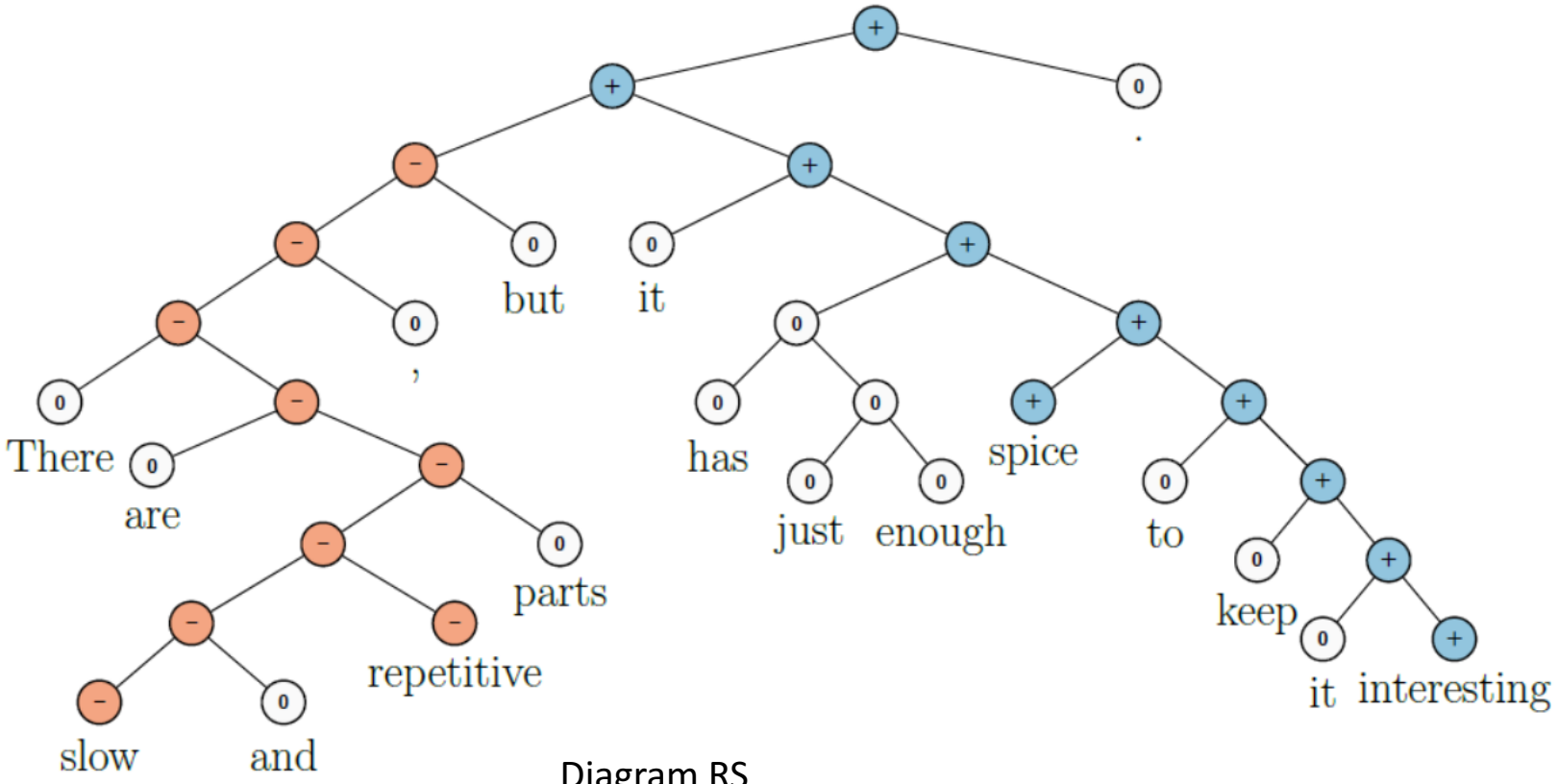


Diagram RS

Summary of Deep Learning on NLP tasks

Also RNN for

- Relationship learning
- Question answering
- Object detection in images

Excellent recent results on Machine Translation has now gone into products such as Google Translate

- Sequence to Sequence Learning with NN
- Sutskever et al 2014, Luong et al 2016

Ongoing research into types of NN and how to apply them to tasks