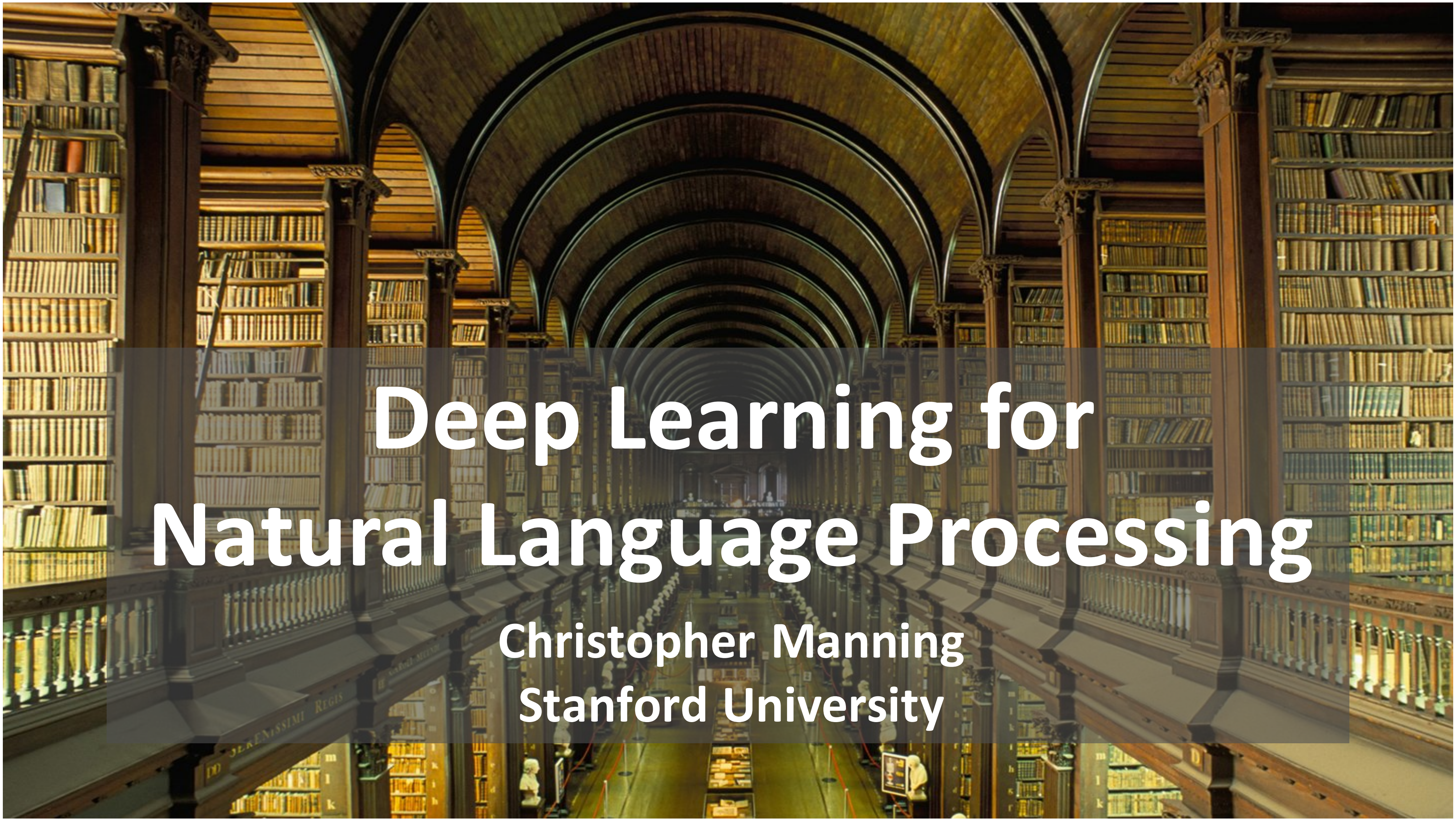


Deep Learning for Natural Language Processing

Christopher Manning
Stanford University

1980s Natural Language Processing

VP →{ V (NP:(↑ OBJ)=↓ (NP:(↑ OBJ2)=↓))

(XP:(↑ XCOMP)=↓)

|@(COORD VP VP)}.

salmon N IRR @(CN SALMON)

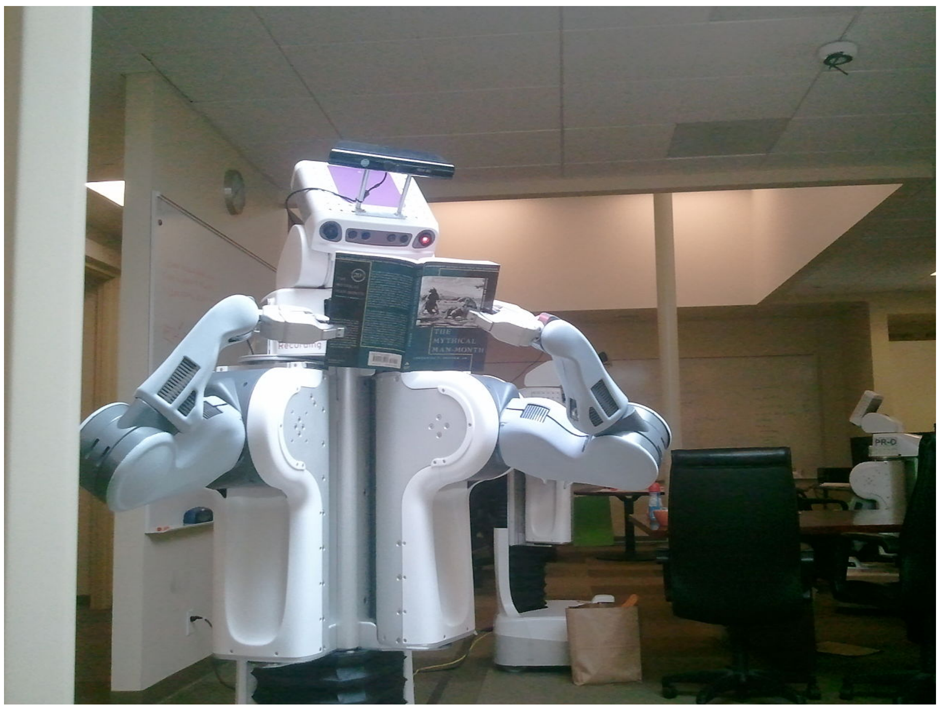
(↑ PERSON)=3

{ (↑ NUM)=SG|(↑ NUM)=PL}.

$$\begin{bmatrix} \text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{“I”} \end{bmatrix} \\ \text{OBJ}_{\text{go}} & \begin{bmatrix} \text{PRED} & \text{“you”} \end{bmatrix} \\ \text{PRED} & \text{‘perfect}\langle\text{XCOMP}\rangle\text{SUBJ’} \\ \text{XCOMP} & \begin{bmatrix} \text{SUBJ} & \text{—} \\ \text{PRED} & \text{‘bring}\langle\text{SUBJ, OBJ, OBJ}_{\text{go}}\rangle\text{’} \\ \text{OBJ} & \begin{bmatrix} \text{PRED} & \text{“candy”} \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

Christopher Manning

1990s, 2000s: Learning language



WRB VBZ DT NN VB TO VB DT
How does a project get to be a
NN JJ . : CD NN IN DT NN .
year late ? … One day at a time .

$$P(\text{late} \mid \text{a, year}) = 0.0087$$

$$P(\text{NN} \mid \text{DT, a, project}) = 0.9$$

Christopher Manning

The traditional word representation

motel

[0 0 0 0 0 0 0 0 0 0 1 0 0 0 0]

Dimensionality: 50K (small domain – speech/PTB) – 13M (web – Google 1T)

motel [0 0 0 0 0 0 0 0 0 0 1 0 0 0 0] AND
hotel [0 0 0 0 0 0 0 1 0 0 0 0 0 0 0] = 0

Christopher Manning

Word distributions → distributed word representations

Through corpus **linguistics**, large chunks
the study of language and **linguistics**.
The field of **linguistics** is concerned
Written like a **linguistics** text book
Phonology is the branch of **linguistics** that

linguistics = $\begin{pmatrix} 0.286 \\ 0.792 \\ -0.177 \\ -0.107 \\ 0.109 \\ -0.542 \\ 0.349 \\ 0.271 \\ 0.487 \end{pmatrix}$

[Bengio et al. 2003, Mnih & Hinton 2008, Collobert & Weston 2008, Turian 2010, Mikolov 2013, etc.]

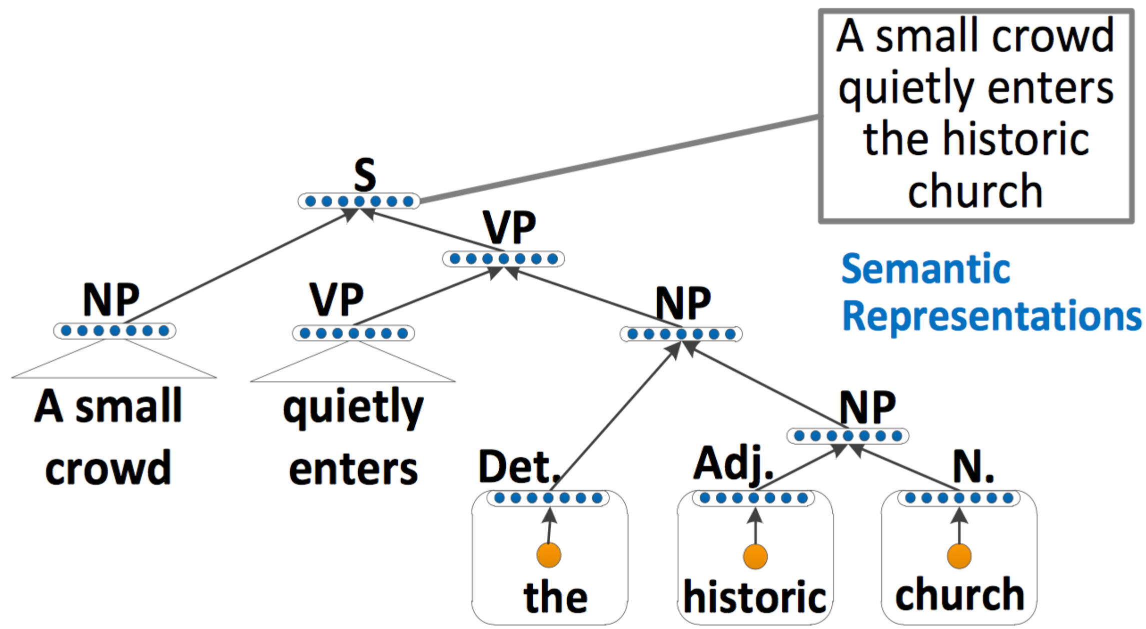
Christopher Manning

Distributed word representations

A foundational component of deep networks in NLP

Base case for meaning composition

Vector space model is widely used for semantic similarity



Matrix-based methods for learning word representations

LSA (SVD), HAL (Lund & Burgess), COALS (Rohde et al), Hellinger-PCA (Lebret & Collobert)

- Fast training
- Efficient usage of statistics
- Primarily used to capture word similarity
- Disproportionate importance given to small counts

“Neural” methods for learning word representations

NNLM, HLBL, RNN, word2vec
Skip-gram/CBOW, (i)vLBL
(Bengio et al; Collobert & Weston; Huang et al; Mnih & Hinton; Mikolov et al; Mnih & Kavukcuoglu)

- Scales with corpus size
- Inefficient usage of statistics
- Generate improved performance on other tasks
- Can capture complex patterns beyond word similarity

Christopher Manning

Matrix-based methods for learning word representations

LSA (SVD), HAL (Lund & Burgess), COALS (Rohde et al), Hellinger-PCA (Lebret & Collobert)

- Fast training
- Efficient usage of statistics
- Primarily used to capture word similarity
- Disproportionate importance given to small counts

NNLM, HLBL, RNN, word2vec Skip-gram/CBOW, (i)vLBL (Bengio et al; Collobert & Weston; Huang et al; Mnih & Hinton; Mikolov et al; Mnih & Kavukcuoglu)

- Scales with corpus size
- Inefficient usage of statistics
- Generate improved performance on other tasks
- Can capture complex patterns beyond word similarity

New, scalable log-bilinear model for word representations

Christopher Manning

Word Analogies

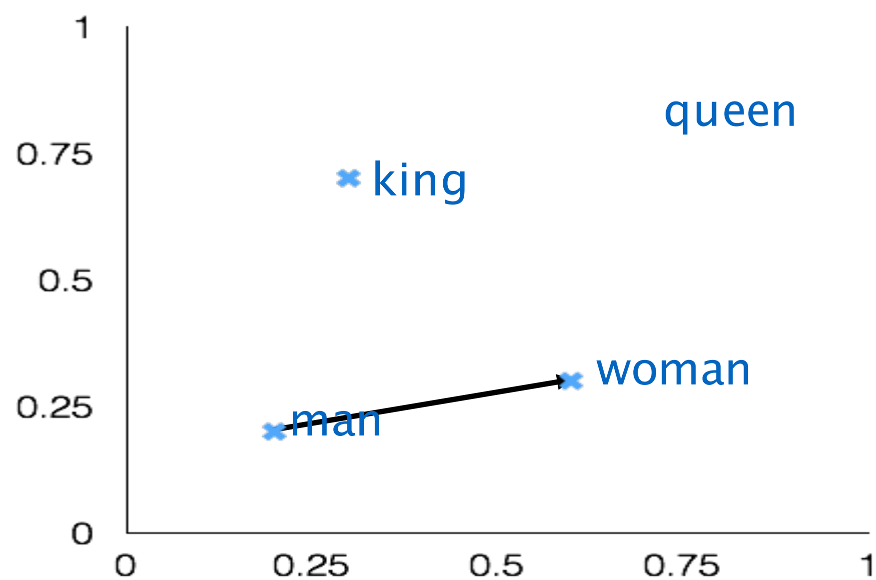
Test for linear relationships, examined by Mikolov et al.

$\boxed{a:b :: c:?}$

$$d = \arg\max_x \frac{(w_b - w_a + w_c)^T w_x}{||w_b - w_a + w_c||}$$

man:woman :: king:?

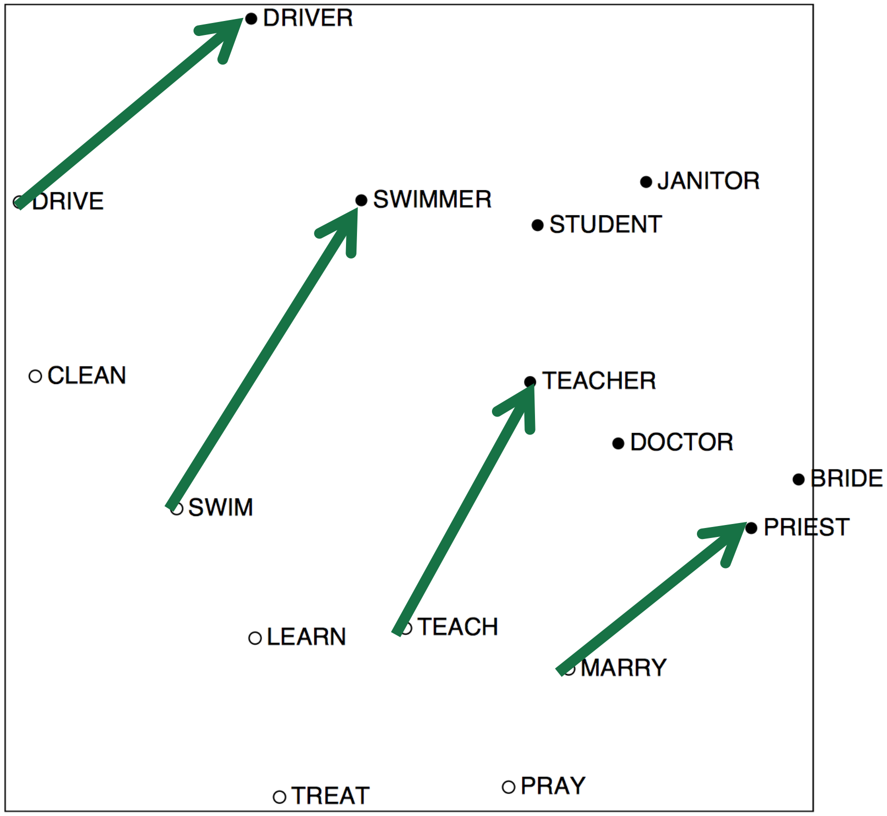
+	king	[0.30 0.70]
−	man	[0.20 0.20]
+	woman	[0.60 0.30]
	queen	[0.70 0.80]



Christopher Manning

COALS model

[Rohde, Gonnerman & Plaut, ms., 2005]



Encoding meaning in vector differences

[Pennington, Socher, and Manning, EMNLP 2014]

Crucial insight: Ratios of co-occurrence probabilities can encode meaning components

	x = solid	x = gas	x = water	x = random
$P(x \mid \text{ice})$	large	small	large	small
$P(x \mid \text{steam})$	small	large	large	small
$\dfrac{P(x \mid \text{ice})}{P(x \mid \text{steam})}$	large	small	~1	~1

Encoding meaning in vector differences

[Pennington et al., EMNLP 2014]

Crucial insight: Ratios of co-occurrence probabilities can encode meaning components

	x = solid	x = gas	x = water	x = fashion
$P(x\|\text{ice})$	1.9×10^{-4}	6.6×10^{-5}	3.0×10^{-3}	1.7×10^{-5}
$P(x\|\text{steam})$	2.2×10^{-5}	7.8×10^{-4}	2.2×10^{-3}	1.8×10^{-5}
$\frac{P(x\|\text{ice})}{P(x\|\text{steam})}$	8.9	8.5×10^{-2}	1.36	0.96

Christopher Manning

GloVe: A new model for learning word representations

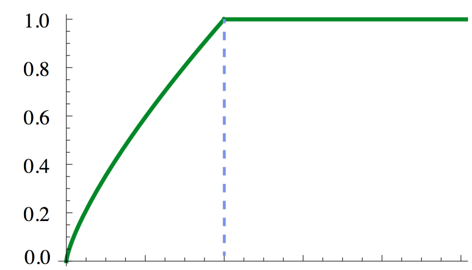
[Pennington et al., EMNLP 2014]

$$w_i \cdot w_j = \log P(i|j)$$

$$w_x \cdot (w_a - w_b) = \log \frac{P(x|a)}{P(x|b)}$$

$$J = \sum_{i,j=1}^{V} f\left(X_{ij}\right)\left(w_i^T \tilde{w}_j + b_i + \tilde{b}_j - \log X_{ij}\right)^2$$

$f \sim$



Christopher Manning

Word similarities

Nearest words to frog:

1. frogs
2. toad
3. litoria
4. leptodactylidae
5. rana
6. lizard
7. eleutherodactylus

litoria

leptodactylidae

rana

eleutherodactylus

Word analogy task [Mikolov, Yih & Zweig 2013a]

Model	Dimensions	Corpus size	Performance (Syn + Sem)
CBOW (Mikolov et al. 2013b)	300	1.6 billion	36.1

Christopher Manning

Named Entity Recognition Performance

Model on CoNLL	CoNLL 2003 dev	CoNLL 2003 test	ACE 2	MUC 7
Categorical CRF	91.0	85.4	77.4	73.4
SVD (log tf)	90.5	84.8	73.6	71.5
HPCA	92.6	**88.7**	81.7	80.7
HSMN (Huang)	90.5	85.7	78.7	74.7
C&W	92.2	87.4	81.7	80.2
CBOW	93.1	88.2	82.2	81.1
GloVe (this work)	**93.2**	88.3	**82.9**	**82.2**

F1 score of CRF trained on CoNLL 2003 English with 50 dim word vectors.

The GloVe Model

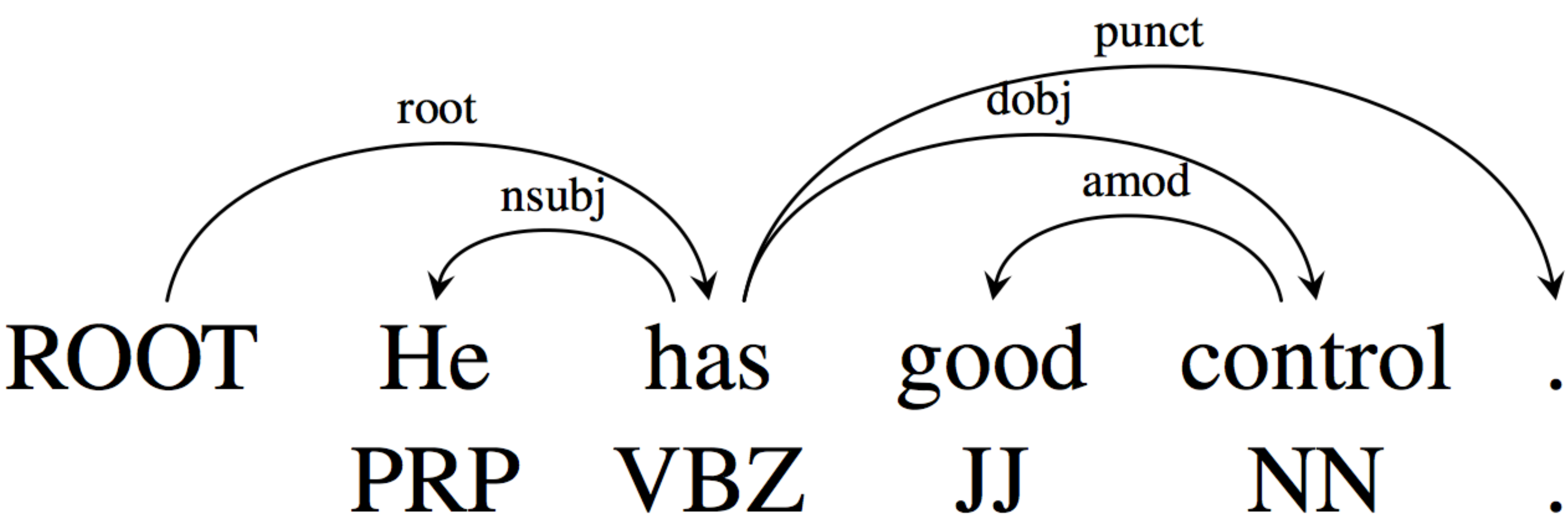
A new global-statistics, unsupervised model for learning word vectors

Design translates relationships between word-word co-occurrence probabilities that encode meaning relationships into linear relations in a word vector space

http://nlp.stanford.edu/projects/glove/

Christopher Manning

Sentence structure: Dependency parsing

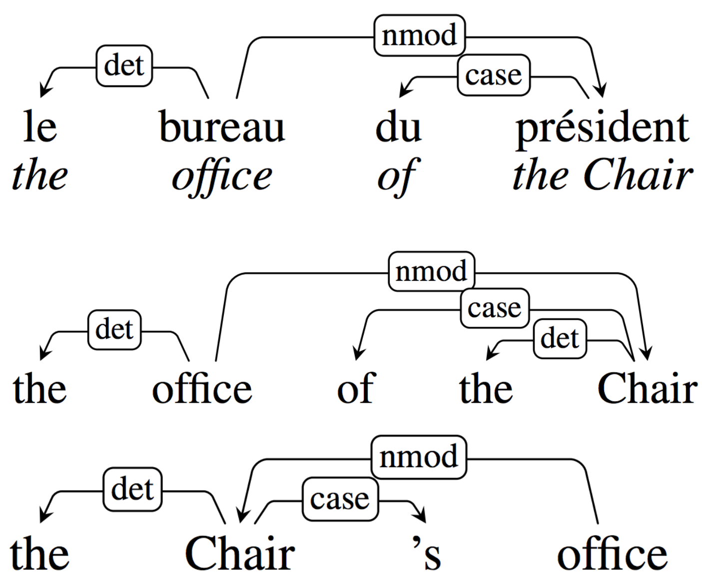


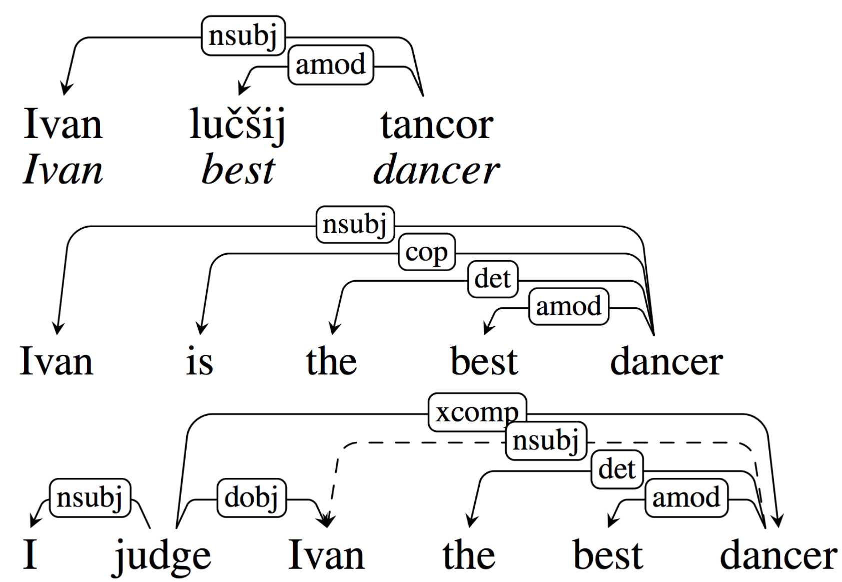
Christopher Manning

Universal (Stanford) Dependencies

[de Marneffe et al., LREC 2014]

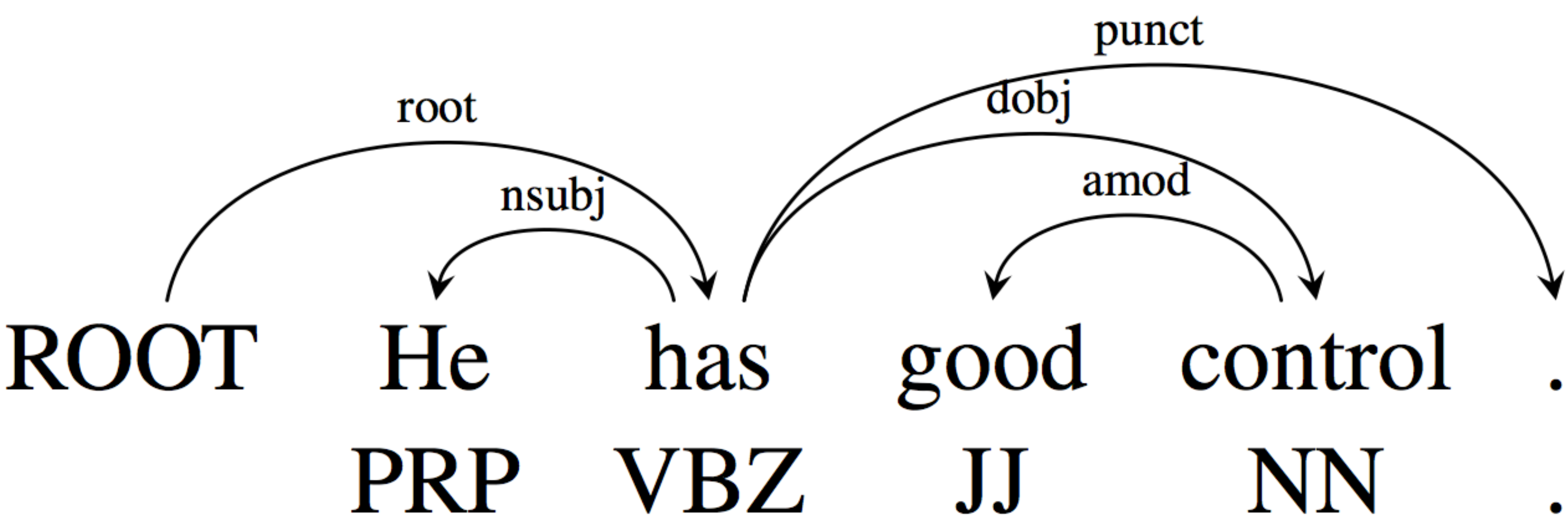
A common dependency representation and label set applicable across languages – http://universaldependencies.github.io/docs/





Christopher Manning

Sentence structure: Dependency parsing



Deep Learning Dependency Parser

[Chen & Manning, EMNLP 2014]

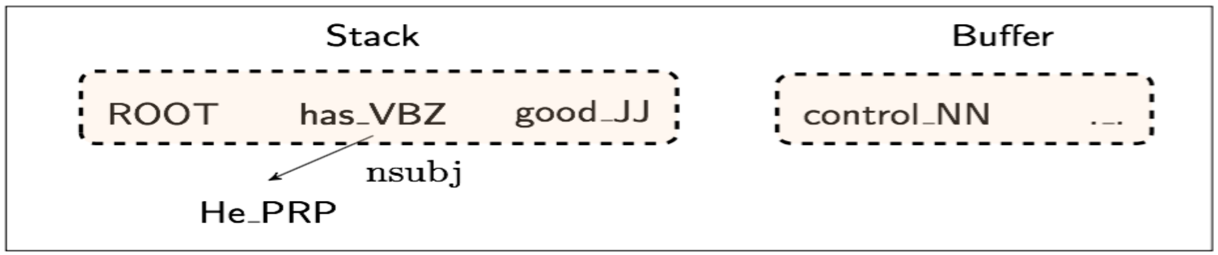
- An accurate and fast neural-network-based dependency parser!
- Parsing to Stanford Dependencies:
  - Unlabeled attachment score (UAS) = head
  - Labeled attachment score (LAS) = head and label

Parser	UAS	LAS	sent / s
MaltParser	89.8	87.2	469

Christopher Manning

Shift-reduce (transition-based) dependency parser feature representation

Configuration



binary, sparse
dim $=10^6 \sim 10^7$

0	0	0	1	0	0	1	0	…	0	0	1	0

Feature templates: usually a combination of 1 ~ 3 elements from the configuration.

Indicator features

$$s1.w = \text{good} \wedge s1.t = \text{JJ}$$

$$s2.w = \text{has} \wedge s2.t = \text{VBZ} \wedge s1.w = \text{good}$$

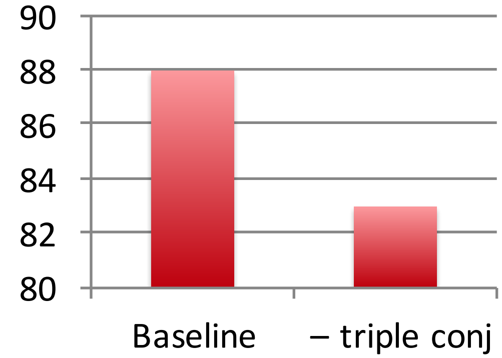
$$lc(s_2).t = \text{PRP} \wedge s_2.t = \text{VBZ} \wedge s_1.t = \text{JJ}$$

$$lc(s_2).w = \text{He} \wedge lc(s_2).l = \text{nsubj} \wedge s_2.w = \text{has}$$

Problems with indicator features

#1 Sparse

Lexicalized interaction terms are important but sparse



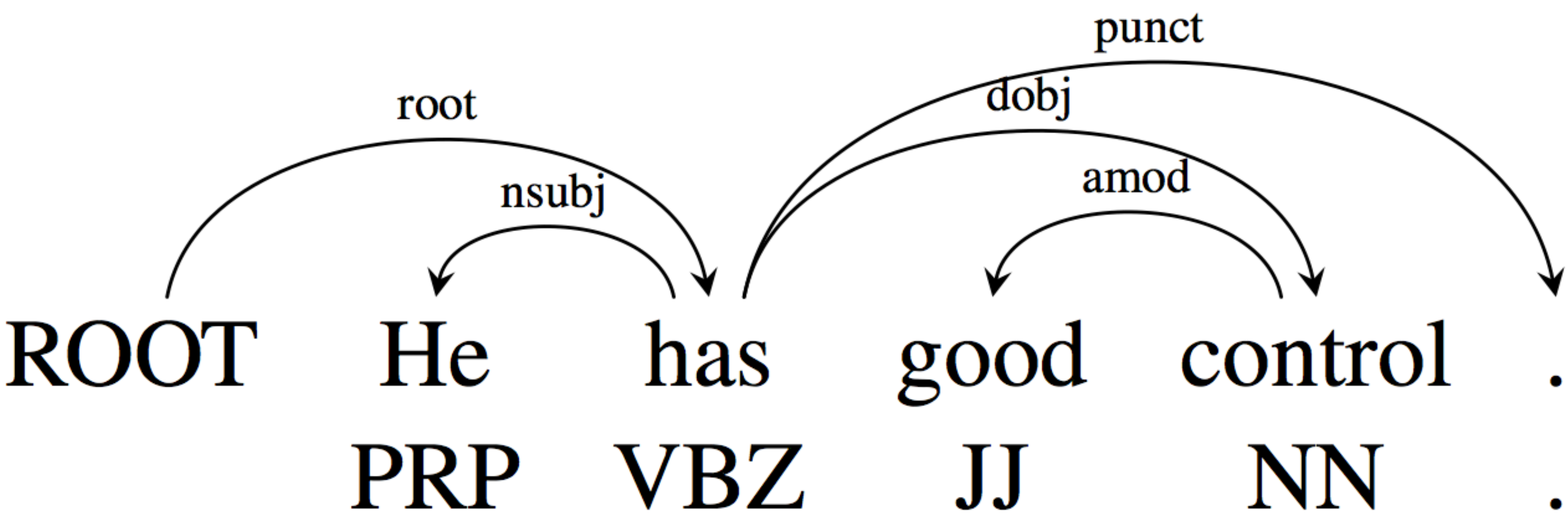
#2 Incomplete

#3 Slow

95% of parsing time is consumed by feature computation

If we encode the configuration with a distributed representation and the model captures interaction terms, all the problems are solved!

Sentence structure: Dependency parsing



"Marginal prepositions"

"There is a continual change going on by which certain participles or adjectives acquire the character of prepositions or adverbs, no longer needing the prop of a noun to cling to" – Fowler (1926)

*They moved slowly, toward the main gate, **following** the wall*

*Repeat the instructions **following** the asterisk*

*This continued most of the week **following** that ill-starred trip to church*

***Following** a telephone call, a little earlier, Winter had said …*

*He bled profusely **following** circumcision*

Deep Learning Dependency Parser

[Chen & Manning, EMNLP 2014]

Softmax layer:

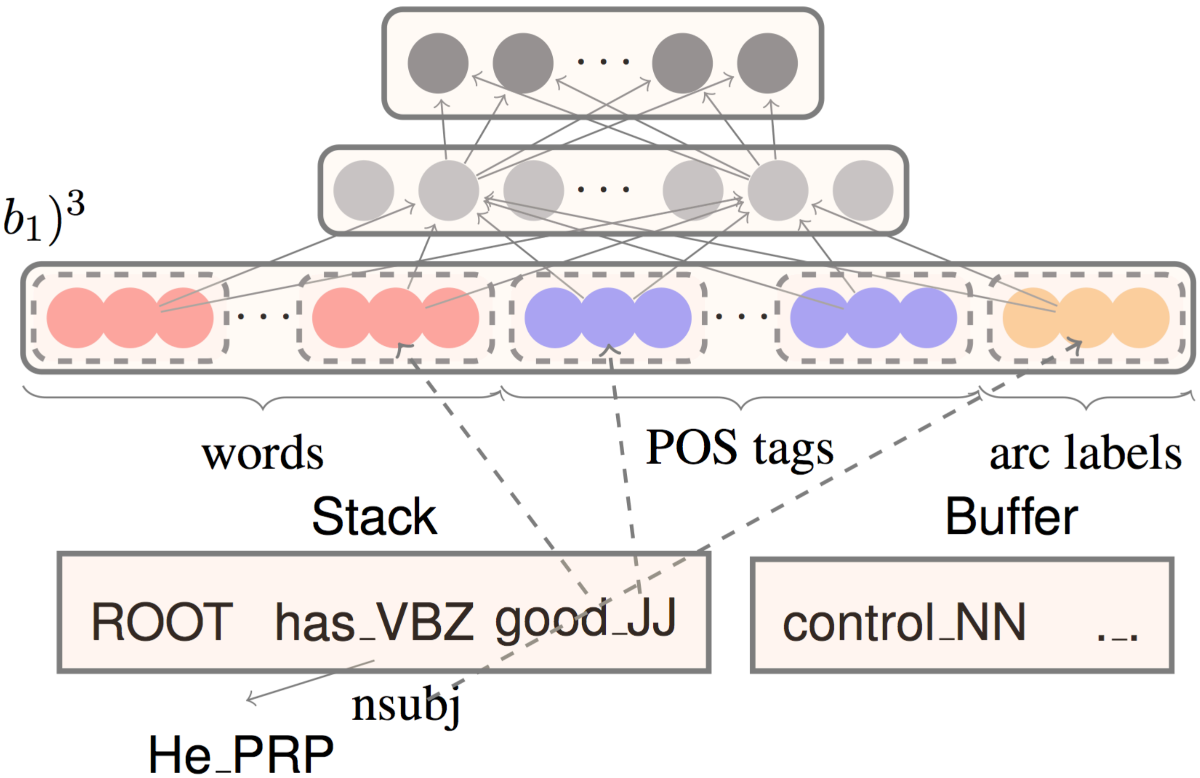
$$p = \texttt{softmax}(W_2 h)$$

Hidden layer:

$$h = (W_1^w x^w + W_1^t x^t + W_1^l x^l + b_1)^3$$

Input layer: $[x^w, x^t, x^l]$

Configuration



Parsing Speed-up

- Pre-computation trick:

	word	POS	dep.
s_1	good	JJ	
s_2	has	VBZ	
b_1	control	NN	
$lc(s_1)$	∅	∅	∅
$rc(s_1)$	∅	∅	∅
$lc(s_2)$	He	PRP	nsubj
$rc(s_2)$	∅	∅	∅
…	…	…	…

- If we have seen (s_1, good) many times in the training set, we can pre-compute matrix multiplications before parsing
  - reducing multiplications to additions.
- 8 ~ 10 times faster. As in [Devlin et al. 2014]

Deep Learning Dependency Parser

[Chen & Manning, EMNLP 2014]

Parser type	Parser	LAS (Label & Attach)	Sentences / sec
Transition-based	MaltParser (stackproj)	86.9	469
Graph-based	MSTParser	87.6	10
	TurboParser (full)	89.7	8

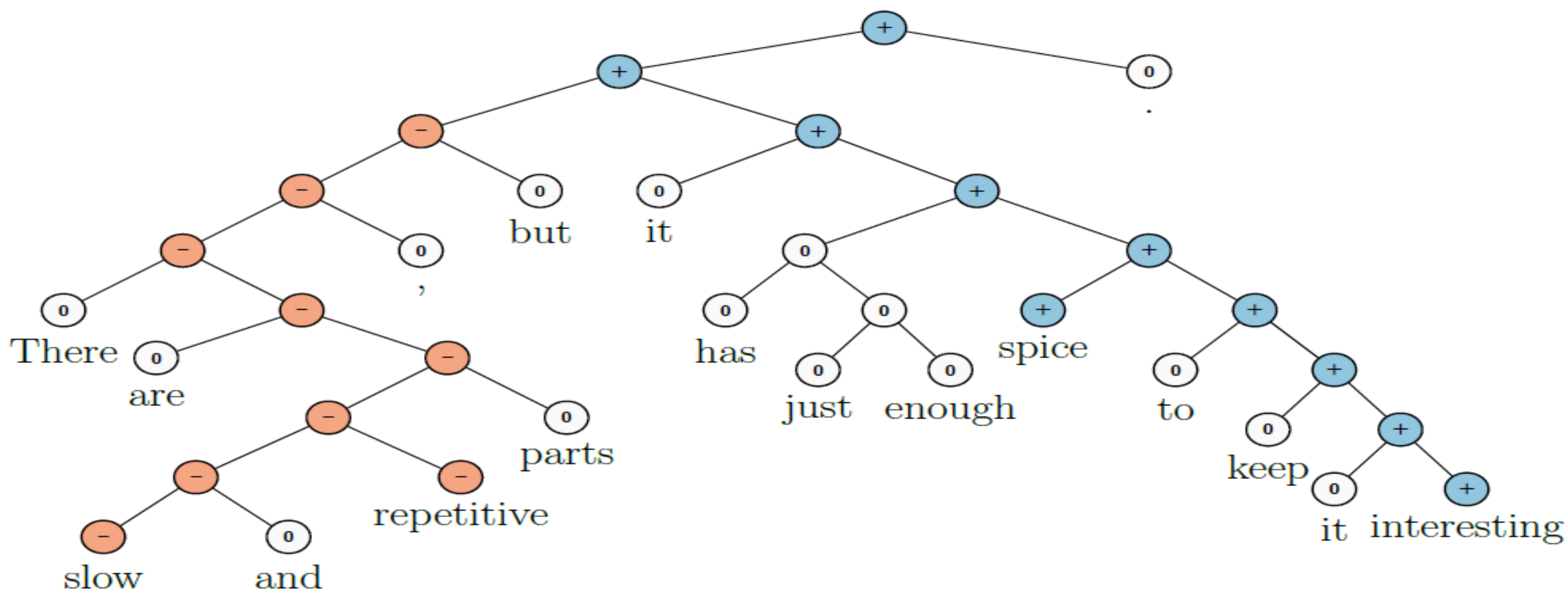
Embedding size 50, hidden size 200, mini-batch AdaGrad α=0.01, 0.5 dropout on hidden, pre-trained C&W word vectors

Christopher Manning

Sentiment Analysis with a Recursive Neural Tensor Network

An RNTN can capture contrastive sentences like *X but Y*

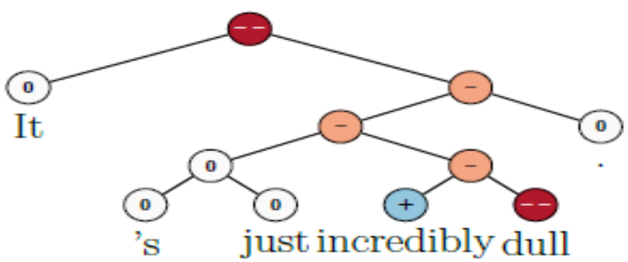
RNTN accuracy of 72%, compared to MV-RNN (65%), biword NB (58%)

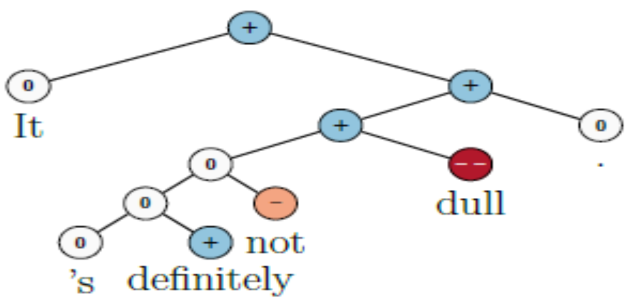


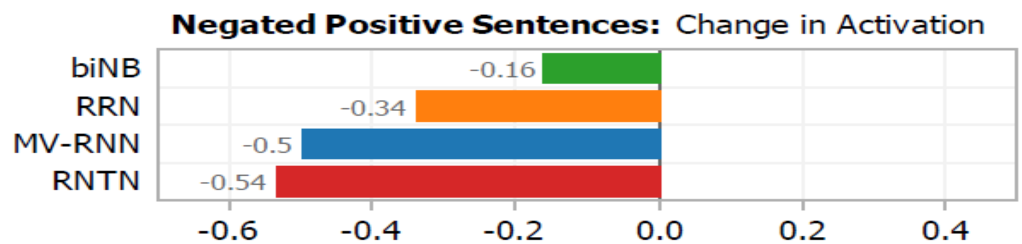
Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Chris Manning, Andrew Ng & Chris Potts. EMNLP 2013.

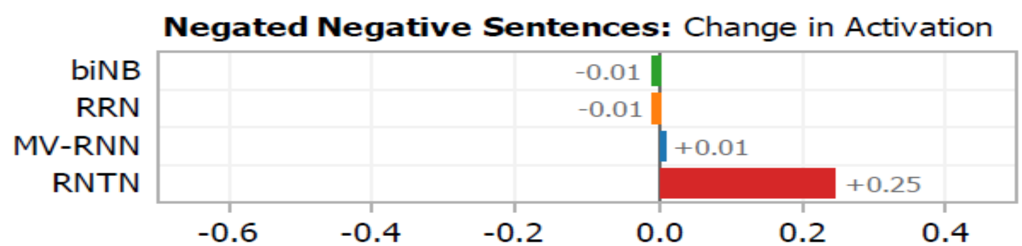
Negation Results

- When negating negatives, positive activation should increase!

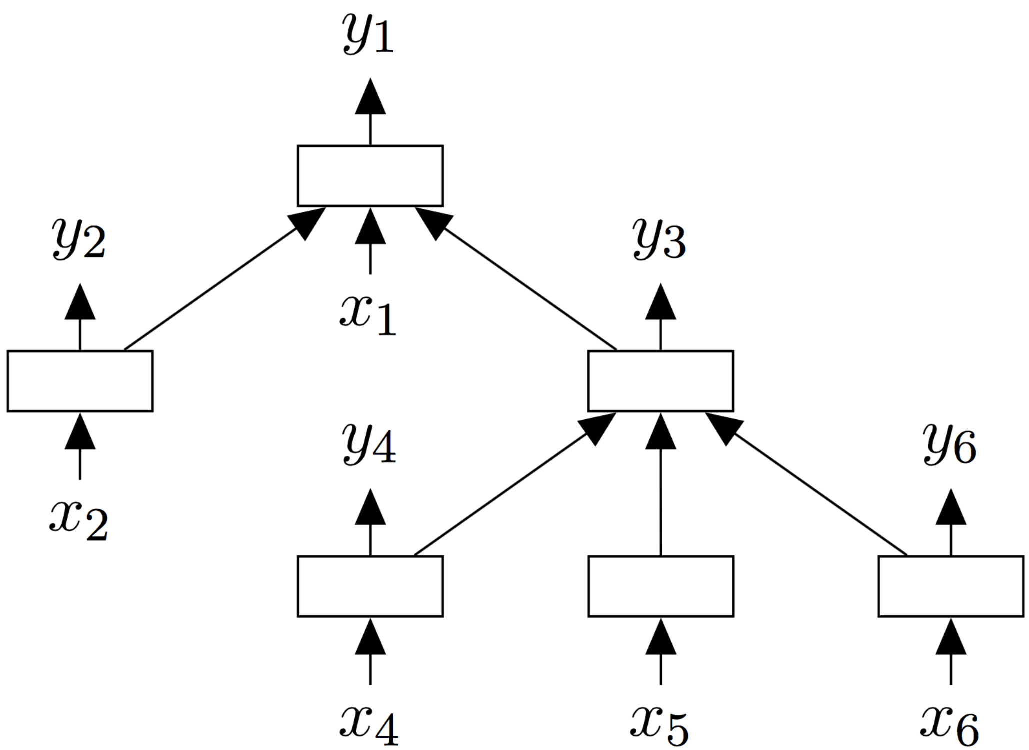








Dependency Tree LSTM similarity



Kai Sheng Tai, RIchard Socher, and Christopher D. Manning, ACL 2015

Structure gives sophisticated similarity

Word vector similarity

two men are playing guitar

some men are playing rugby

two men are talking

two dogs are playing with each other

Dependency Tree LSTM

two men are playing guitar

the man is singing and playing the guitar

the man is opening the guitar for donations and plays with the case

two men are dancing and singing in front of a crowd

Envoi

A new understanding of good word vectors

An accurate – and fast – neural network dependency parser

A sentence understanding model of sentiment analysis

Available in Stanford CoreNLP …

http://nlp.stanford.edu/software/corenlp.shtml

The key tools for building intelligent systems that can recognize and exploit the compositional semantic structure of language

Thank you!

AI2
IPsoft
Bloomberg
DARPA
NSF
Google
ebay

SAIL

The Stanford Artificial Intelligence Laboratory