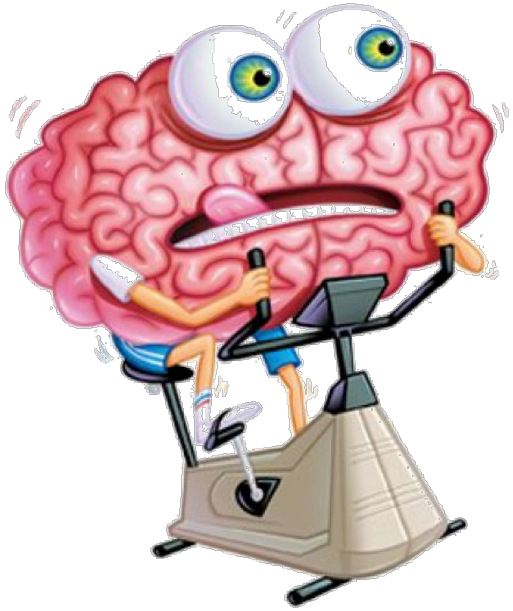
UNIVERSITEIT STELLENBOSCH UNIVERSITY

mih media lab.

A Tutorial on Deep Learning for NLP



Stephan Gouws

Several slides with kind permission of Richard Socher, Yoshua Bengio and Chris Manning [ACL, 2012]

stephan@ml.sun.ac.za

UNIVERSITEIT STELLENBOSCH UNIVERSITY

mih media lab.

Introduction

- Why the interest in deep learning?
  - Good results (POS, NER, Sentiment, ..)
  - In spite of ignoring much of NLP prior work
- "Deep" as opposed to..?
  - Not to be confused with deep vs shallow semantics

UNIVERSITEIT STELLENBOSCH UNIVERSITY

mih media lab.

1. Motivation
2. Background Theory
3. Neural Language Models
4. Other NLP Tasks
5. My Work
6. End

UNIVERSITEIT STELLENBOSCH UNIVERSITY

mih media lab.

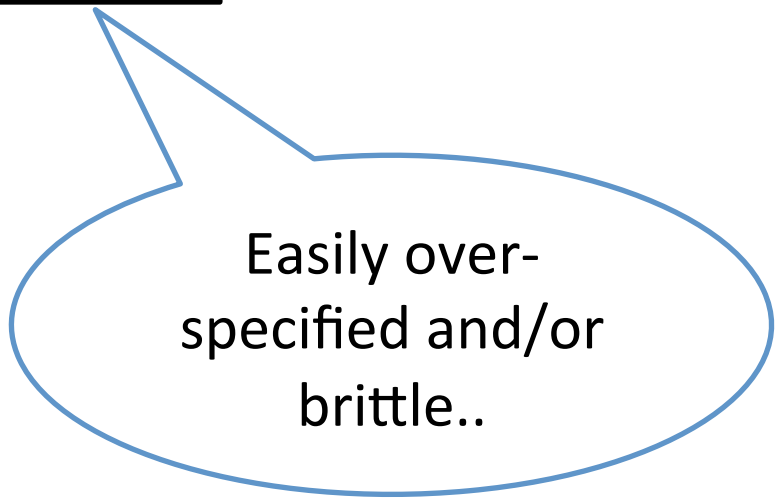
1. Motivation
2. Background Theory
3. Neural Language Models
4. Other NLP Tasks
5. My Work
6. End

UNIVERSITEIT STELLENBOSCH UNIVERSITY

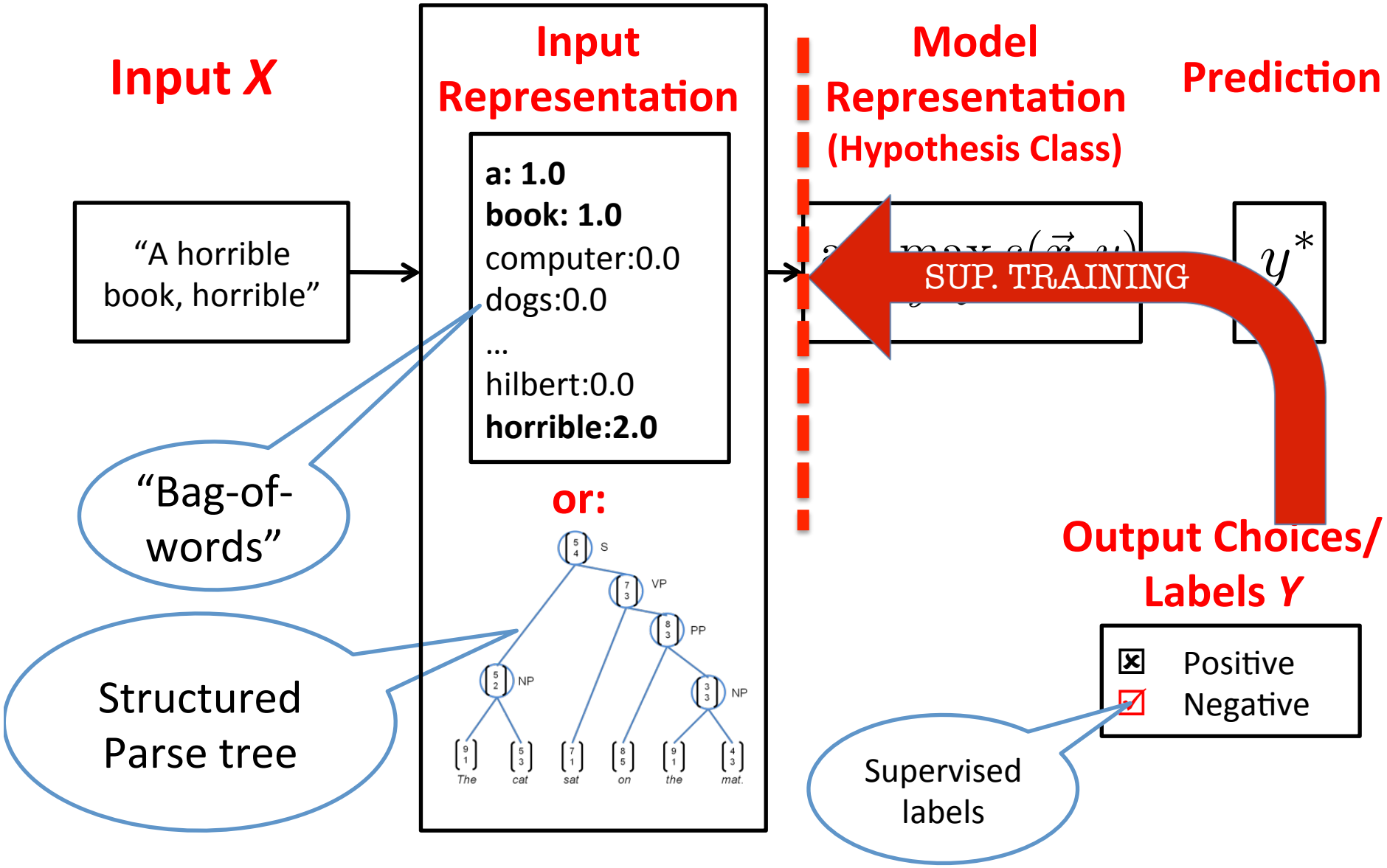
mih media lab.

NLP at a High Level

- Most approaches to NLP learn the predictive values (weights) of human-defined features w.r.t. some (un)structured classification task
- Input ➔ features ➔ classifier ➔ output



Example: Sentiment Classification



Word Representations

What is the main source of information in NLP?

Words.

How do most systems handle words?

Not very well.

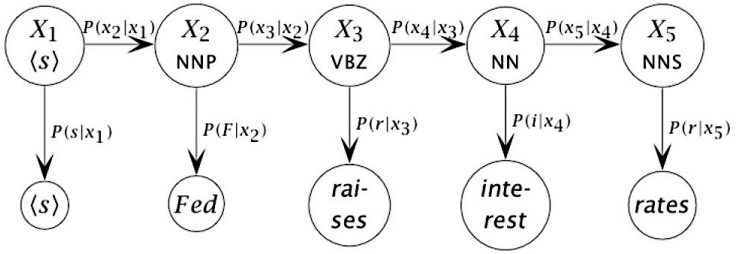
UNIVERSITEIT STELLENBOSCH UNIVERSITY

mih media lab.

Discrete Representations

- In NLP we work with *words*, *phrases* and *sentences*
- Traditionally, we treat words **categorically**
- "cat" no more similar to "dog" than to "car"
- *Consider*:

  Probabilistic models with discrete states/symbols:



***Intuitively**: humans don't operate on such a coarsely discretized representation of the world, but on a meaning continuum*

"Local" Word Representations

Discrete Symbols in Vector Form

The vast majority of NLP work regards words as atomic symbols:
hotel, conference, walk

Regardless of whether it is rule-based NLP or statistical NLP

In vector space terms, this is a vector with one 1 and **a lot** of zeroes

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

Dimensionality: 20K (speech) – 50K (PTB) – 500K (big vocab) – 13M (Google 1T)

We call this a "one-hot" representation

Can one do better?

"Local" Word Representations

Discrete Symbols in Vector Form

A symbolic representation looks ridiculous as a vector …

But it's what vector space model IR actually uses (conceptually)

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

UNIVERSITEIT STELLENBOSCH UNIVERSITY

mih media lab.

Continuous-space Representations

- Vector space models represent words and documents within a continuous, latent (LSA) or explicit, syntactic-semantic feature space.
- Distributional semantics
  - E.g. TF-IDF [Salton & McGill, 1983]
- Vector space defines a simple, principled distance metric over representations
  - E.g. now: sim(cat, dog) > sim(cat, car)

Vector Representations

In all of these approaches, including deep learning models, a word is represented as a dense vector:

"linguistics" =

$$\begin{bmatrix} 0.286 \\ 0.792 \\ -0.177 \\ -0.107 \\ 0.109 \\ -0.542 \\ 0.349 \\ 0.271 \end{bmatrix}$$

Some approaches aim for some sparsity in the vector
In probabilistic approaches, all the numbers are non-negative.

UNIVERSITEIT STELLENBOSCH UNIVERSITY

mih media lab.

Distributed Representations

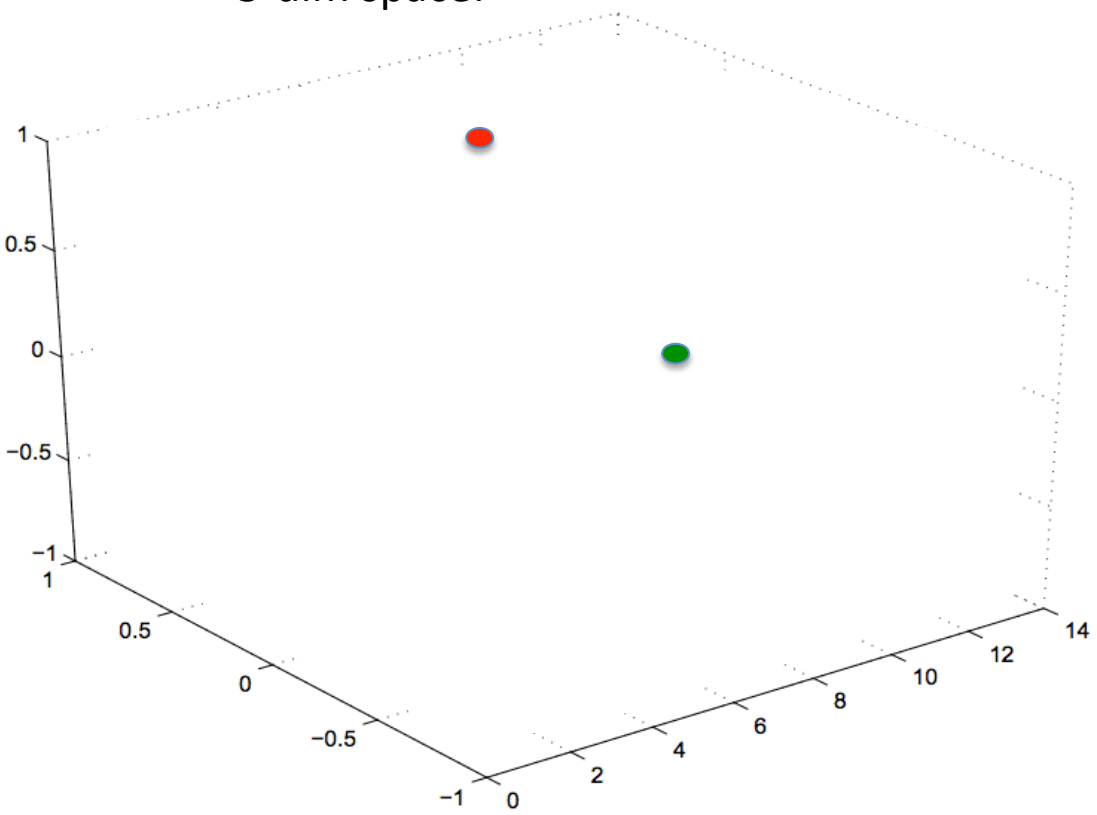
[Hinton 1984]

- "[…] each entity is represented by a pattern of activity *distributed* over many computing elements, and each computing element is involved in representing many different entities".
- Leads to compact, more efficient encodings than ***local*** representations

The Manifold Hypothesis

[Cayton, 2005]

Input: 2 points in 3-dim space:



UNIVERSITEIT STELLENBOSCH UNIVERSITY

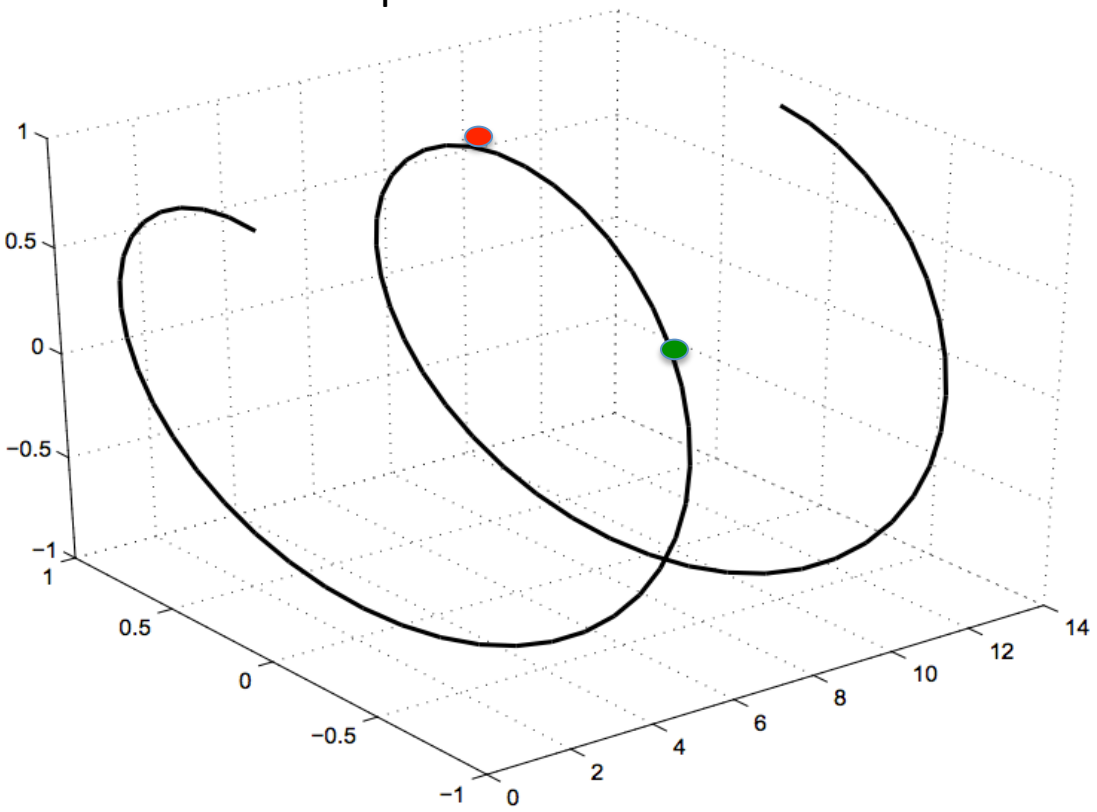
mih media lab.

The Manifold Hypothesis

[Cayton, 2005]

Underlying data generating mechanism:

3-dim space:

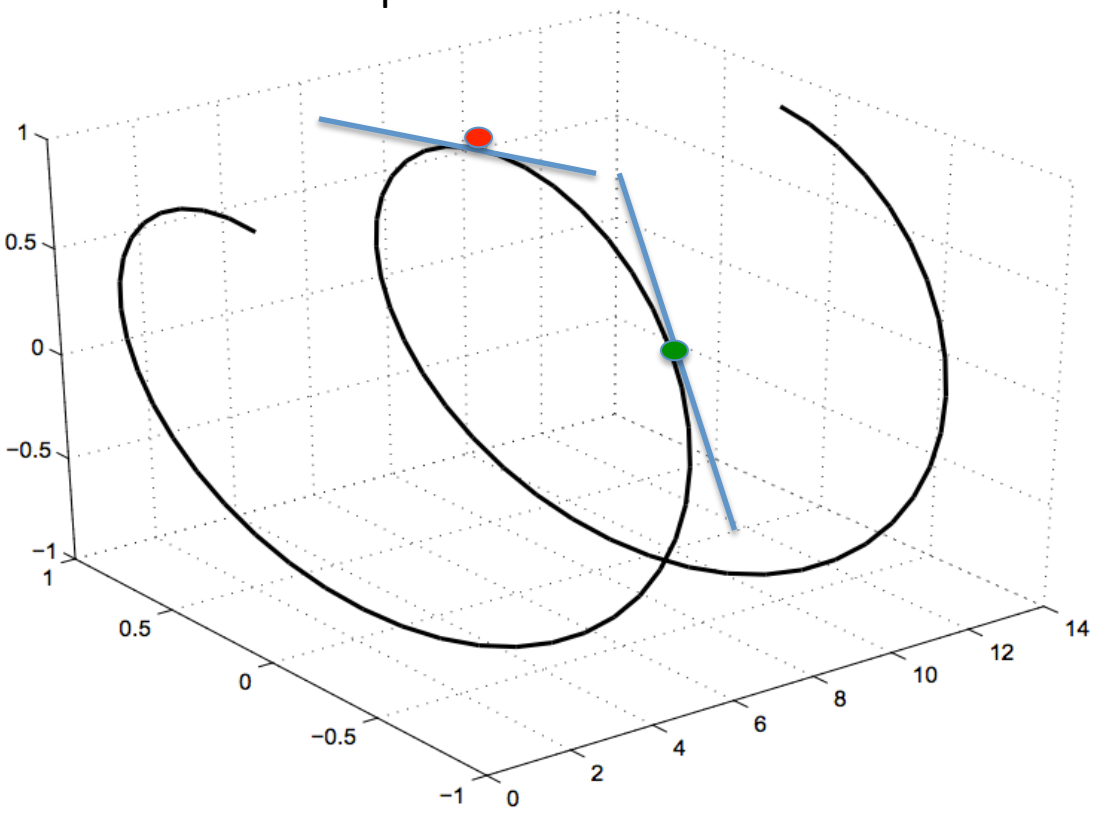


The Manifold Hypothesis

[Cayton, 2005]

1-dim manifold ***embedded*** in 3-dim space:

3-dim space:



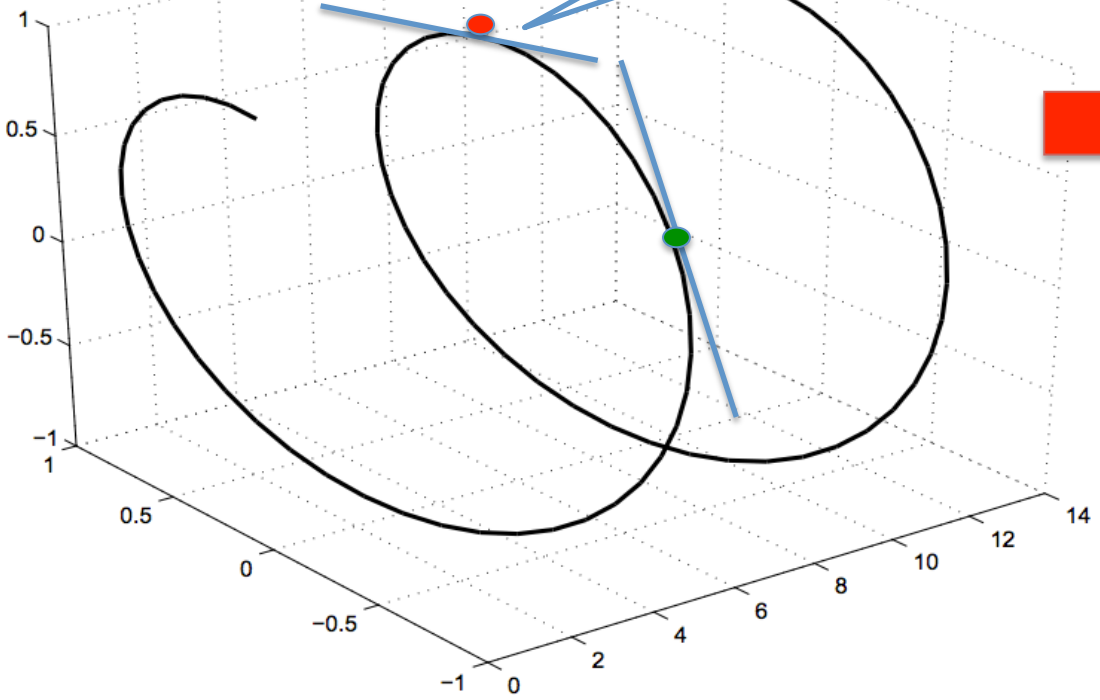
The Manifold Hypothesis

[Cayton, 2005]

1-dim manifold ***embedded*** in 3-dim space:

3-dim space:

Locally resembles $\mathcal{R}^1$



1-dim space:

"Structure" of the data/ (distance between points) is preserved when mapped into the lower dim space

UNIVERSITEIT STELLENBOSCH UNIVERSITY

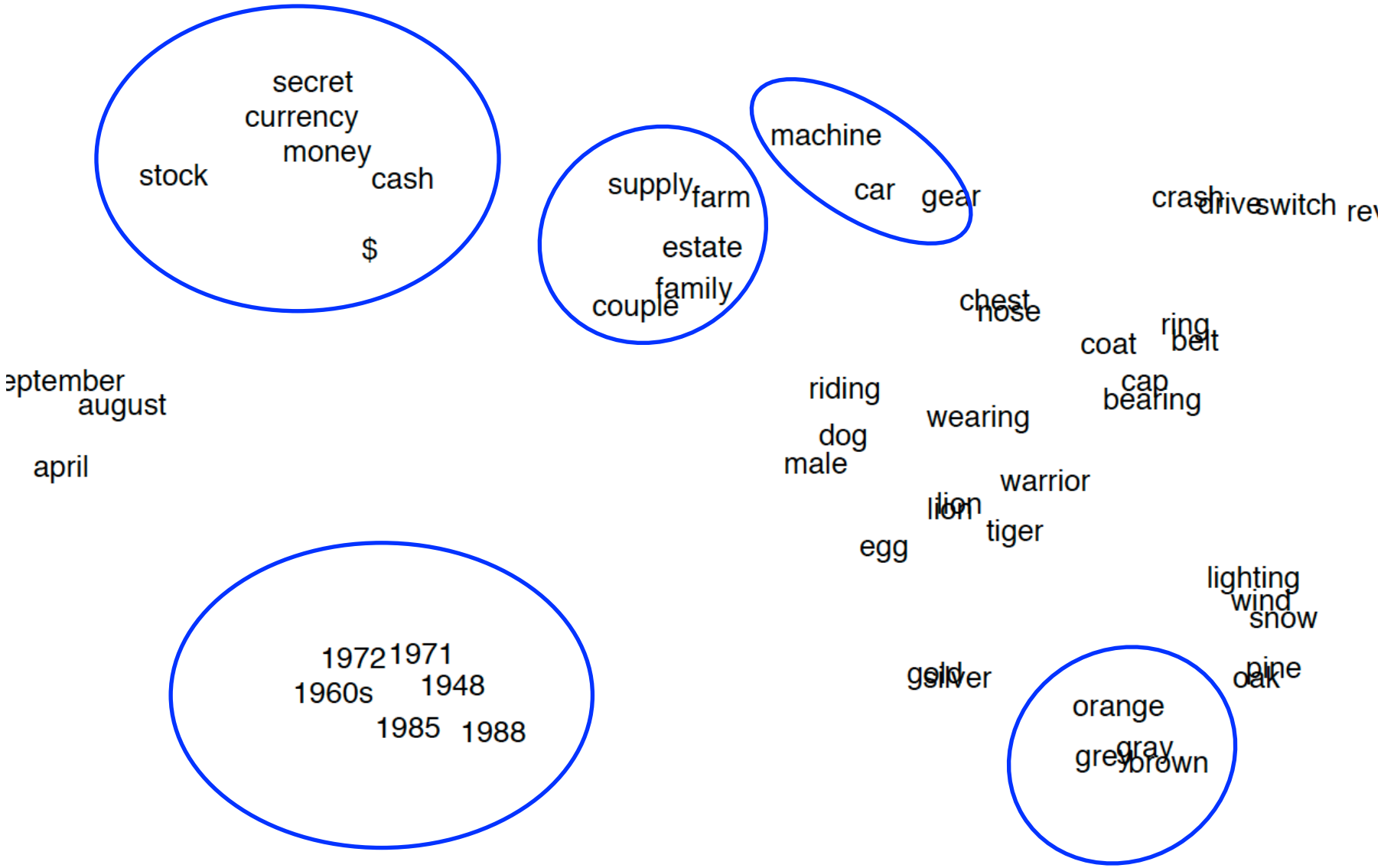
mih media lab.

The Manifold Hypothesis

[Cayton, 2005]

"The probability mass P(x) of real-world data x presented in a high-dimensional input space is expected to concentrate in the vicinity of a manifold of ***much lower dimensionality****, embedded in the high-dimensional input space".*

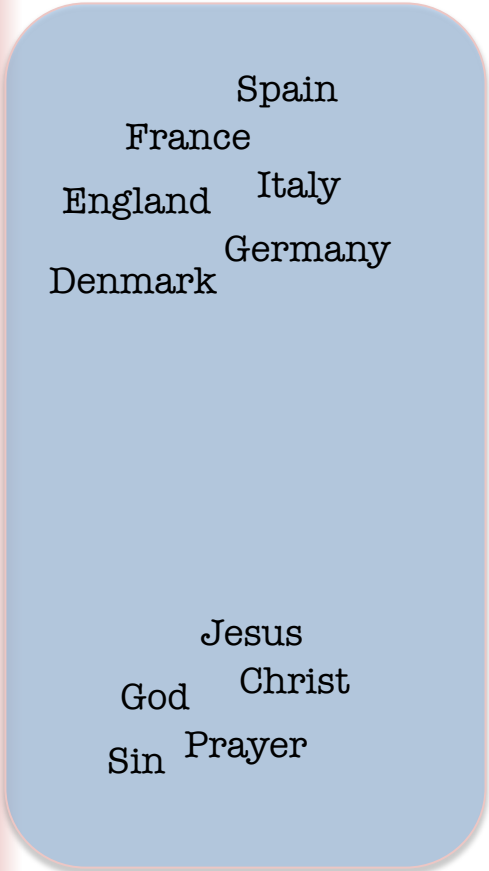
Neural Word Embeddings



Neural Word Embeddings

[Collobert + Weston, ICML 2008]

Most similar words to a few words



France	Jesus	XBOX	Reddish	Scratched
Spain	Christ	Playstation	Yellowish	Smashed
Italy	God	Dreamcast	Greenish	Ripped
Russia	Resurrection	PS###	Brownish	Brushed
Poland	Prayer	SNES	Bluish	Hurled
England	Yahweh	WH	Creamy	Grabbed
Denmark	Josephus	NES	Whitish	Tossed
Germany	Moses	Nintendo	Blackish	Squeezed
Portugal	Sin	Gamecube	Silvery	Blasted
Sweden	Heaven	PSP	Greyish	Tangled
Austria	Salvation	Amiga	Paler	Slashed

UNIVERSITEIT STELLENBOSCH UNIVERSITY

mih media lab.

Advantages of the Neural Word Embedding Approach

Compared to a method like LSA:

- Higher-level features (beyond the word)
- Forces a representation on the words themselves that is better and more meaningful
  - Because everything is trained together
- By adding supervision from either one task or multiple tasks simultaneously, we can improve the representation of the words for handling language analysis tasks

UNIVERSITEIT STELLENBOSCH UNIVERSITY

mih media lab.

From Learning ***Weights*** to Learning ***Representations***

- Usually we learn the predictive *weights* of human-defined features
- In deep learning we aim to learn **new representations** of the data at increasing levels of abstraction, which may then be used as features

Hypothesis: Features good for characterizing **P(input)** also good for **P(label|input)**

*"We have linguistic representations? And LSA/LDA/… can **learn** representations?"*

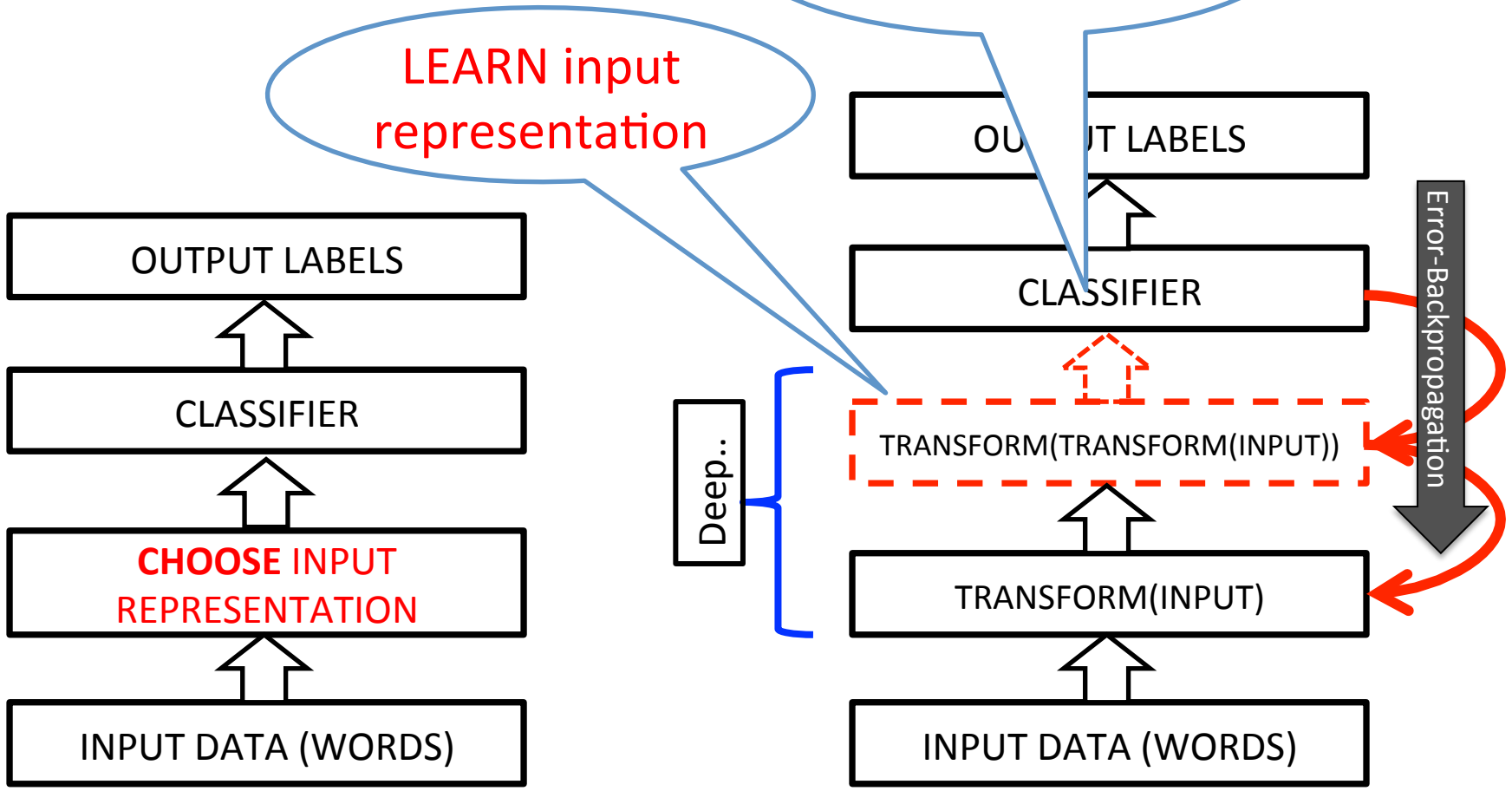
- Manual:
  - POS tags, HPSG, other grammar formalisms..
- Unsupervised:
  - Latent Semantic Analysis: SVD(Term-Doc Matrix)
  - pLSA/Latent Dirichlet Allocation: Prob dim reduction
  - Principal components analysis
  - Brown clustering – Hierarchical clustering over words based on average mutual information

Deep learning is an ***adaptive*** (task-dependent), ***non-linear*** framework for learning representations *at increasing levels of generality*.

Shallow Hal

vs

Deep L

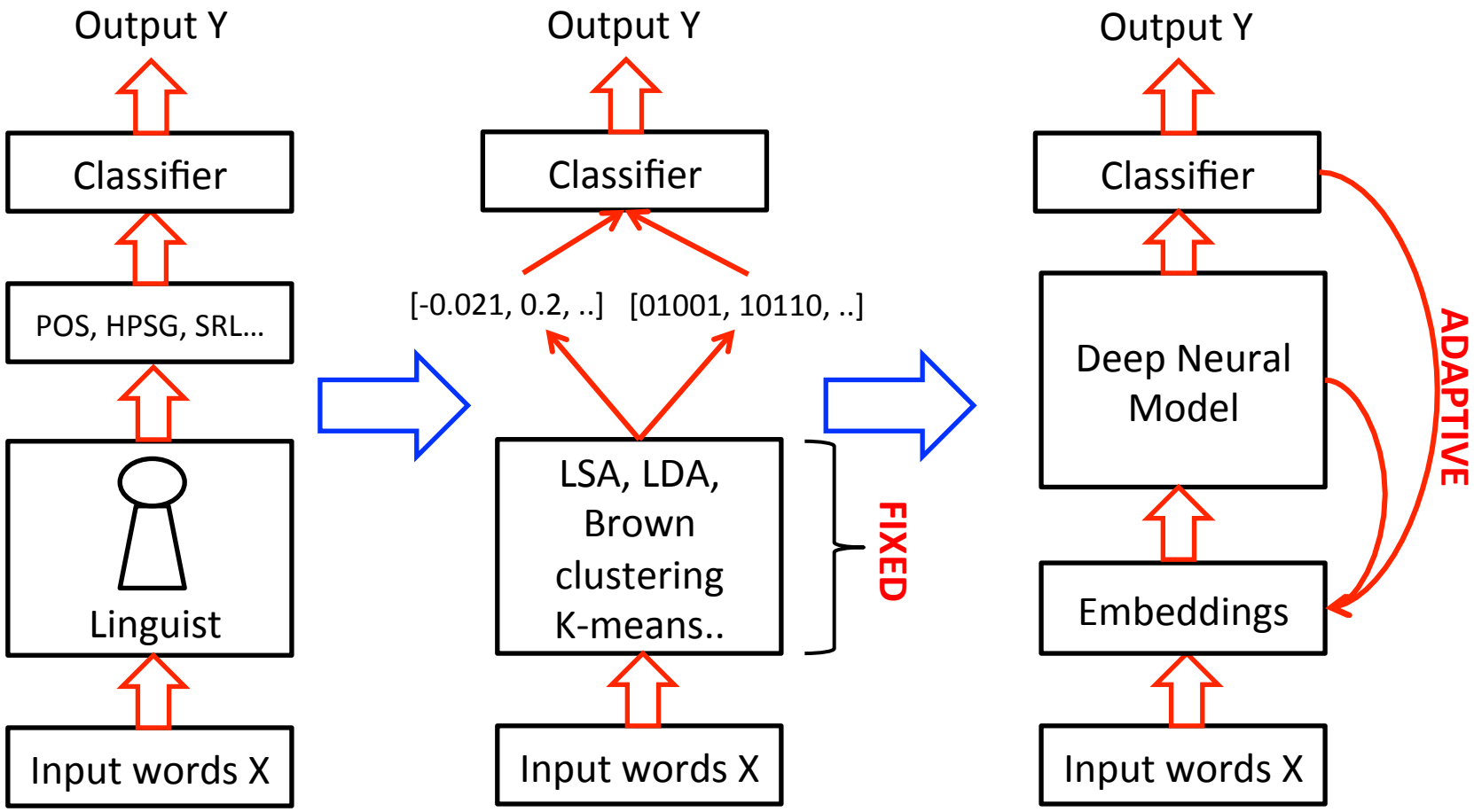


DL(Co-occurrence) ➔ Meaning?

- The Distributional Hypothesis [Firth, 1957] claims some relationship $f : W \times C \rightarrow M$ between the co-occurrence statistics *C* and meaning *M* of some word *W*.
- One can view methods for inducing word representations as attempts to uncover *M* from *C*:
  - LSA and other linear methods assume *f* is a linear function
  - Brown/k-means clustering make stronger, non-parametric *a priori* assumptions on *f*

Deep learning may be a useful **data-driven** approach for learning more informative, task-dependent representations of language

Evolution of Linguistic Representations



"Linguist-in-a-box"

Standard Distributional Representations

Deep Learning: Task-adaptive, Distributed Representations

UNIVERSITEIT STELLENBOSCH UNIVERSITY

mih media lab.

Pfft.. LSA also gives..“vectors”!

Yes, but LSA is deterministic, these models are adaptive..

UNIVERSITEIT STELLENBOSCH UNIVERSITY

mih media lab.

1. Motivation
2. Background Theory
    1. Neural Networks: The Basics
    2. Inference: Forward propagation
    3. Training: Error Backpropagation
3. Neural Language Models
4. Other NLP Tasks
5. My Work
6. End

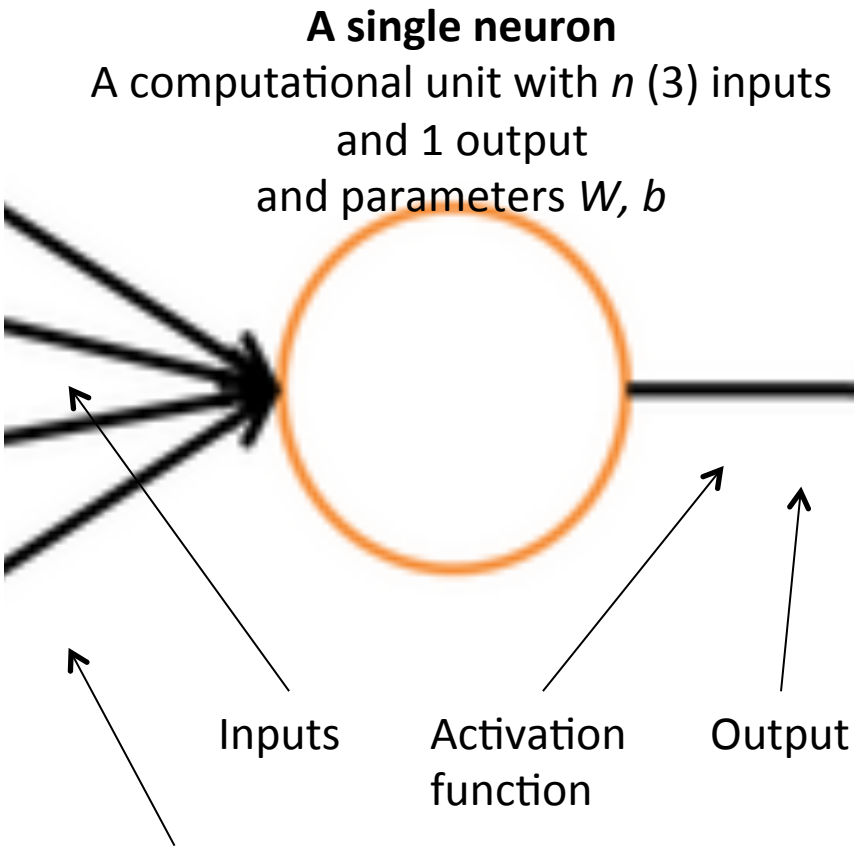
UNIVERSITEIT STELLENBOSCH UNIVERSITY

mih media lab.

1. Motivation
2. Background Theory
    1. Neural Networks: The Basics
    2. Inference: Forward propagation
    3. Training: Error Backpropagation
3. Neural Language Models
4. Other NLP Tasks
5. My Work
6. End

Demystifying Neural Networks

- Neural networks come with their own terminological baggage (of “neurons”, “activation functions”, and “weight decay”)
- But if you understand how maxent/logistic regression models work
  - Then you already understand the operation of a basic neural network neuron!

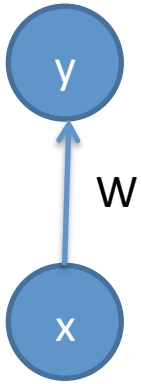


UNIVERSITEIT STELLENBOSCH UNIVERSITY

mih media lab.

Linear Regression

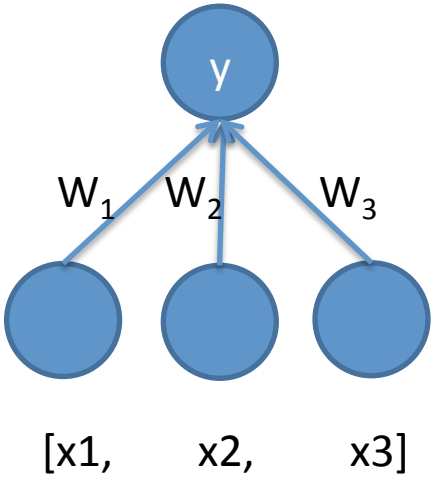
- $y = Wx + b$

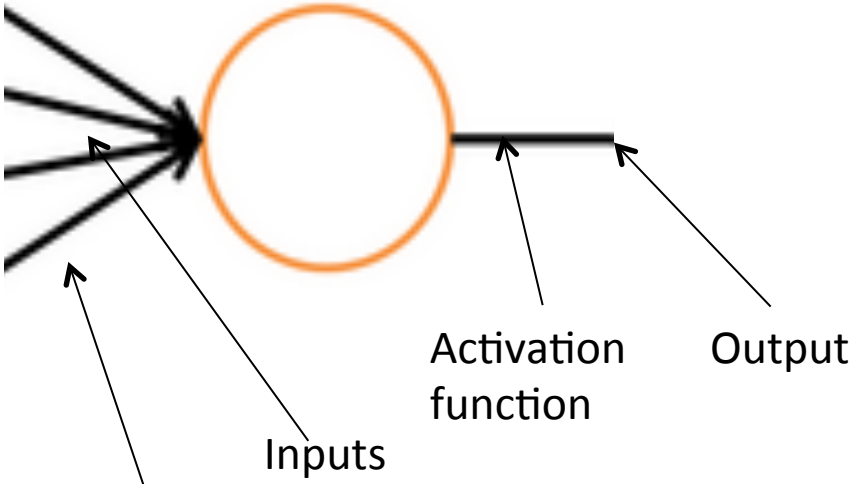


From Linear Regression to Linear NN

- $y = \mathbf{W}\mathbf{x} + b$

  $= w_1x_1 + w_2x_2 + w_3x_3 + b$





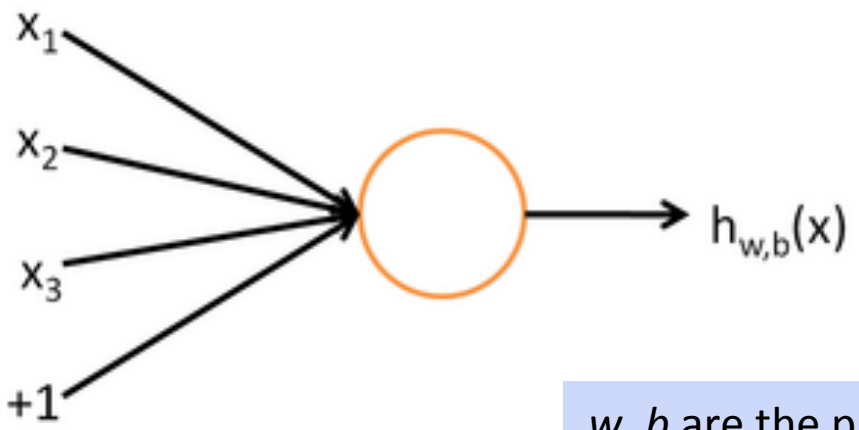
UNIVERSITEIT STELLENBOSCH UNIVERSITY

mih media lab.

A Neuron as Logistic Regression

$$h_{w,b}(x) = f(w \cdot x + b)$$

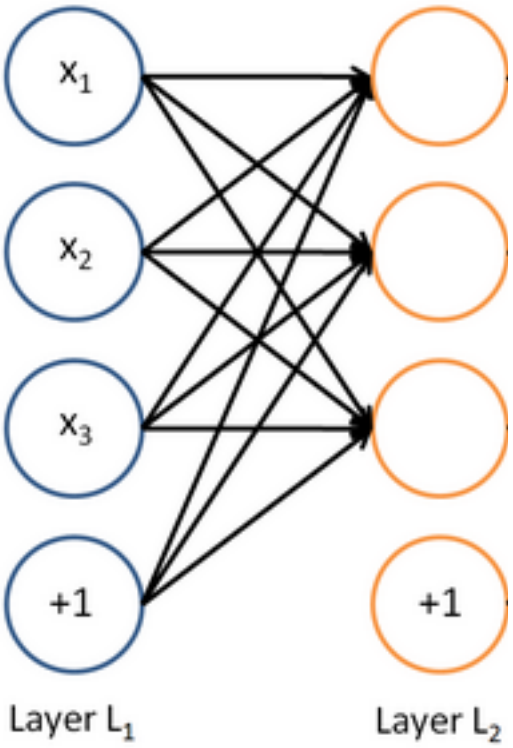
$$f(z) = \frac{1}{1 + e^{-z}}$$



w, *b* are the parameters of this neuron
i.e., this logistic regression model

A neural network = running several logistic regressions at the same time

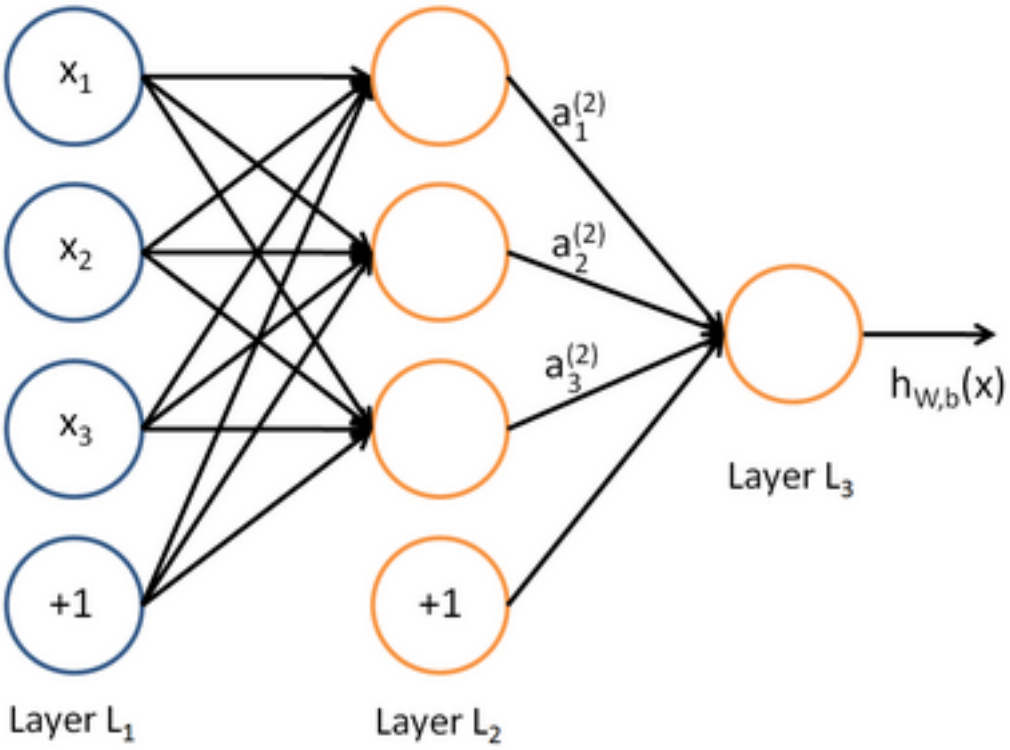
If we feed a vector of inputs through a bunch of logistic regression functions, then we get a vector of outputs



But we don't have to decide ahead of time what variables these logistic regressions are trying to predict!

A neural network = running several logistic regressions at the same time

… which we can feed into another logistic regression function



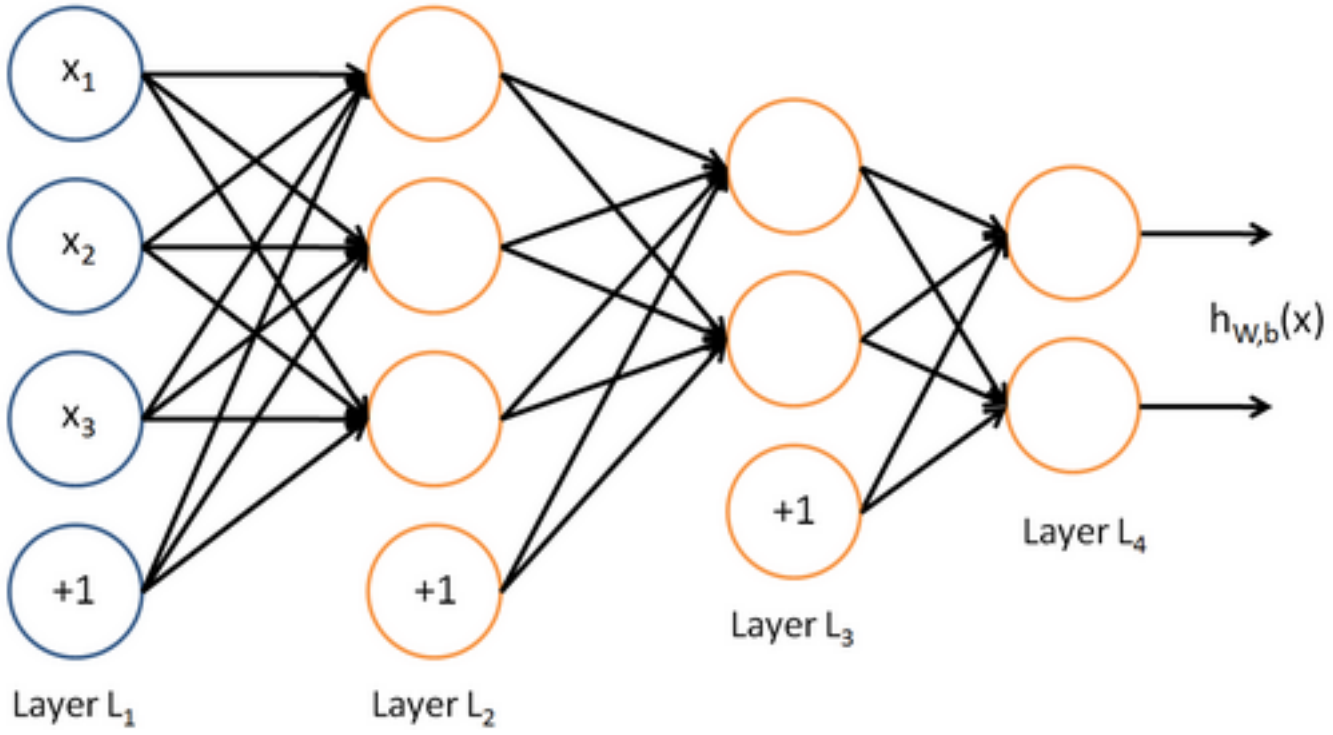
and it is the training criterion that will decide what those intermediate binary target variables should be, so as to make a good job of predicting the targets for the next layer, etc.

UNIVERSITEIT STELLENBOSCH UNIVERSITY

mih media lab.

A neural network = running several logistic regressions at the same time

Before we know it, we have a multilayer neural network….



UNIVERSITEIT STELLENBOSCH UNIVERSITY

mih media lab.

Non-linearities: Why they're needed

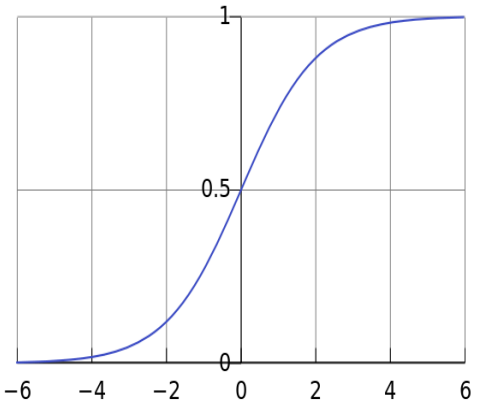
- For logistic regression, they're motivated by mapping to probabilities: [0,1] ("squashing functions")

- Here, they're motivated by being able to do *function approximation*:
  - Without non-linearities, neural networks can't do anything more than a linear transform: extra layers could just be compiled down into a single linear transform
  - The probabilistic interpretation for hidden units is usually unnecessary except in the Boltzmann machine models.

Non-linearities: What's used

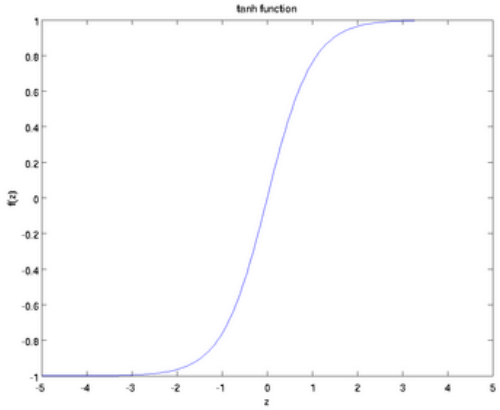
logistic ("sigmoid")

$$f(z) = \frac{1}{1 + \exp(-z)}.$$



tanh

$$f(z) = \tanh(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}},$$



- tanh() is just a rescaled and shifted sigmoid (2 × as steep, [−1,1]):

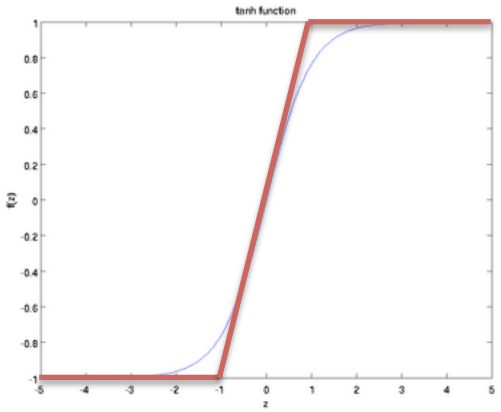
$$\tanh(z) = 2\text{logistic}(2z) - 1$$

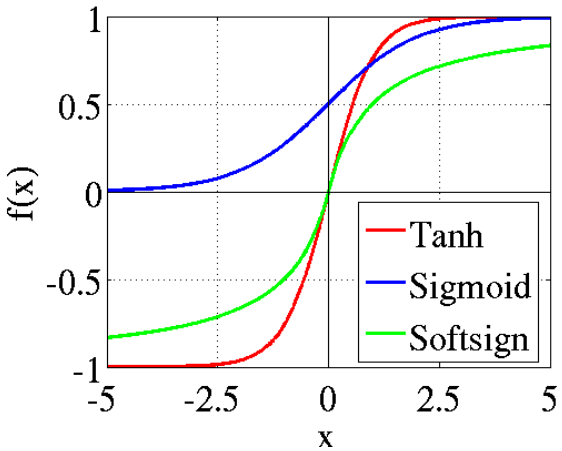
- tanh is what is most used and often performs best for deep nets
  [Glorot and Bengio *AISTATS* 2010]

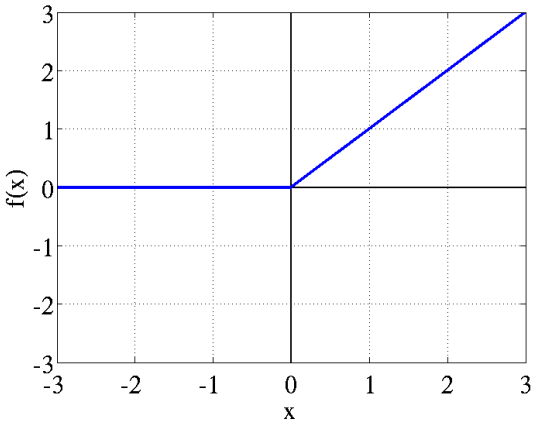
Non-linearities: There are various other choices

hard tanh soft sign rectifier

$$\text{HardTanh}(x) = \begin{cases} -1 & \text{if } x < -1 \\ x & \text{if } -1 <= x <= 1 \\ 1 & \text{if } x > 1 \end{cases} \qquad \text{softsign}(z) = \frac{a}{1+|a|} \qquad \text{rect}(z) = \max(z, 0)$$







- hard tanh is a mathematically awkward but computationally cheap tanh
- [Glorot and Bengio, *AISTATS*, 2010] discuss uses of softsign and rectifier

UNIVERSITEIT STELLENBOSCH UNIVERSITY

mih media lab.

Summary of Terminology

- "*Neuron*" = logistic regression or similar function
- "*Input layer*" = input training/test vector
- "*Bias unit*" = intercept term
- "*Activation function*" is a sigmoid (or similar nonlinearity)
- "*Activation*" = response
- "*Backpropagation*" = running stochastic gradient descent across a multilayer network
- "*Weight decay*" = regularization / Bayesian prior

UNIVERSITEIT STELLENBOSCH UNIVERSITY

mih media lab.

1. Motivation
2. Background Theory
    1. Neural Networks: The Basics
    2. Inference: Forward propagation
    3. Training: Error Backpropagation
3. Neural Language Models
4. Other NLP Tasks
5. My Work
6. End

Fprop: Forward Propagation

We have

$$a_1 = f(W_{11}x_1 + W_{12}x_2 + W_{13}x_3 + b_1)$$

$$a_2 = f(W_{21}x_1 + W_{22}x_2 + W_{23}x_3 + b_2)$$

etc.

In matrix notation

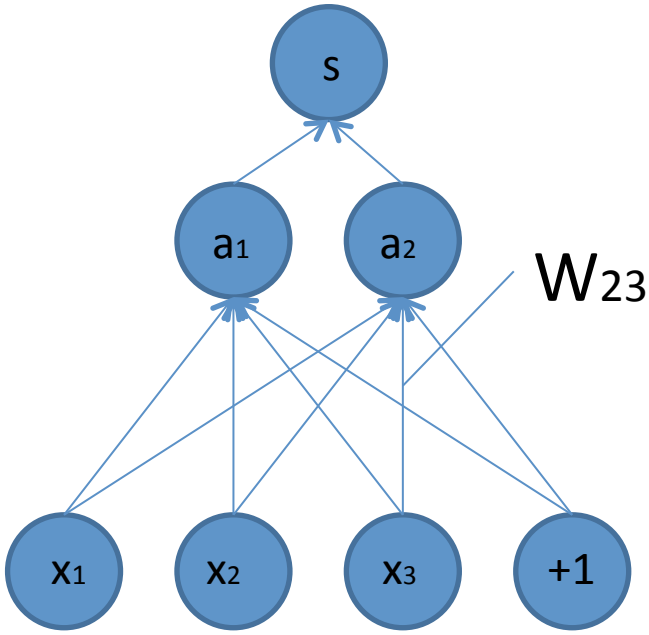
$$z = Wx + b$$

$$a = f(z)$$

Called "forward propagation" / fprop

where *f* is applied element-wise:

$$f([z_1, z_2, z_3]) = [f(z_1), f(z_2), f(z_3)]$$



UNIVERSITEIT STELLENBOSCH UNIVERSITY

mih media lab.

1. Motivation
2. Background Theory
    1. Neural Networks: The Basics
    2. Inference: Forward propagation
    3. Training: Error Backpropagation
3. Neural Language Models
4. Other NLP Tasks
5. My Work
6. End

UNIVERSITEIT STELLENBOSCH UNIVERSITY

mih media lab.

Training

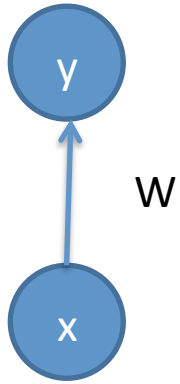
Error functions

- Fprop on input *x* using parameters *W* gets us a prediction *y'*
- What we really want is the target *y**
- How do we "tune" the model to do better?
- First, we need an indication of how badly the model is doing (a cost/error function *J*):
  - Sum-of-squared errors (real-valued outputs)
  - Negative log-likelihood (NLL, for probabilities)
  - Cross-entropy

Training:

A Single Linear Unit

- Consider a **single** linear unit:
  - $y = W^T x + b$, or $W^T x'$ for $x' = [x, 1]$
- Let the input be X and the desired real-valued output Y
- The parameters of this model is fully specified by the vector W
- How do we "tune" W to obtain a "better" model?



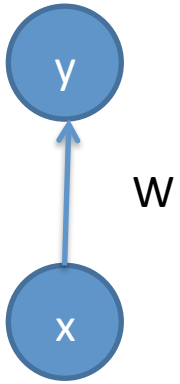
UNIVERSITEIT STELLENBOSCH UNIVERSITY

mih media lab.

Training:

A Single Linear Unit

- Approach: Define a cost function *J* i.t.o. W
  - Since y is a *real number*, we use the sum-of-squared errors
  - Let $J = \frac{1}{2} \Sigma(y' - y^*)^2 = \frac{1}{2} \Sigma(W^T x - y^*)^2$
- Compute the derivatives of J wrt W:



$$\text{Let } J = \frac{1}{2} \sum_{i=1}^{N} (y_i' - y_i)^2 \text{ and } y_i' = W^T x_i$$

$$\frac{\partial J}{W} = \frac{\partial J}{y'} \frac{\partial y'}{W}$$

$$= (\sum_{i=1}^{N} (y_i' - y_i))(x_i)$$

$$\frac{\partial J}{\partial w_i} = (w_i x_i - y)(x_i)$$

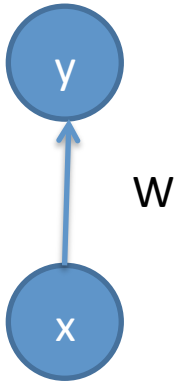
UNIVERSITEIT STELLENBOSCH UNIVERSITY

mih media lab.

Training:

A Single Logistic Unit

- Consider a single binary logistic unit:
  - $P(y=1|x) = 1 / (1 + \exp(-z)) = \text{sigm}(z)$
  - $z = W^T x + b$, or $W^T x'$ for $x' = [x, 1]$
- Let the input be X and the desired output Y
- The parameters of this model is fully specified by the vector W
- How do we "tune" W to obtain a "better" model?



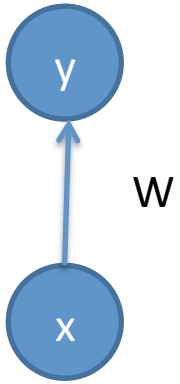
UNIVERSITEIT STELLENBOSCH UNIVERSITY

mih media lab.

Training:

A Single Logistic Unit

- Approach: Define a cost function *J* i.t.o W
  - Since *y* is a *probability*, use neg. log likelihood
  - Let J = NLL(y) = - log P($y|x$) = - log(sigm($W^T x$))
- Compute the derivatives of J wrt W:
  - Let $z = W^T x$, then by the chain rule:



$$\frac{\partial J}{\partial W} = \frac{\partial J}{\partial z}\frac{\partial z}{\partial W}$$

$$= -\frac{\partial \log(\sigma(z))}{\partial z}\frac{\partial z}{\partial W}$$

$$= -\frac{1}{\sigma(z)}\frac{\partial \sigma(z)}{\partial z}x$$

$$= -(1-\sigma(W^T x))x$$

Useful fact:

$$\frac{\partial \sigma(z)}{\partial z} = \sigma(z)(1-\sigma(z))$$

UNIVERSITEIT STELLENBOSCH UNIVERSITY

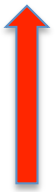
mih media lab.

Training

Stochastic Gradient Descent

- Once we know how badly the model is doing, and how its parameters influence its error, we can make updates to the parameters to decrease this error:

$$\Delta W^{t+1} \leftarrow W^t - \alpha \frac{\partial J}{\partial W}$$



"learning rate"

UNIVERSITEIT STELLENBOSCH UNIVERSITY

mih media lab.

Training with Backprop

- Compute gradient of example-wise loss wrt parameters

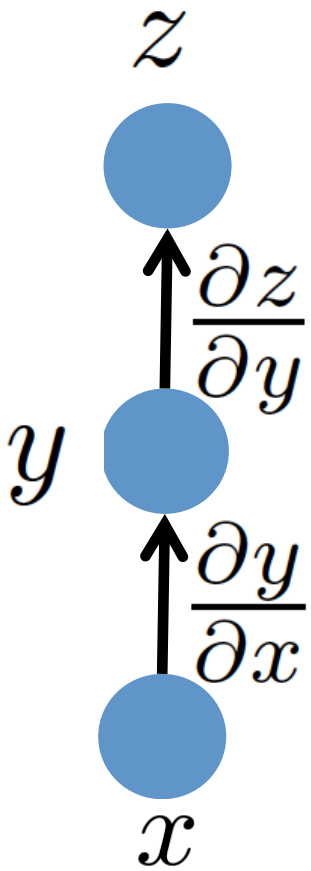
- Simply applying the derivatives chain rule wisely

$$z = f(y) \quad y = g(x) \quad \frac{\partial z}{\partial x} = \frac{\partial z}{\partial y}\frac{\partial y}{\partial x}$$

- If computing the loss(example,parameters) is O(n) computation then so is computing the gradient

Simple Chain Rule

Let: z=f(y), y=g(x), i.e. z=f(g(x))



$$\Delta z = \frac{\partial z}{\partial y}\Delta y$$

$$\Delta y = \frac{\partial y}{\partial x}\Delta x$$

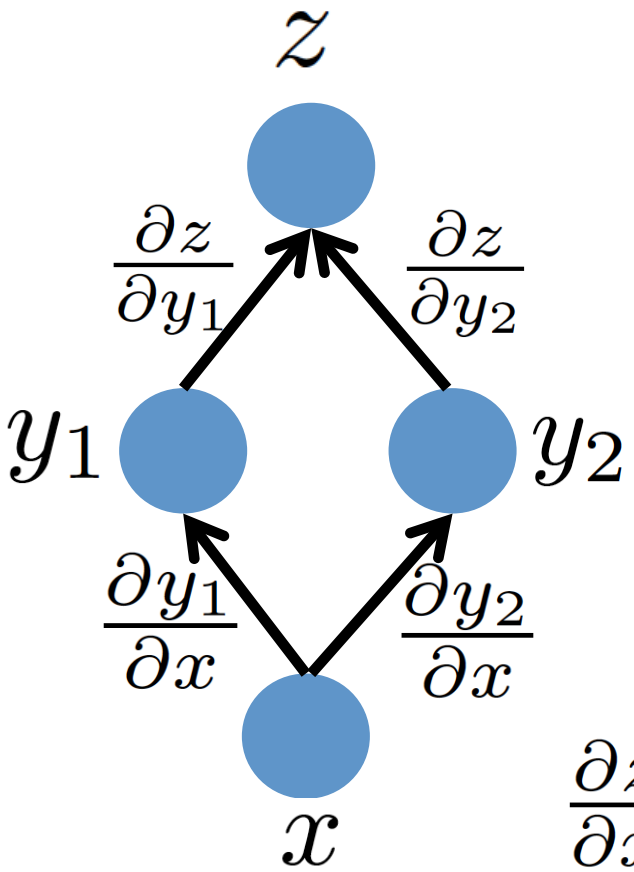
$$\Delta z = \frac{\partial z}{\partial y}\frac{\partial y}{\partial x}\Delta x$$

$$\frac{\partial z}{\partial x} = \frac{\partial z}{\partial y}\frac{\partial y}{\partial x}$$

UNIVERSITEIT STELLENBOSCH UNIVERSITY

mih media lab.

Multiple Paths Chain Rule

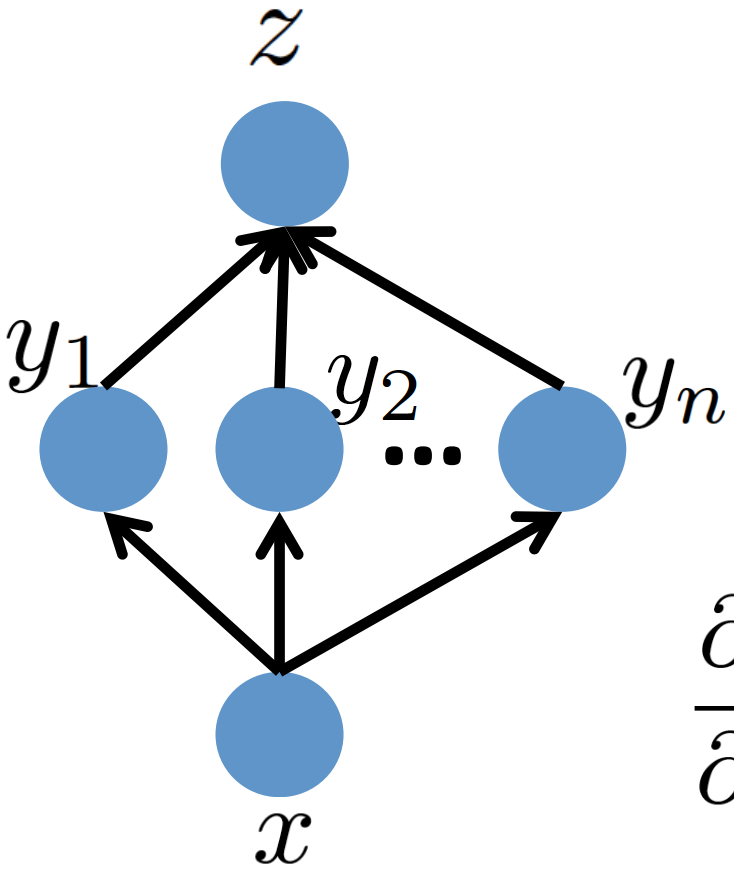


$$\frac{\partial z}{\partial x} = \frac{\partial z}{\partial y_1}\frac{\partial y_1}{\partial x} + \frac{\partial z}{\partial y_2}\frac{\partial y_2}{\partial x}$$

UNIVERSITEIT STELLENBOSCH UNIVERSITY

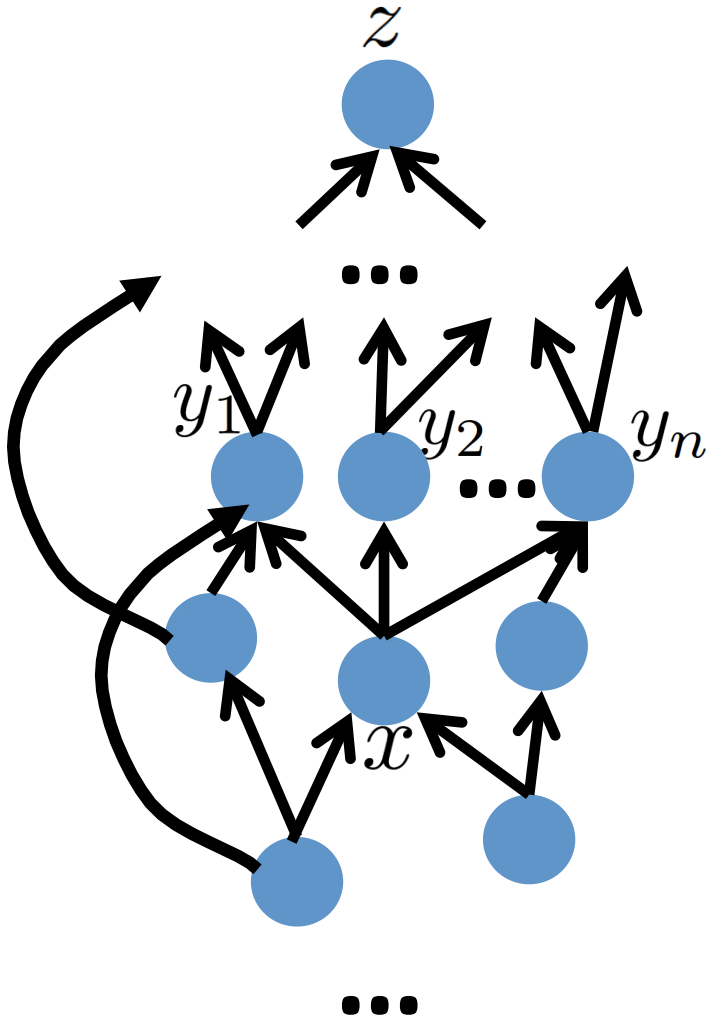
mih media lab.

Chain Rule in Flow Graph



$$\frac{\partial z}{\partial x} = \sum_{i=1}^{n} \frac{\partial z}{\partial y_i} \frac{\partial y_i}{\partial x}$$

Chain Rule in Flow Graph

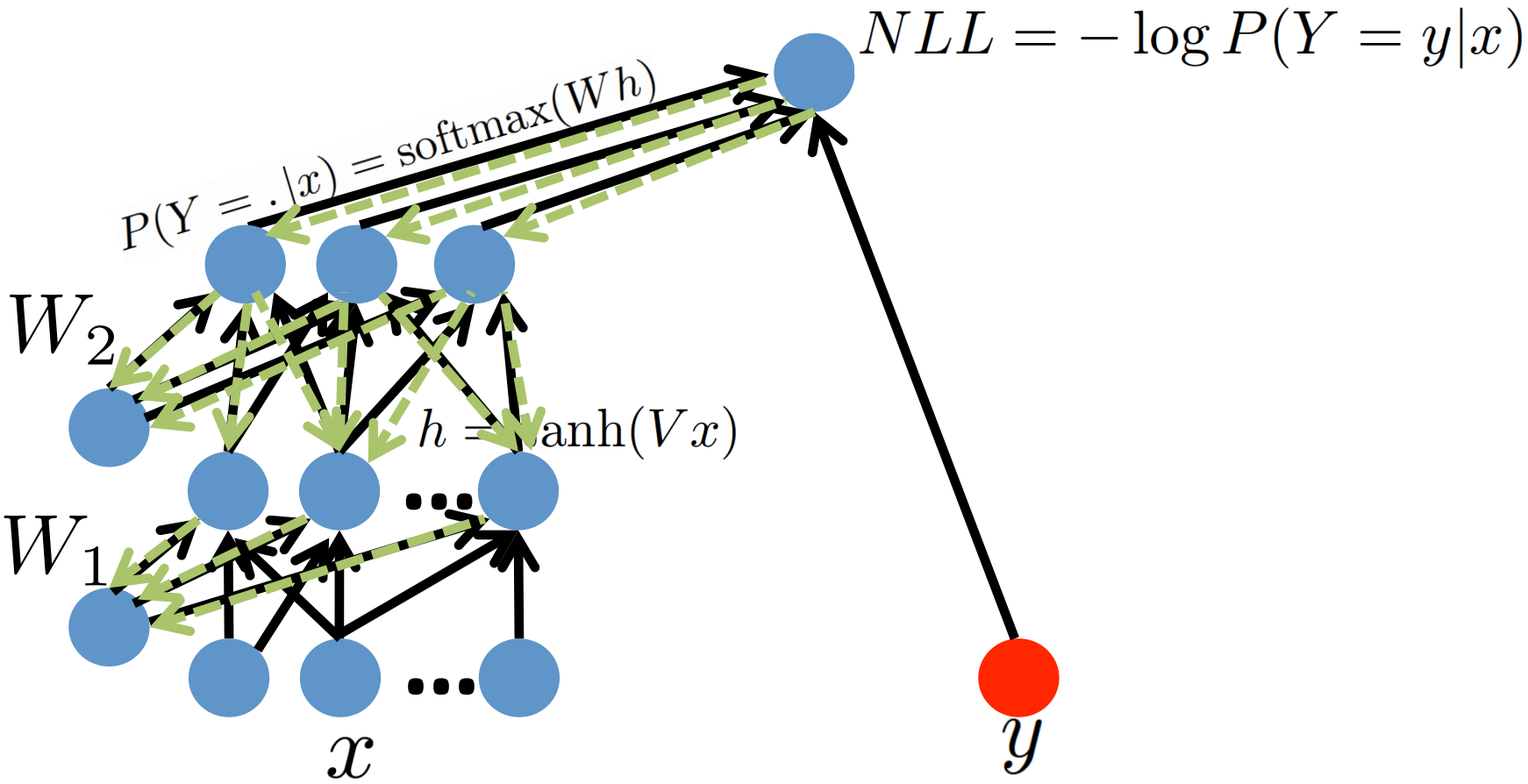


Flow graph: any directed acyclic graph
node = computation result
arc = computation dependency

$\{y_1, y_2, \ldots y_n\}$ = successors of x

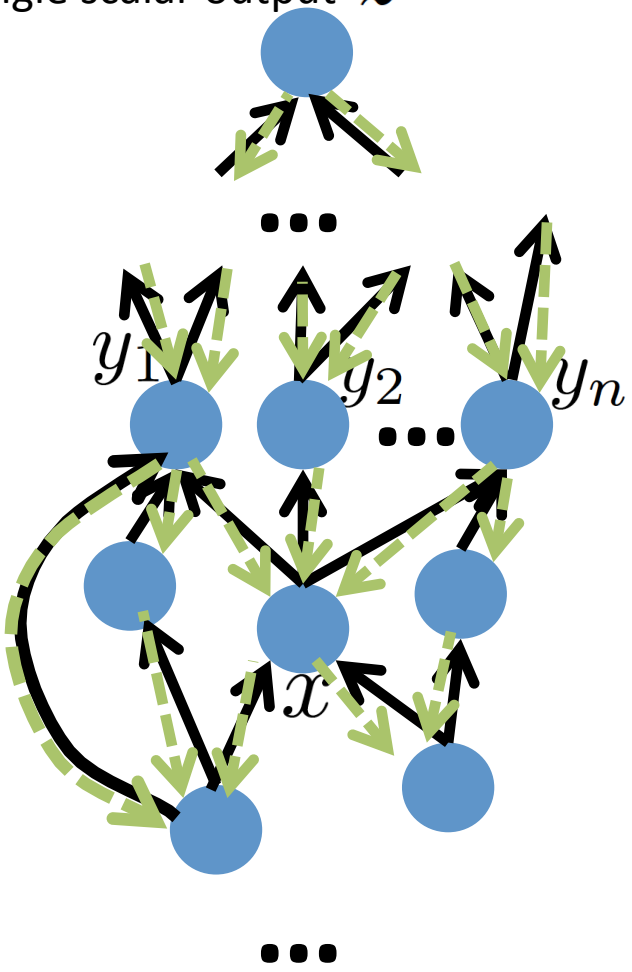
$$\frac{\partial z}{\partial x} = \sum_{i=1}^{n} \frac{\partial z}{\partial y_i} \frac{\partial y_i}{\partial x}$$

Back-Prop in Multi-Layer Net



Back-Prop in General Flow Graph

Single scalar output z



1. Fprop: visit nodes in topo-sort order
   - Compute value of node given predecessors
2. Bprop:
   - initialize output gradient = 1
   - visit nodes in reverse order:
     Compute gradient wrt each node using gradient wrt successors

$\{y_1, y_2, \ldots y_n\}$ = successors of x

$$\frac{\partial z}{\partial x} = \sum_{i=1}^{n} \frac{\partial z}{\partial y_i} \frac{\partial y_i}{\partial x}$$

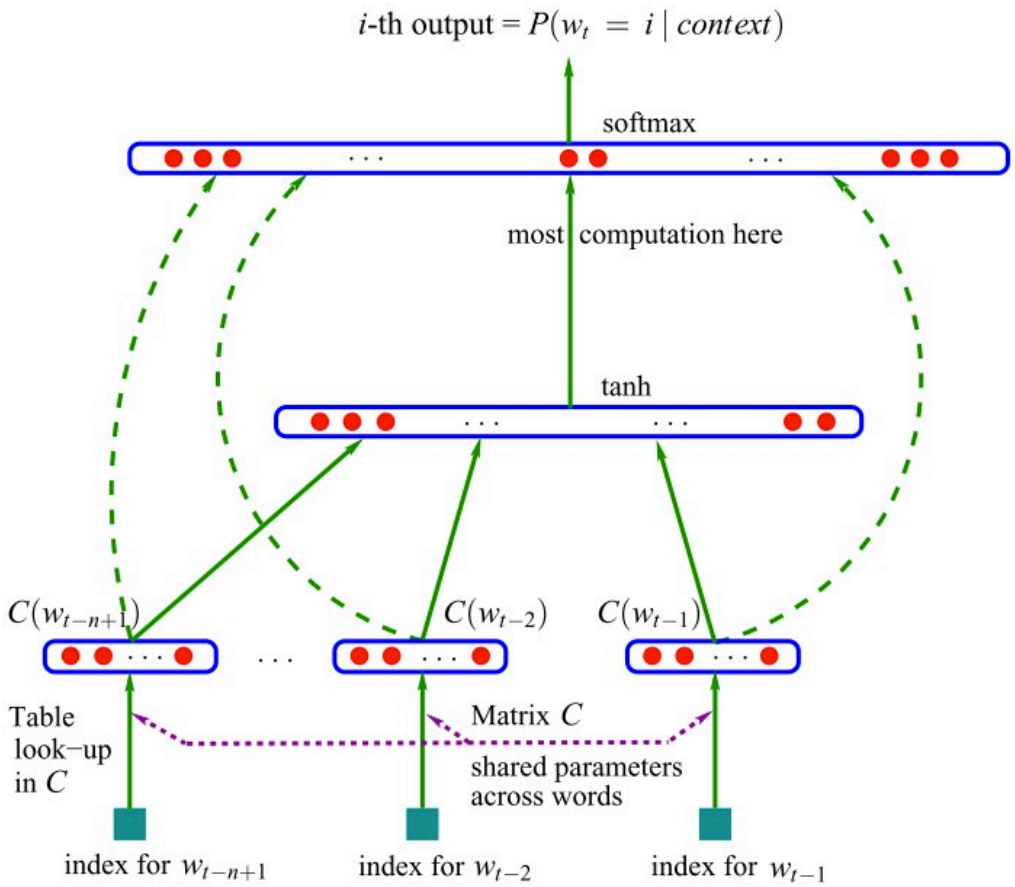
UNIVERSITEIT STELLENBOSCH UNIVERSITY

mih media lab.

1. Motivation
2. Background Theory
3. Neural Language Models
    1. Bengio et al. 2003
    2. Mnih + Hinton 2009
    3. Collobert + Weston 2011
    4. Mikolov et al. 2010
4. Other NLP Tasks
5. My Work
6. End

Neural Probabilistic Language Model

[Bengio *et al.* 2003]

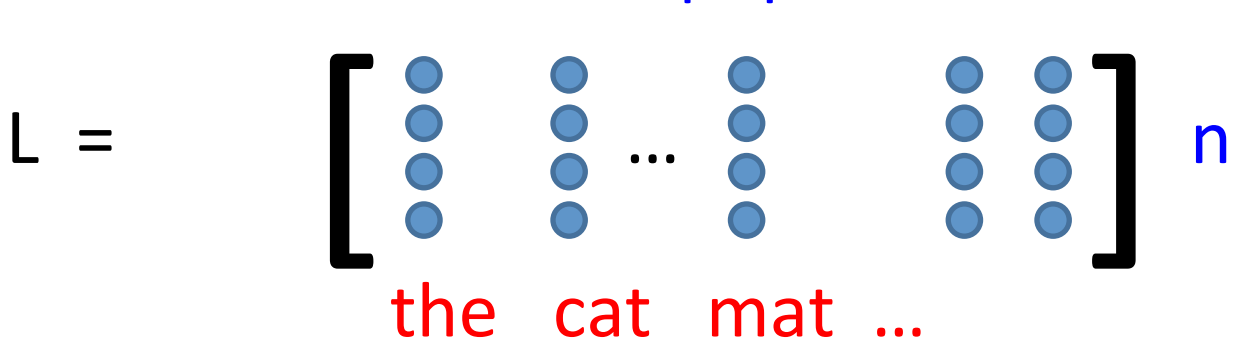


UNIVERSITEIT STELLENBOSCH UNIVERSITY

mih media lab.

Word Embedding Matrix

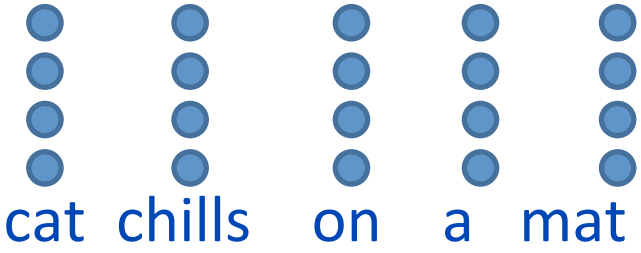
- Initialize all word vectors randomly to form a word embedding matrix $L \in \mathbb{R}^{n \times |V|}$



- These are the *word features* we want to learn
- Also called a look-up table
  - Conceptually you get a word's vector by left multiplying a one-hot vector *o* by L: $x = Lo$

Word Vectors as Input to a NN

- NPLM computes Pr(cat, chills, on, a, mat) in *softmax* layer
- To describe a phrase, retrieve (via index) the corresponding vectors from L



- Then concatenate them to (5n) vector:
- x =[●●●● ●●●● ●●●● ●●●● ●●●●]

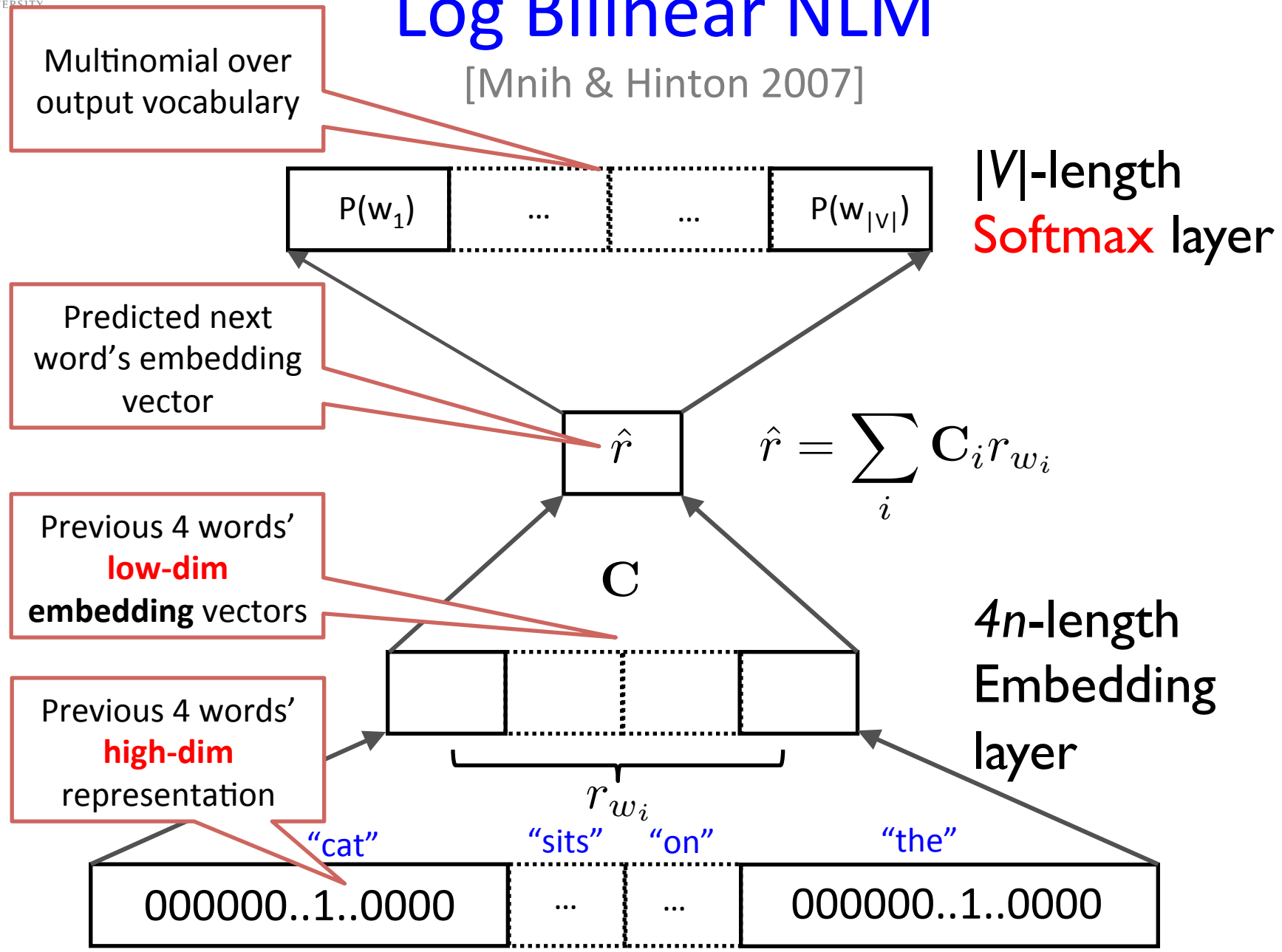
UNIVERSITEIT STELLENBOSCH UNIVERSITY

mih media lab.

1. Motivation
2. Background Theory
3. Neural Language Models
    1. Bengio et al. 2003
    2. Mnih + Hinton 2009
    3. Collobert + Weston 2011
    4. Mikolov et al. 2010
4. Other NLP Tasks
5. My Work
6. End

Log Bilinear NLM

[Mnih & Hinton 2007]



Language Modelling Output Bottleneck

- NPLM computes P(cat chills on a mat) in softmax layer

- Computing the softmax multinomial is O(kV) where k is the dimensionality of the embedding vectors (during inference & each training step)
  - Hierarchical decomposition
  - Sampling methods

UNIVERSITEIT STELLENBOSCH UNIVERSITY

mih media lab.

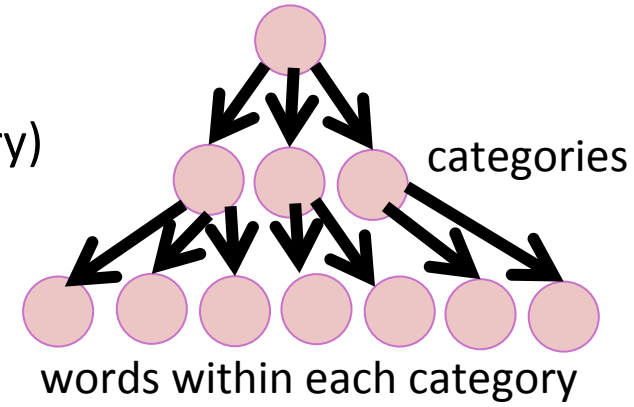
Language Modeling Output Bottleneck

- Only predict most frequent words (short list) and use n-gram for the others [Schwenk et al 2002]

- Hierarchical representations [Morin & Bengio 2005; Blitzer et al 2005; Mnih & Hinton 2007,2009; Mikolov et al 2011]:
  Multiple output groups, conditionally computed, predict
  - P(word category | context)
  - P(sub-category | context, category)
  - P(word | context, sub-category, category)

  – With 2 levels, O(|V|) becomes O(sqrt(|V|))
  – With d levels, O(|V|) becomes O(log(|V|))

  – Hard categories, can be arbitrary [Mikolov et al 2011]



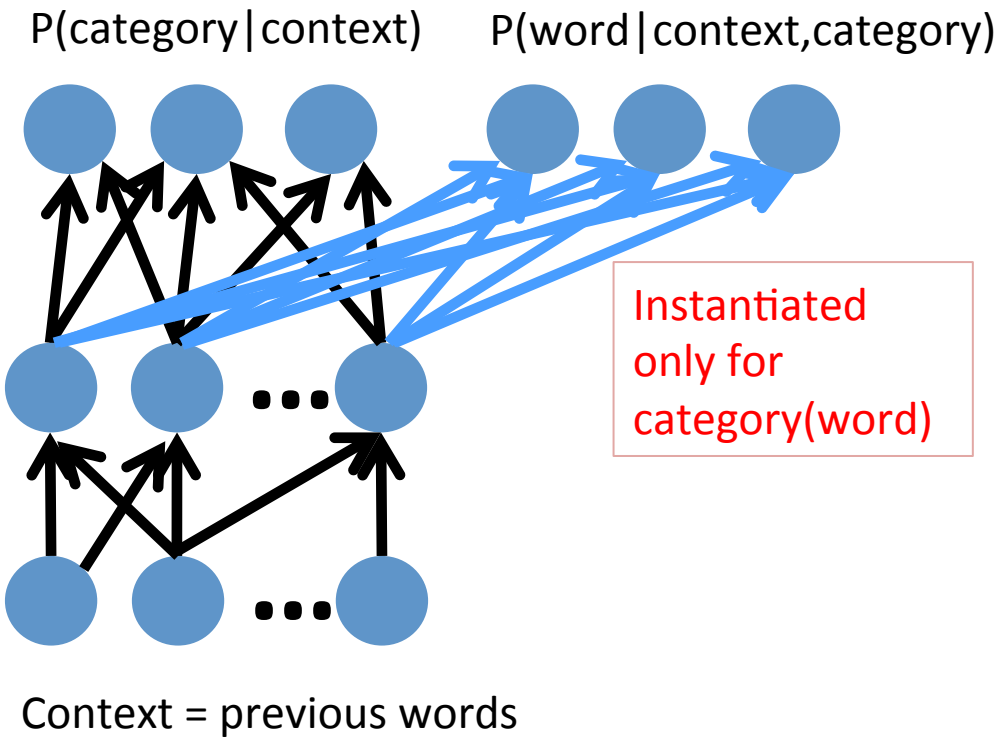
Language Modelling Output Bottleneck:
Hierarchical Word Categories

Compute:

P(word|category,context)

only for

category=category(word)



UNIVERSITEIT STELLENBOSCH UNIVERSITY

mih media lab.

Language Modelling Output Bottleneck: Sampling Methods

- Importance sampling to recover next-word probabilities [Bengio & Senecal 2003, 2008]
- Sampling a ranking loss [Collobert et al, 2008, 2011]
  - Increase score of observed word's output
  - Decrease score of randomly selected word (negative example)
  - Not anymore outputting probabilities (OK if the goal is just to learn word embeddings)
- Importance sampling for reconstructing bag-of-words [Dauphin et al 2011]
- MCMC-based sampling strategy for RBM language model [Dahl et al. 2012]

UNIVERSITEIT STELLENBOSCH UNIVERSITY

mih media lab.

1. Motivation
2. Background Theory
3. Neural Language Models
    1. Bengio et al. 2003
    2. Mnih + Hinton 2009
    3. Collobert + Weston 2011
    4. Mikolov et al. 2010
4. Other NLP Tasks
5. My Work
6. End

UNIVERSITEIT STELLENBOSCH UNIVERSITY

mih media lab.

A NN for learning word vectors

- The output bottleneck makes training SLOW
- Idea: A word and its context is a positive training sample, a random word in that same context is a negative training sample.
  - score(cat chills on a mat) > score(cat chills Jeju a mat)
- How to compute the score?
  - With a neural network
  - Each word is associated with an *n*-dimensional vector

A NN for learning word vectors

[Collobert & Weston, JMLR, 2011]

Idea: A word and its context is a positive training sample; a random word in that same context gives a negative training sample:

+ cat chills on a mat

− cat chills Jeju a mat

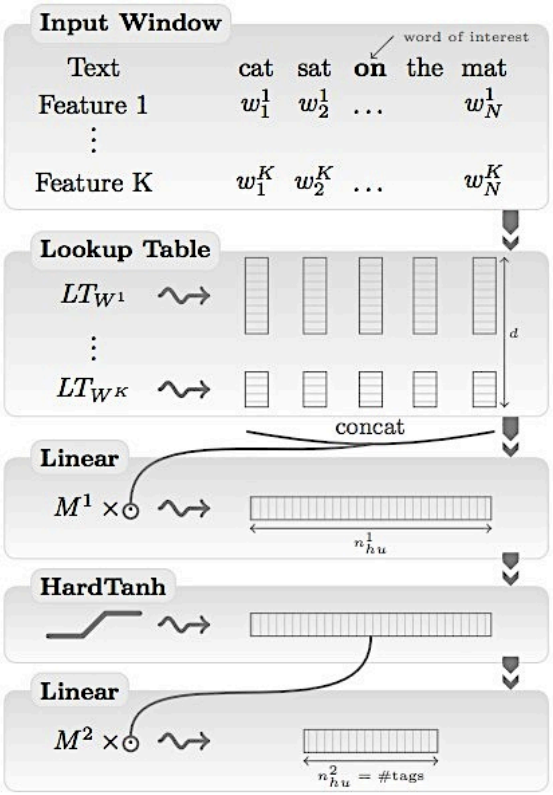
Similar: Implicit negative evidence in Contrastive Estimation, Smith and Eisner (2005)

UNIVERSITEIT STELLENBOSCH UNIVERSITY

mih media lab.

Collobert and Weston

[Collobert & Weston, JMLR 2011]



"Window" approach

UNIVERSITEIT STELLENBOSCH UNIVERSITY

mih media lab.

A Single Layer of the Neural Network

- A single layer is a combination of a linear layer and a nonlinearity:

$$z = Wx + b$$
$$a = f(z)$$

- The neural activations can then be used to compute some function.
- For instance, to score input context x:

$$score(x) = W_{score}^T a \in \mathbb{R}$$

Computing a Window's Score

- Computing a window's score with a 3-layer Neural Net: s = score(cat chills on a mat)

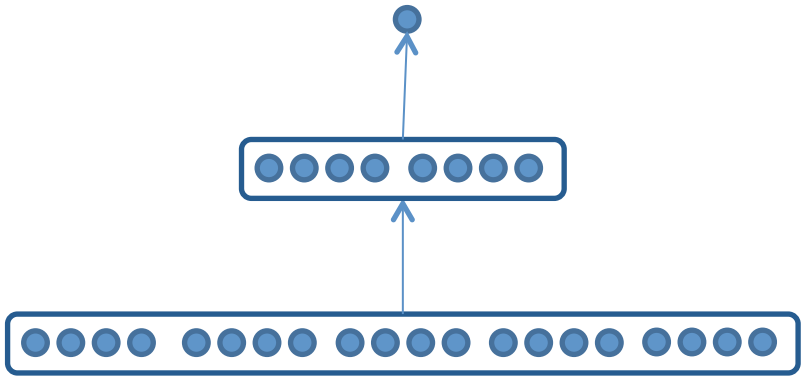
$$s = W_{score}^T f(Wx + b) \qquad x \in \mathbb{R}^{20\times 1}, W \in \mathbb{R}^{8\times 20}, W_{score} \in \mathbb{R}^{8\times 1}$$

$$s = W_{score}^T a$$

$$\begin{aligned} z &= Wx + b \\ a &= f(z) \end{aligned}$$

$$x = [x_{cat} \quad x_{chills} \quad x_{on} \quad x_a \quad x_{mat}]$$

$$L \in \mathbb{R}^{n\times|V|}$$



cat chills on a mat

UNIVERSITEIT STELLENBOSCH UNIVERSITY

mih media lab.

Computing a Window's Score

- s = score(cat chills on a mat)
- s_c = score(cat chills Jeju a mat)

- Idea for training objective *J*: make score of true window larger and corrupt window's score lower (until they're good enough):
- I.e. minimize a hinge loss

$$J = \max(0, 1 - s + s_c)$$

UNIVERSITEIT STELLENBOSCH UNIVERSITY

mih media lab.

1. Motivation
2. Background Theory
3. Neural Language Models
    1. Bengio et al. 2003
    2. Mnih + Hinton 2009
    3. Collobert + Weston 2011
    4. Mikolov et al. 2010
4. Other NLP Tasks
5. My Work
6. End

UNIVERSITEIT STELLENBOSCH UNIVERSITY

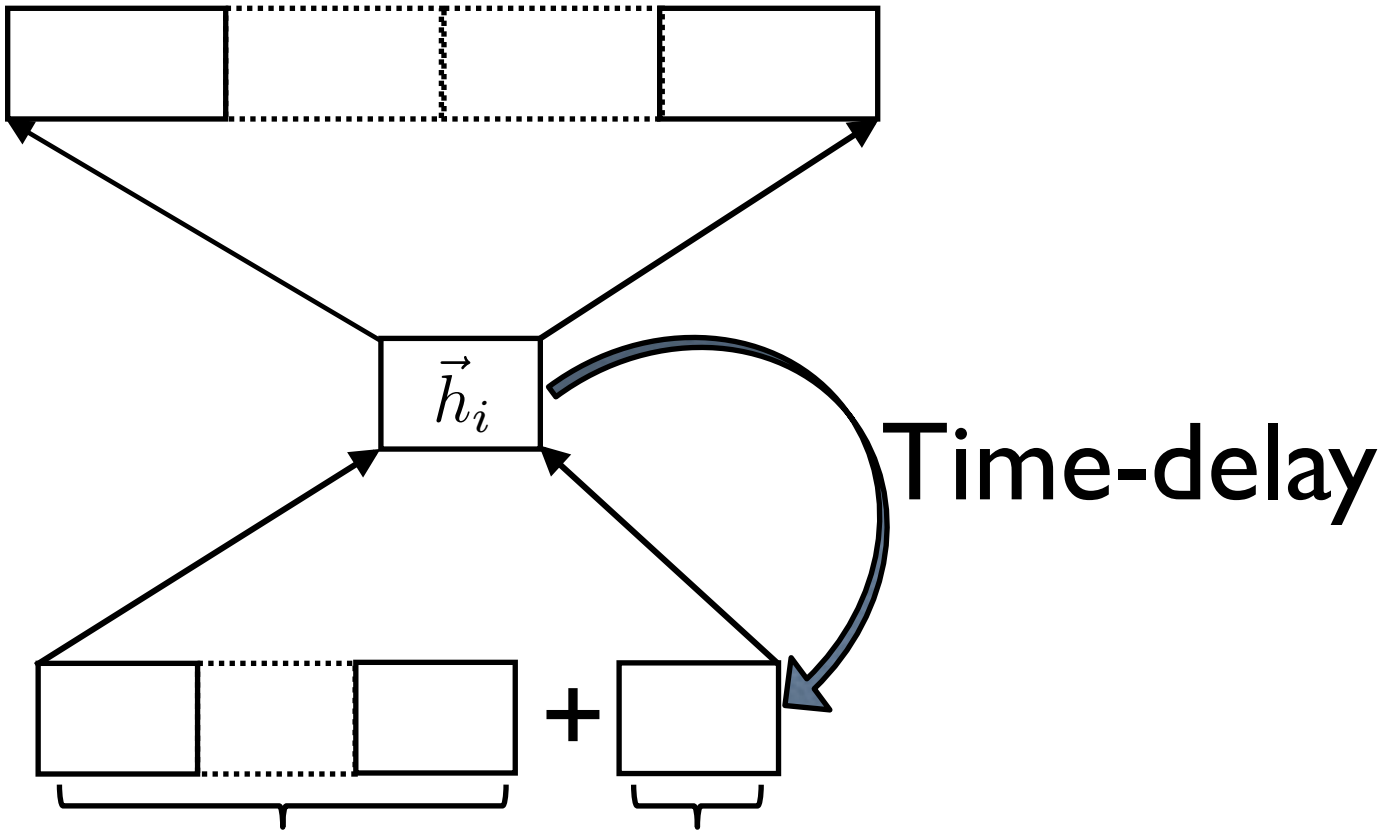
mih media lab.

The Exponential Blow-up of N-gram LMs

- Context size is important for accurately predicting next word
- Precisely why trigram LM > bigram LM, etc.
- However, the parameters (conditional probability tables) of an n-gram LM grows as V^c for a V-length vocabulary and c-length context..

Recurrent NNs

[Mikolov *et al.* 2010; Sutskever *et al.* 2011]



UNIVERSITEIT STELLENBOSCH UNIVERSITY

mih media lab.

RNN Language modelling results

[Mikolov *et al.* 2010]

	Penn Corpus		Switchboard	
Model	NN	NN+KN	NN	NN+KN
KN5 (baseline)	-	141	-	92.9
feedforward NN	141	118	85.1	77.5
RNN trained by BP	137	113	81.3	75.4
RNN trained by BPTT	123	106	77.5	72.5

UNIVERSITEIT STELLENBOSCH UNIVERSITY

mih media lab.

1. Motivation
2. Background Theory
3. Neural Language Models
4. Other NLP Tasks
    1. POS, Chunking, NER, SRL
    2. Sentiment Analysis
    3. Paraphrase Detection
    4. Semantic Parsing
5. My Work
6. End

UNIVERSITEIT STELLENBOSCH UNIVERSITY

mih media lab.

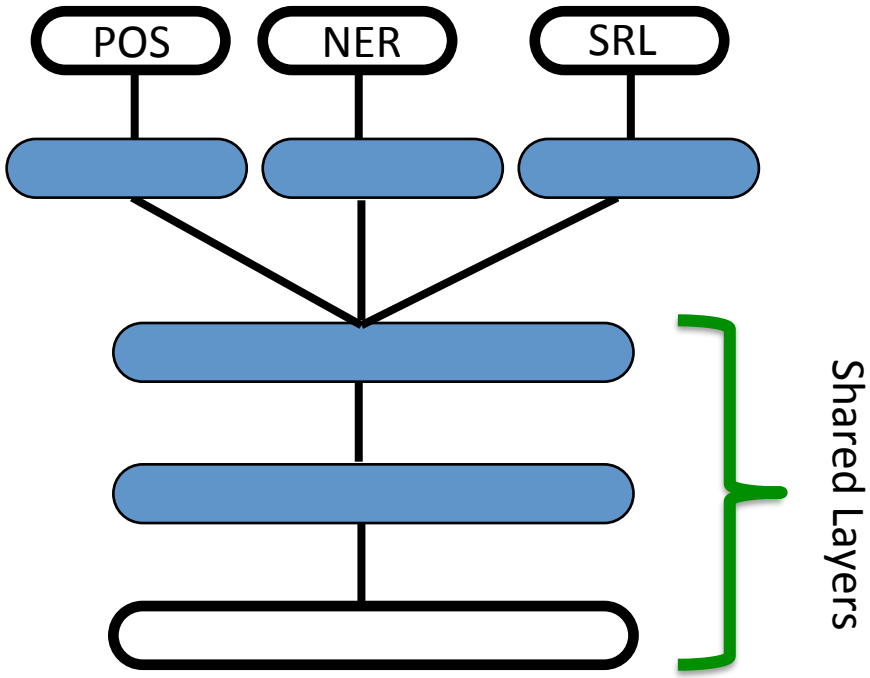
1. Motivation
2. Background Theory
3. Neural Language Models
4. Other NLP Tasks
    1. POS, Chunking, NER, SRL
    2. Sentiment Analysis
    3. Paraphrase Detection
    4. Semantic Parsing
5. My Work
6. End

UNIVERSITEIT STELLENBOSCH UNIVERSITY

mih media lab.

Multi-Task Learning

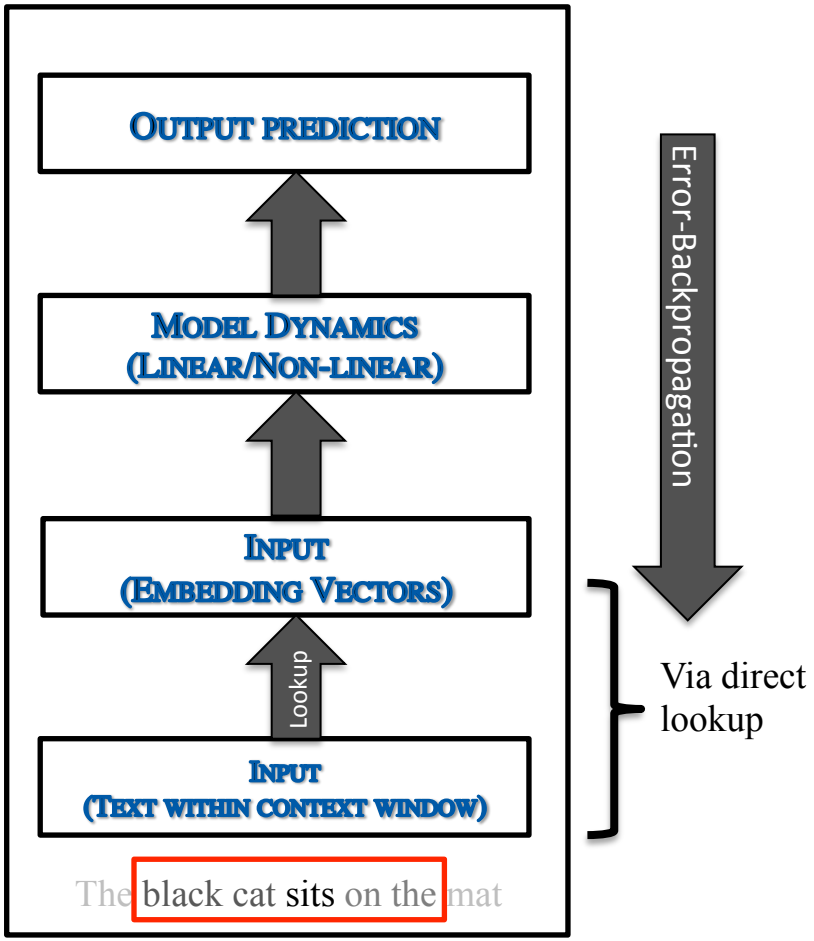
- Deep architectures learn good intermediate representations that can be shared across tasks
- Good representations make sense for many tasks



UNIVERSITEIT STELLENBOSCH UNIVERSITY

mih media lab.

Squinty-eyed picture so far..



UNIVERSITEIT STELLENBOSCH UNIVERSITY

mih media lab.

The semantic boomerang effect..

"[…] [the window approach] works well for most NLP tasks we're interested in. However this approach fails with SRL where the tag of a word depends on a verb […]. If the verb falls outside the window one cannot expect this word to be tagged correctly."

[Collobert & Weston, 2011]

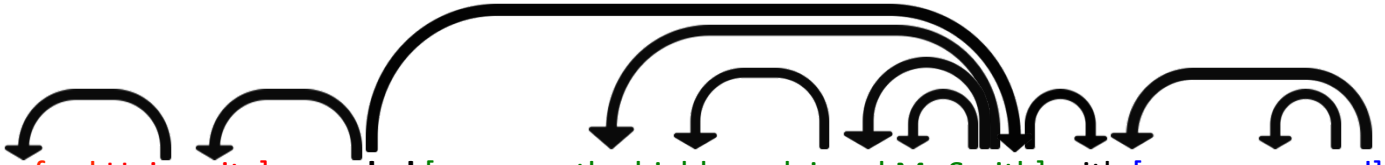
UNIVERSITEIT STELLENBOSCH UNIVERSITY

mih media lab.

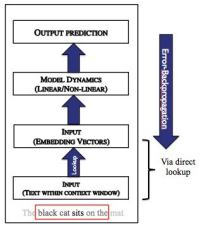
The semantic boomerang effect..

Stanford University awarded the highly acclaimed Mr Smith with an award…

The semantic boomerang effect..



???

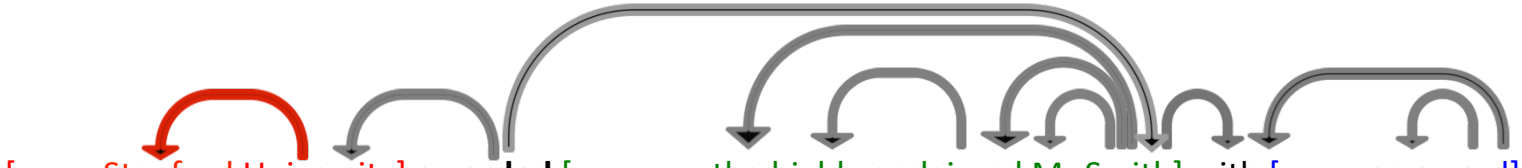


	Stanford	University awarded the highly acclaimed Mr Smith with an award …

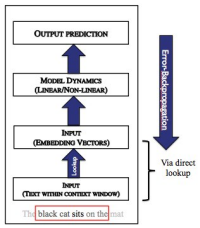
UNIVERSITEIT STELLENBOSCH UNIVERSITY

mih media lab.

The semantic boomerang effect..



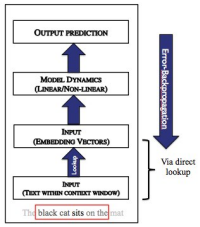
???



	Stanford	University awarded the highly acclaimed Mr Smith with an award …

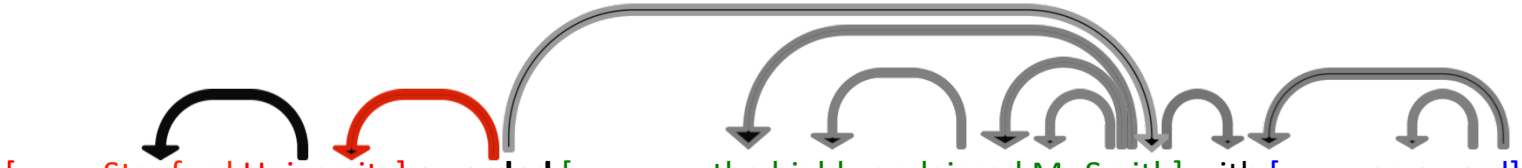
The semantic boomerang effect..

B-AGENT

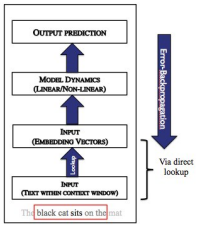


	Stanford	University awarded the highly acclaimed Mr Smith with an award …

The semantic boomerang effect..

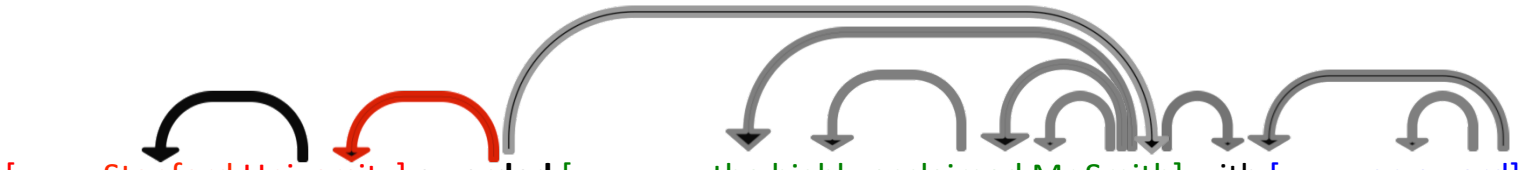


B-AGENT ????

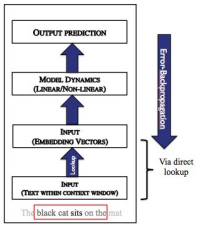


	Stanford University	awarded the highly acclaimed Mr Smith with an award …

The semantic boomerang effect..

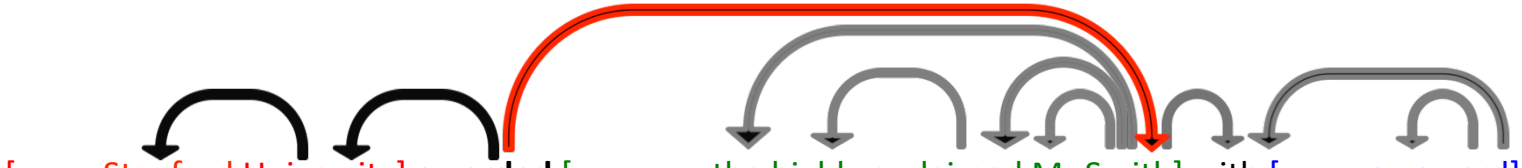


B-AGENT I-AGENT



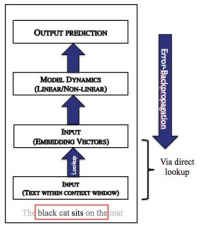
	Stanford University	awarded the highly acclaimed Mr Smith with an award …

The semantic boomerang effect..



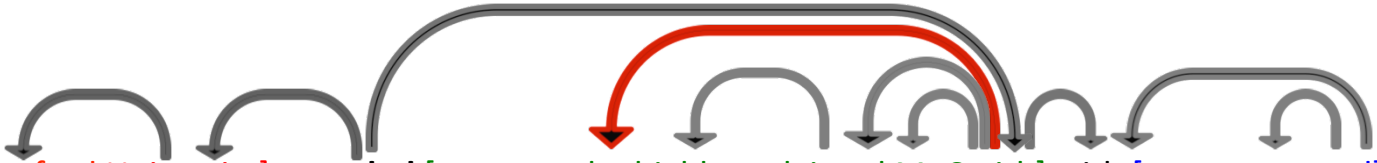
[AGENT Stanford University] **awarded** [RECIPIENT the highly acclaimed Mr Smith] with [THEME an award] …

B-AGENT I-AGENT B-PRED



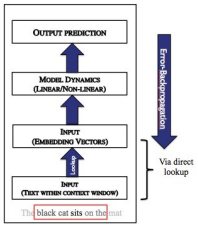
	Stanford University awarded	the highly acclaimed Mr Smith with an award …

The semantic boomerang effect..



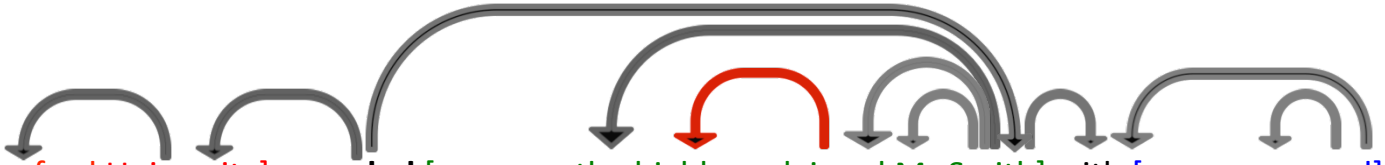
[AGENT Stanford University] **awarded** [RECIPIENT the highly acclaimed Mr Smith] with [THEME an award] …

B-AGENT I-AGENT B-PRED B-REC



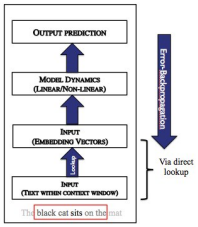
	Stanford University awarded the	highly acclaimed Mr Smith with an award …

The semantic boomerang effect..



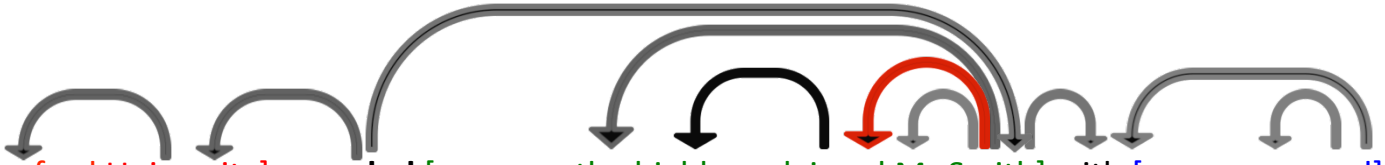
[AGENT Stanford University] **awarded** [RECIPIENT the highly acclaimed Mr Smith] with [THEME an award] …

B-AGENT I-AGENT B-PRED B-REC I-REC



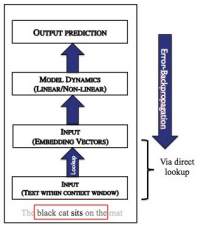
	Stanford University awarded the highly	acclaimed Mr Smith with an award …

The semantic boomerang effect..



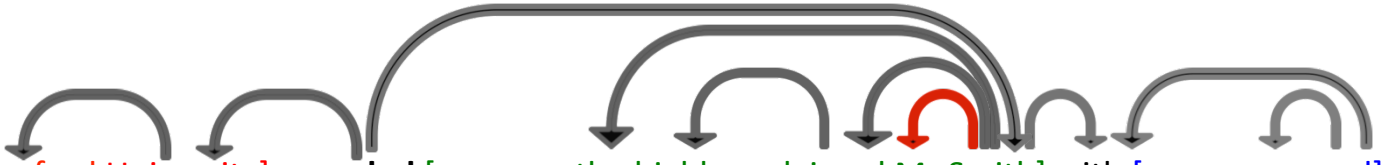
[AGENT Stanford University] **awarded** [RECIPIENT the highly acclaimed Mr Smith] with [THEME an award] …

B-AGENT I-AGENT B-PRED B-REC I-REC I-REC



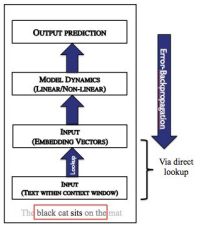
	University awarded the highly acclaimed	Mr Smith with an award …

The semantic boomerang effect..



[AGENT Stanford University] **awarded** [RECIPIENT the highly acclaimed Mr Smith] with [THEME an award] …

B-AGENT I-AGENT B-PRED B-REC I-REC I-REC I-REC



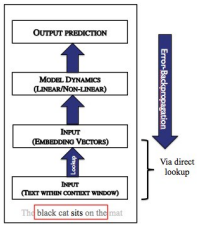
	awarded the highly acclaimed Mr	Smith with an award …

The semantic boomerang effect..

[AGENT Stanford University] **awarded** [RECIPIENT the highly acclaimed Mr Smith] with [THEME an award] …

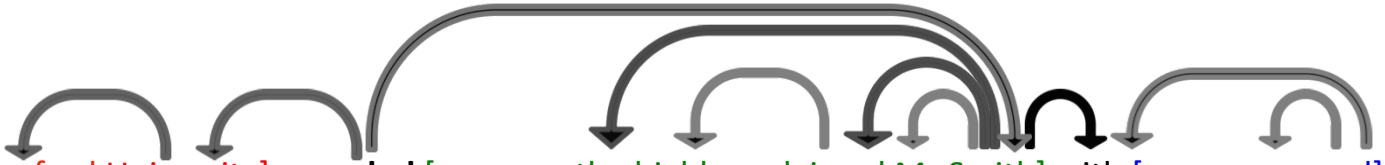
Main predicate "awarded" outside model context-window

B-AGENT I-AGENT B-PRED B-REC I-REC I-REC I-REC I-REC



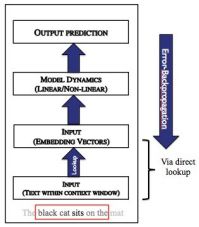
	the highly acclaimed Mr Smith	with an award …

The semantic boomerang effect..



[AGENT Stanford University] **awarded** [RECIPIENT the highly acclaimed Mr Smith] with [THEME an award] …

B-AGENT I-AGENT B-PRED B-REC I-REC I-REC I-REC I-REC O



highly acclaimed Mr Smith with an award …

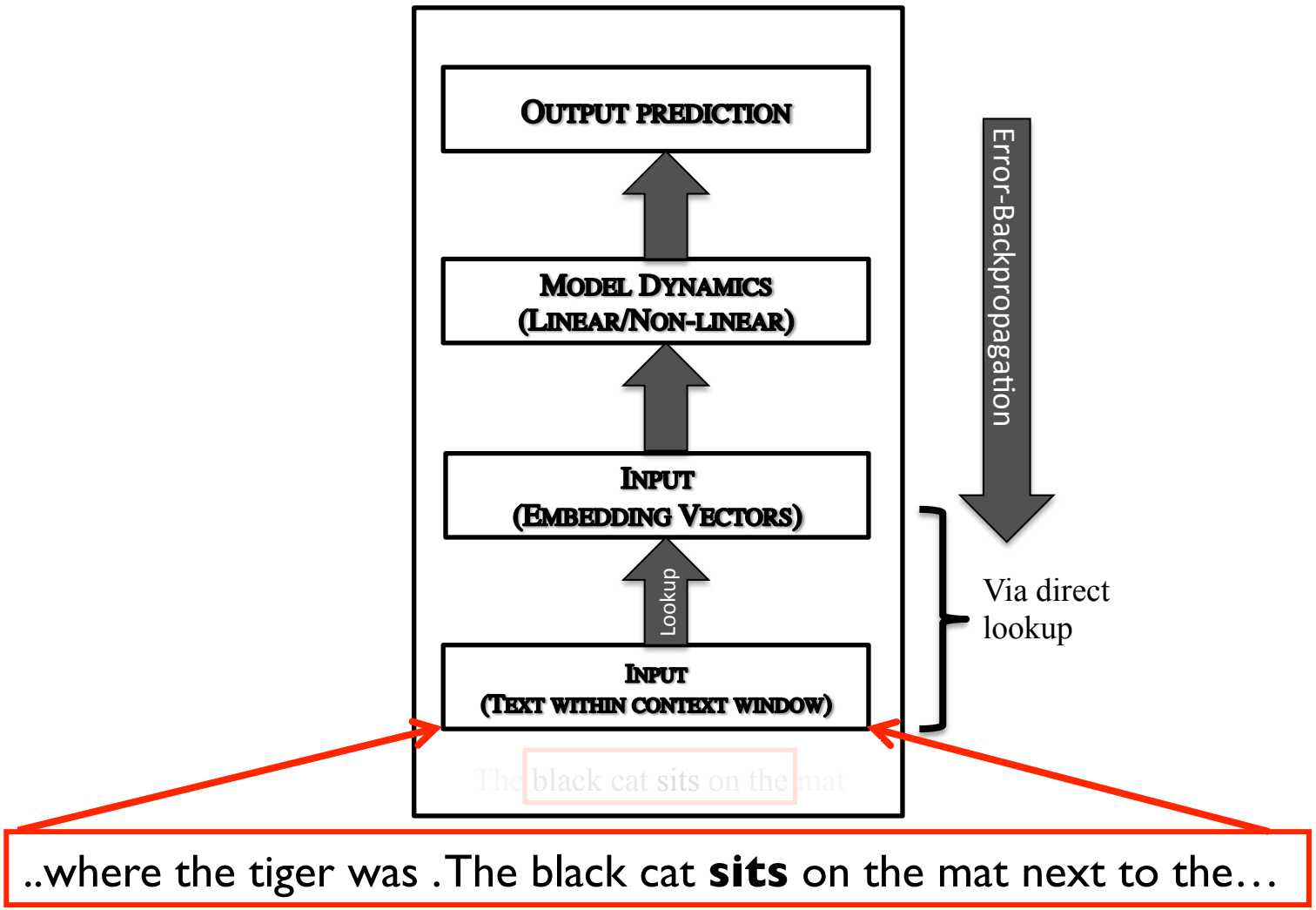
UNIVERSITEIT STELLENBOSCH UNIVERSITY

mih media lab.

Some issues..

- Window-based models cannot capture long-range dependencies necessary for semantic disambiguation
- Models essentially learn *local lexical distributional semantics*
- Models not informed by our knowledge of language
- We like what they're offering, but.. we want more

Increase context!



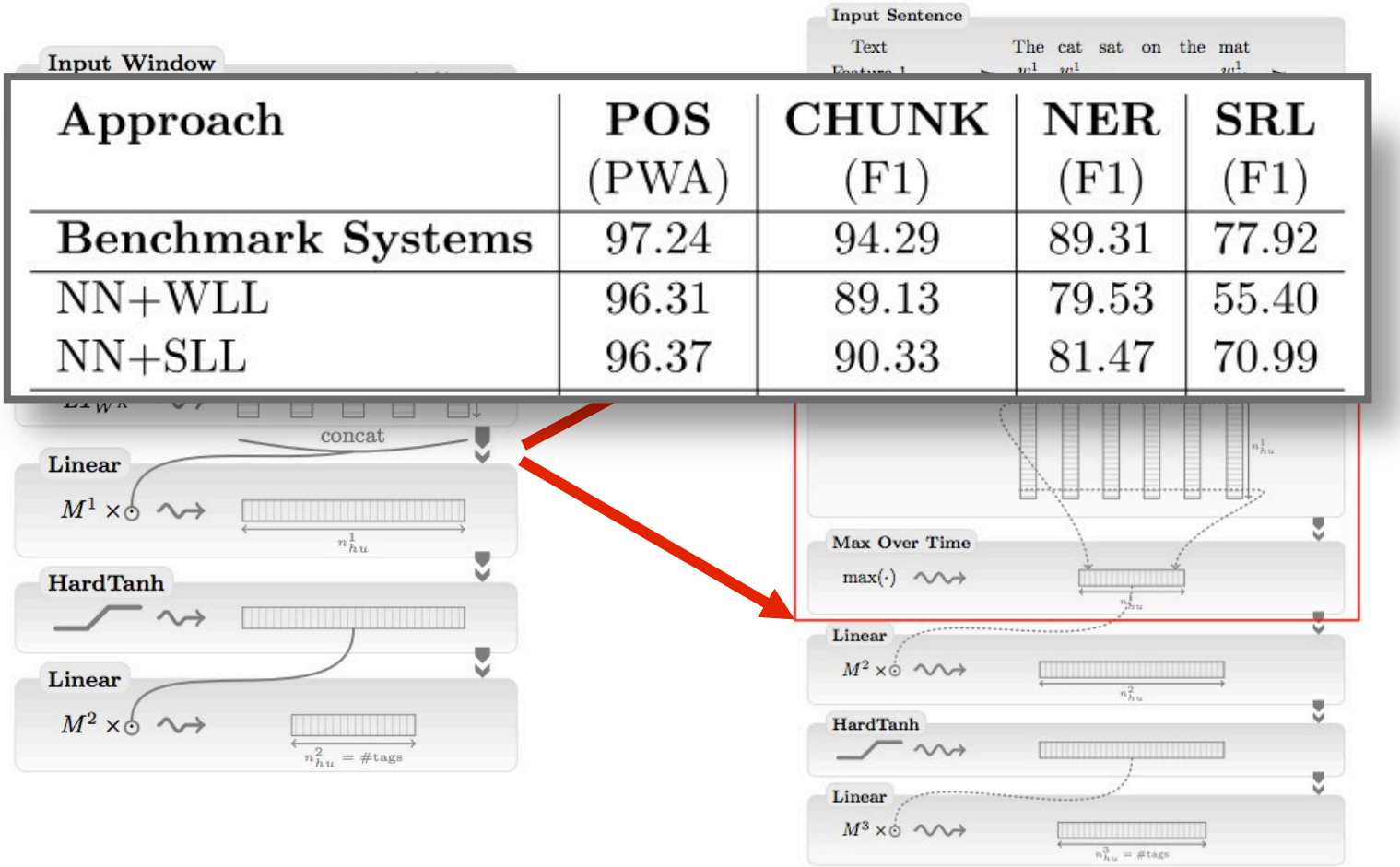
UNIVERSITEIT STELLENBOSCH UNIVERSITY

mih media lab.

Collobert and Weston

[Collobert & Weston 2011]

Approach	POS (PWA)	CHUNK (F1)	NER (F1)	SRL (F1)
Benchmark Systems	97.24	94.29	89.31	77.92
NN+WLL	96.31	89.13	79.53	55.40
NN+SLL	96.37	90.33	81.47	70.99

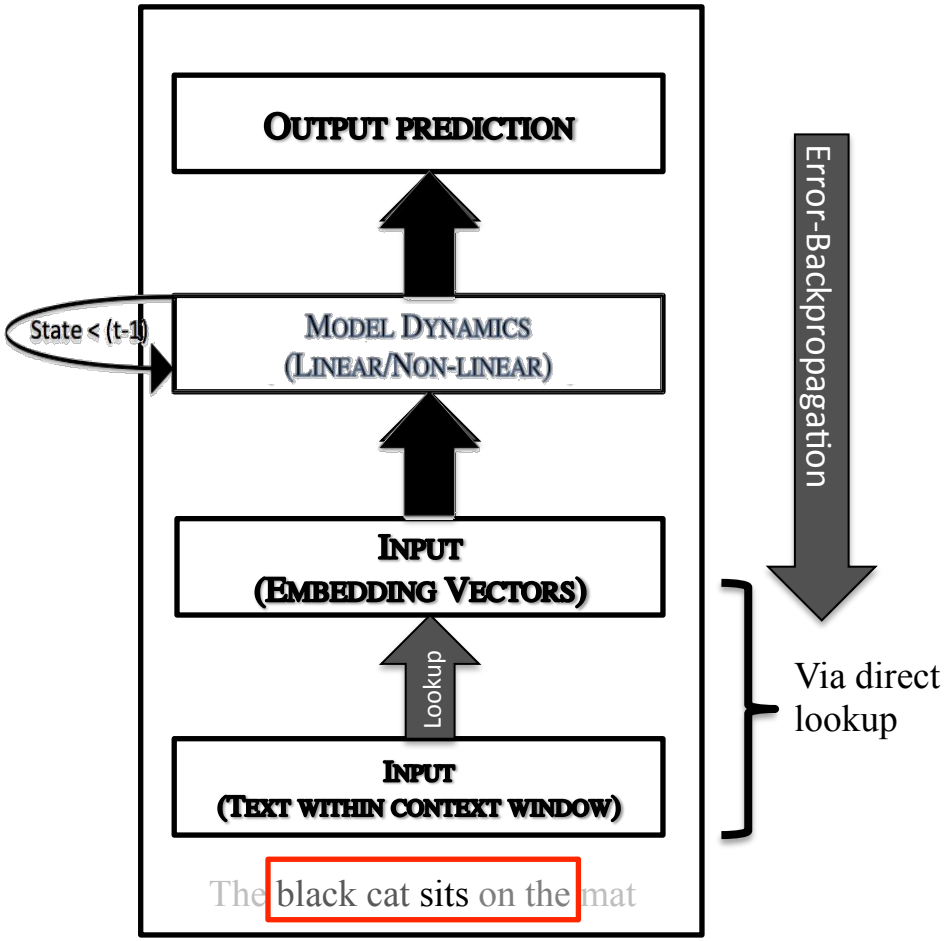


Window-approach → **Sentence-approach**

UNIVERSITEIT STELLENBOSCH UNIVERSITY

mih media lab.

Increase context!!

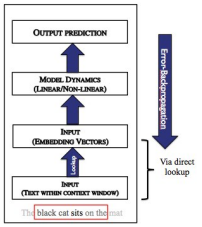


The semantic boomerang effect..

[AGENT Stanford University] **awarded** [RECIPIENT the highly acclaimed Mr Smith] with [THEME an award] …

Main predicate "awarded" outside model context-window

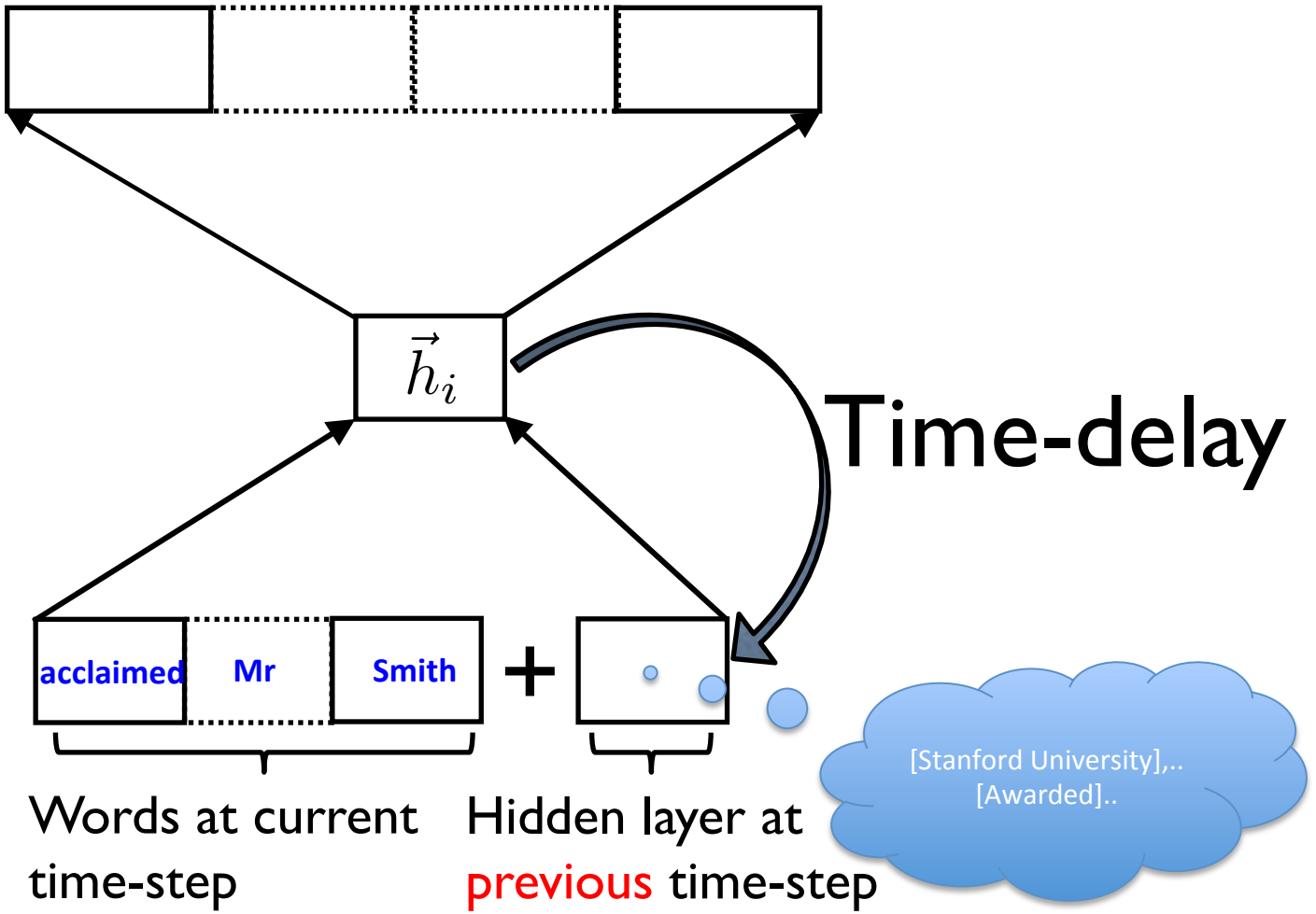
B-AGENT I-AGENT B-PRED B-REC I-REC I-REC I-REC I-REC



	the highly acclaimed Mr Smith	with an award …

Recurrent NNs

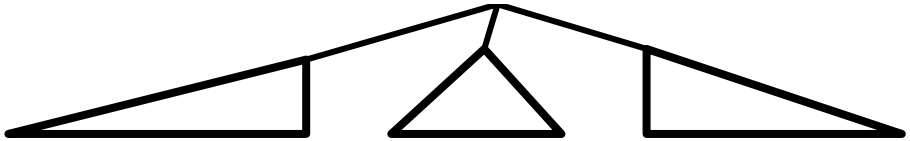
[Mikolov *et al.* 2010; Sutskever *et al.* 2011]



UNIVERSITEIT STELLENBOSCH UNIVERSITY

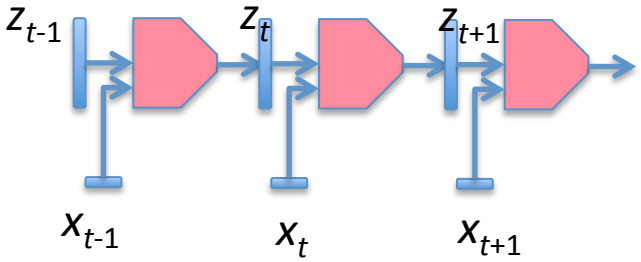
mih media lab.

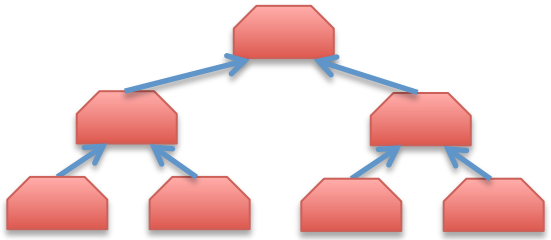
1. Motivation
2. Background Theory
3. Neural Language Models
4. Other NLP Tasks
    1. POS, Chunking, NER, SRL
    2. Sentiment Analysis
    3. Paraphrase Detection
    4. Semantic Parsing
5. My Work
6. End



(Compo(sitio(nali ty)))?

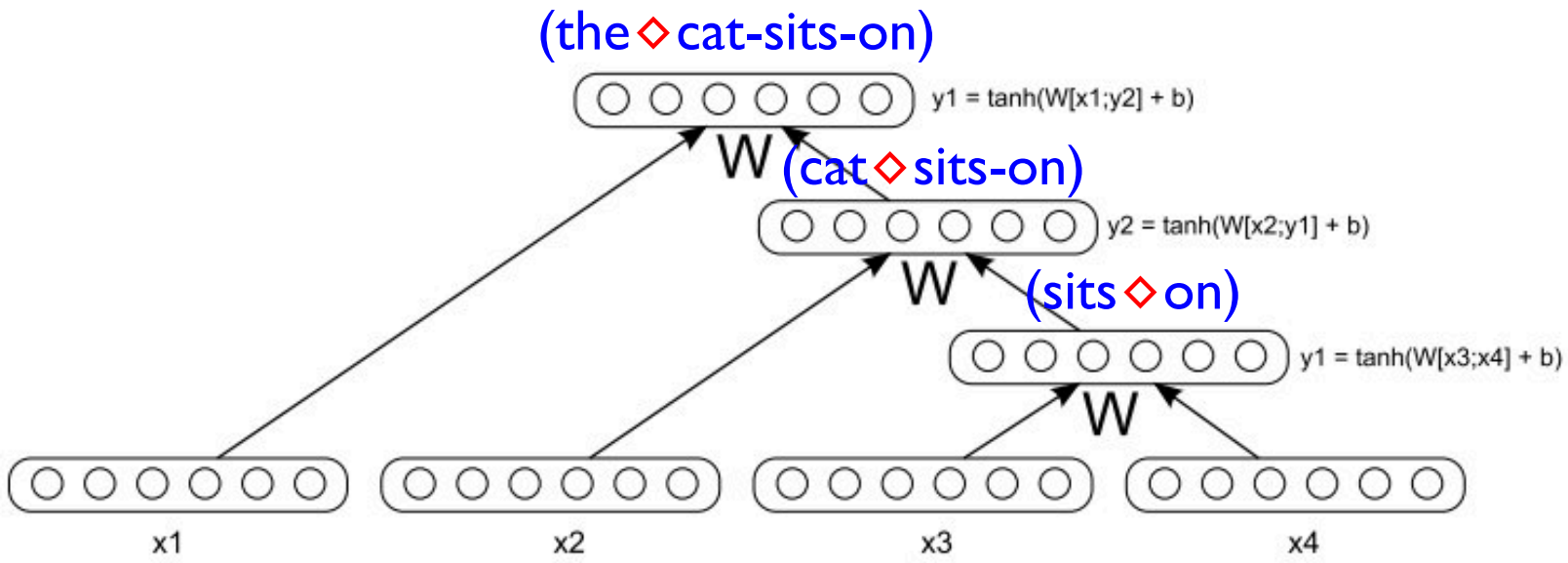
- Natural language is compositional
- E.g. head nouns and pronominal modifiers:
  "My/PRP bank/**NN**" vs "I/PRP bank/**VBP**"
- Differentiates the *sense meanings* of words
- Recurrent nets can model sequential (chain-structure) compositionality
- Recursive nets generalize this to arbitrary DAGs





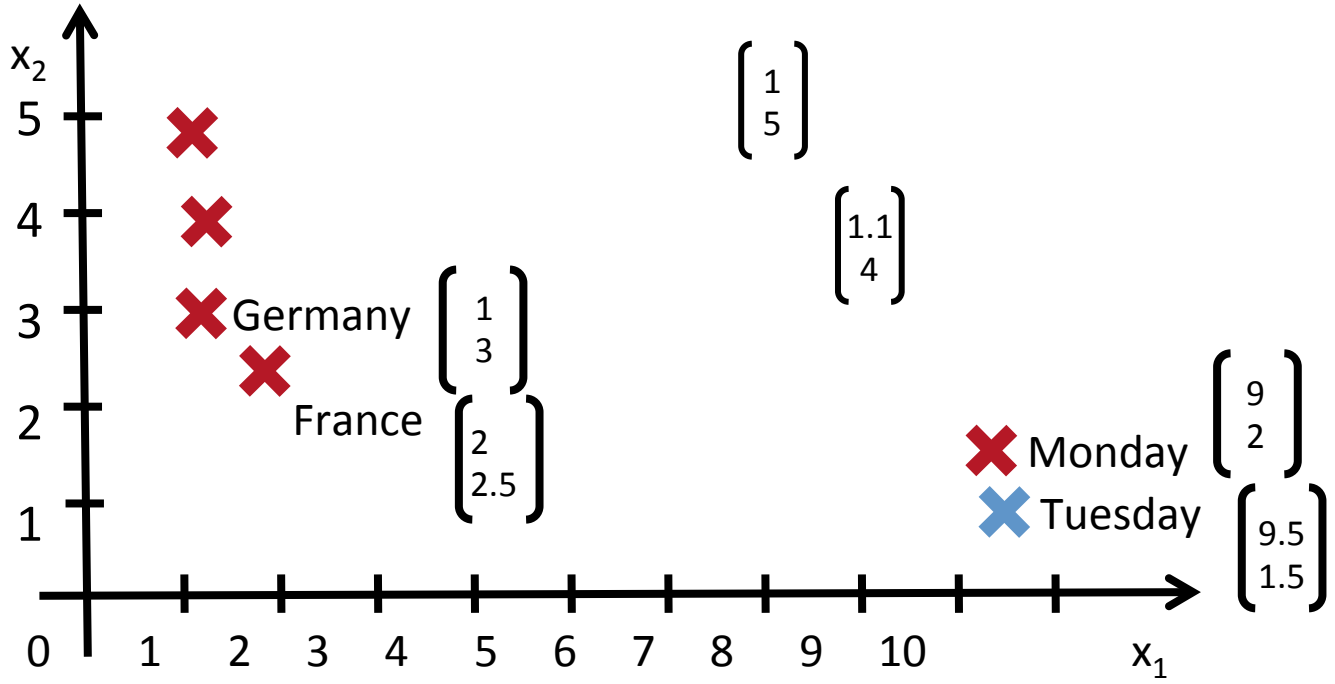
Recursive NNs

[Pollack 1996; Costa et al. 2001; Socher et al. 2010,2011]



the cat sits on …

Building on Word Vector Space Models



the country of my birth
the place where I was born

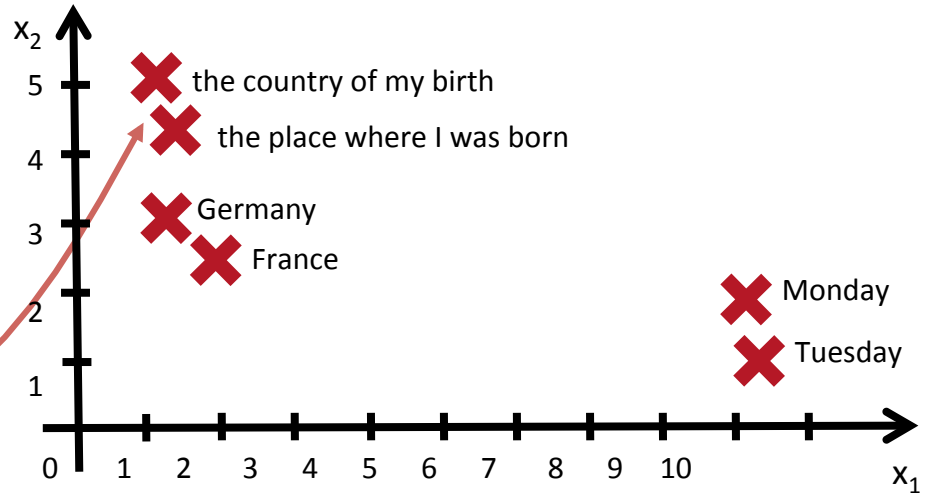
But how can we represent the meaning of longer phrases?

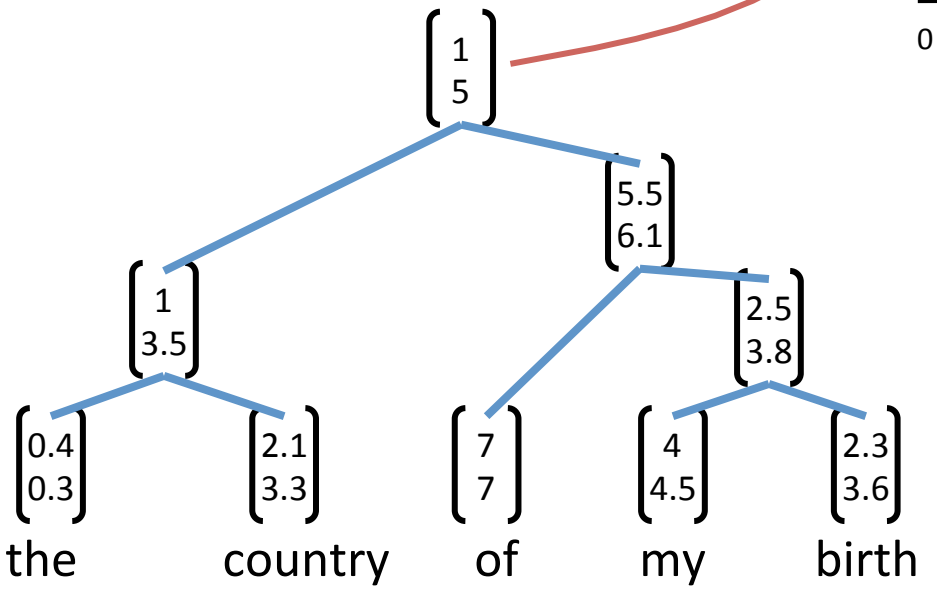
By mapping them into the same vector space!

How should we map phrases into a vector space?

Use 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.





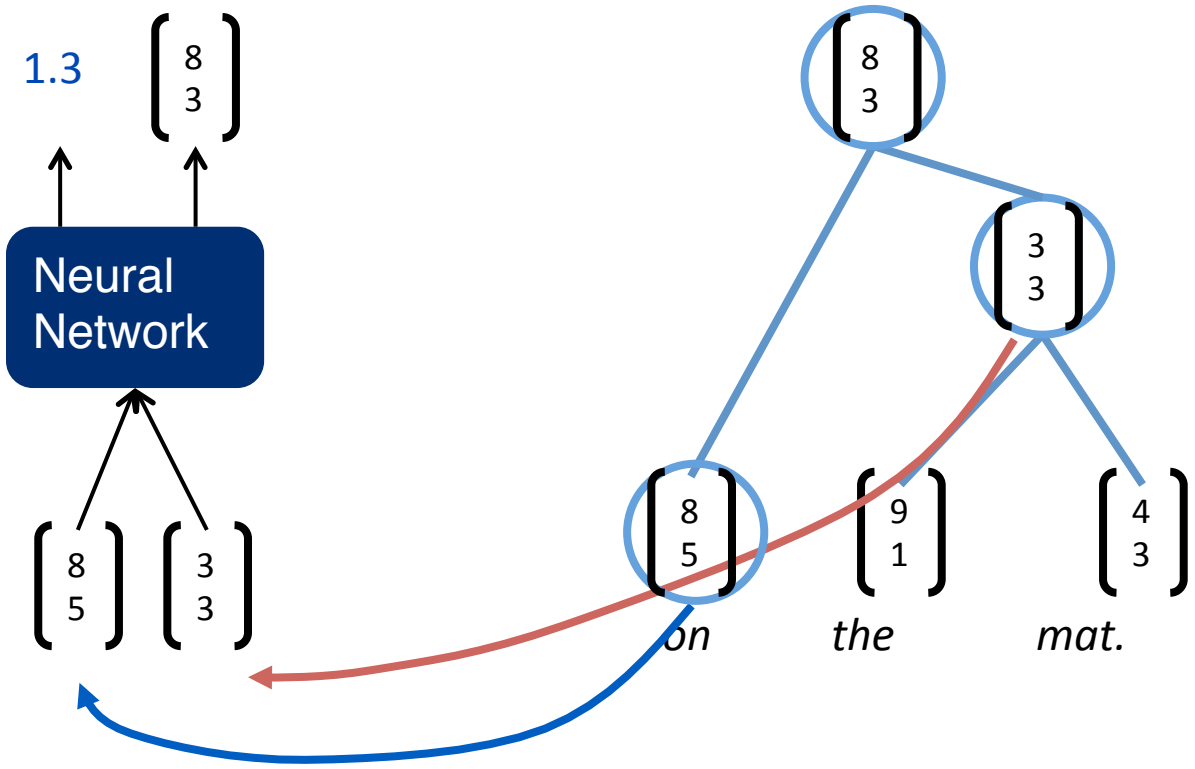
Recursive Neural Nets can jointly learn compositional vector representations and parse trees

Recursive Neural Networks for Structure Prediction

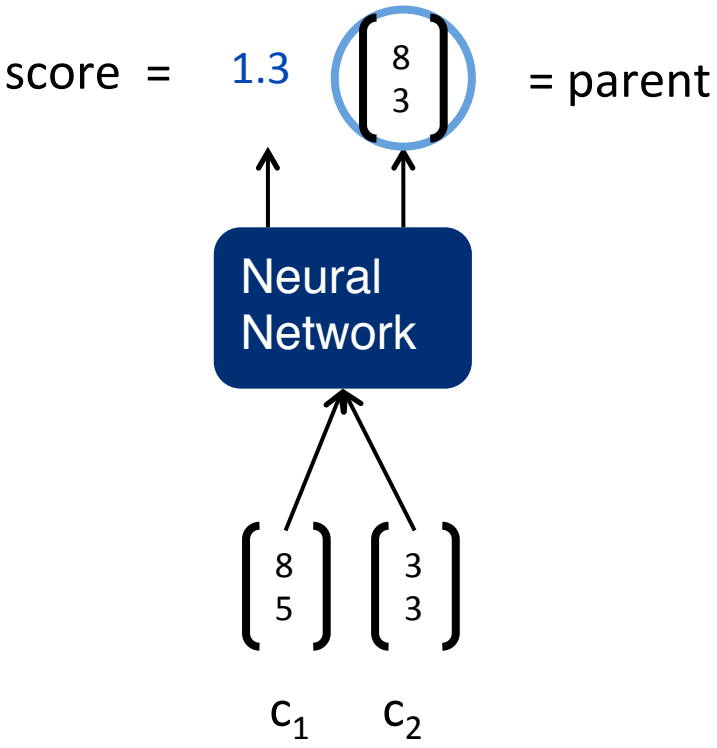
Inputs: two candidate children's representations

Outputs:

1. The semantic representation if the two nodes are merged.
2. Score of how plausible the new node would be.



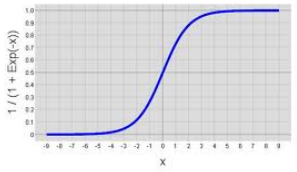
Recursive Neural Network Definition



$$score = W^T_{score}\, p$$

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

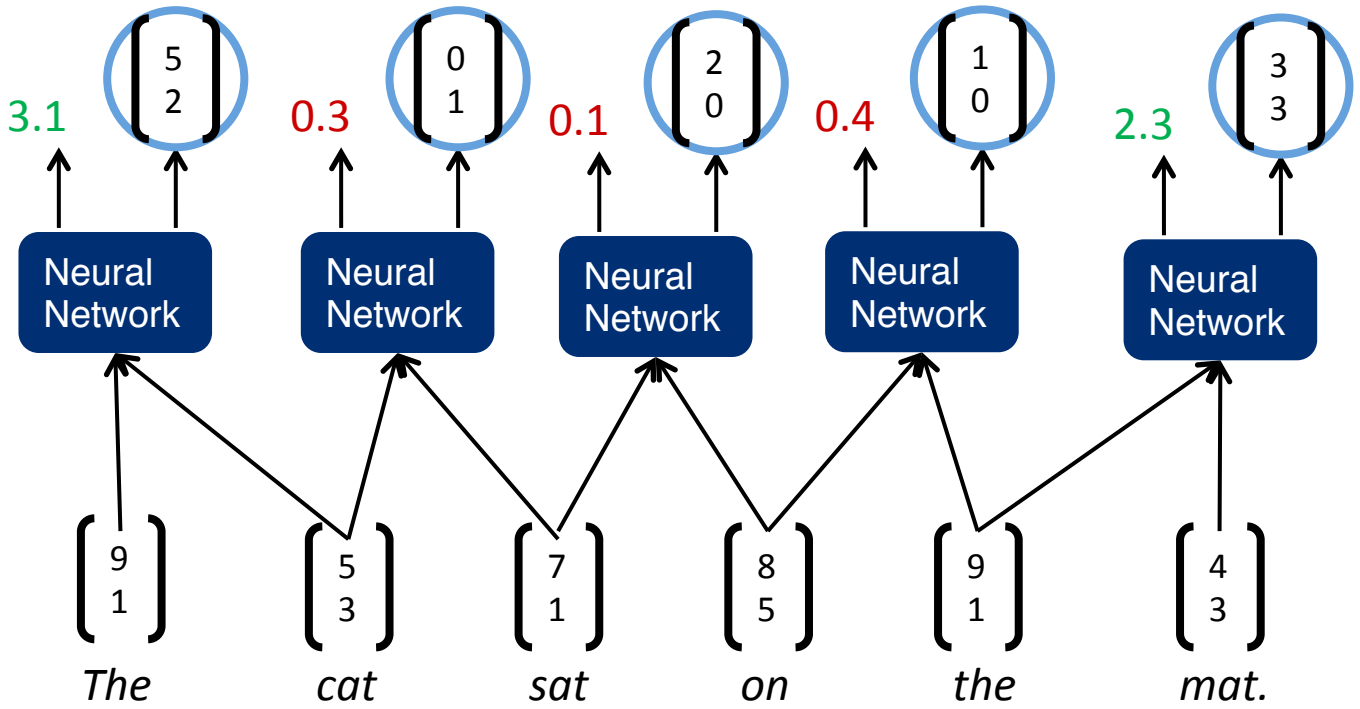
where sigmoid:



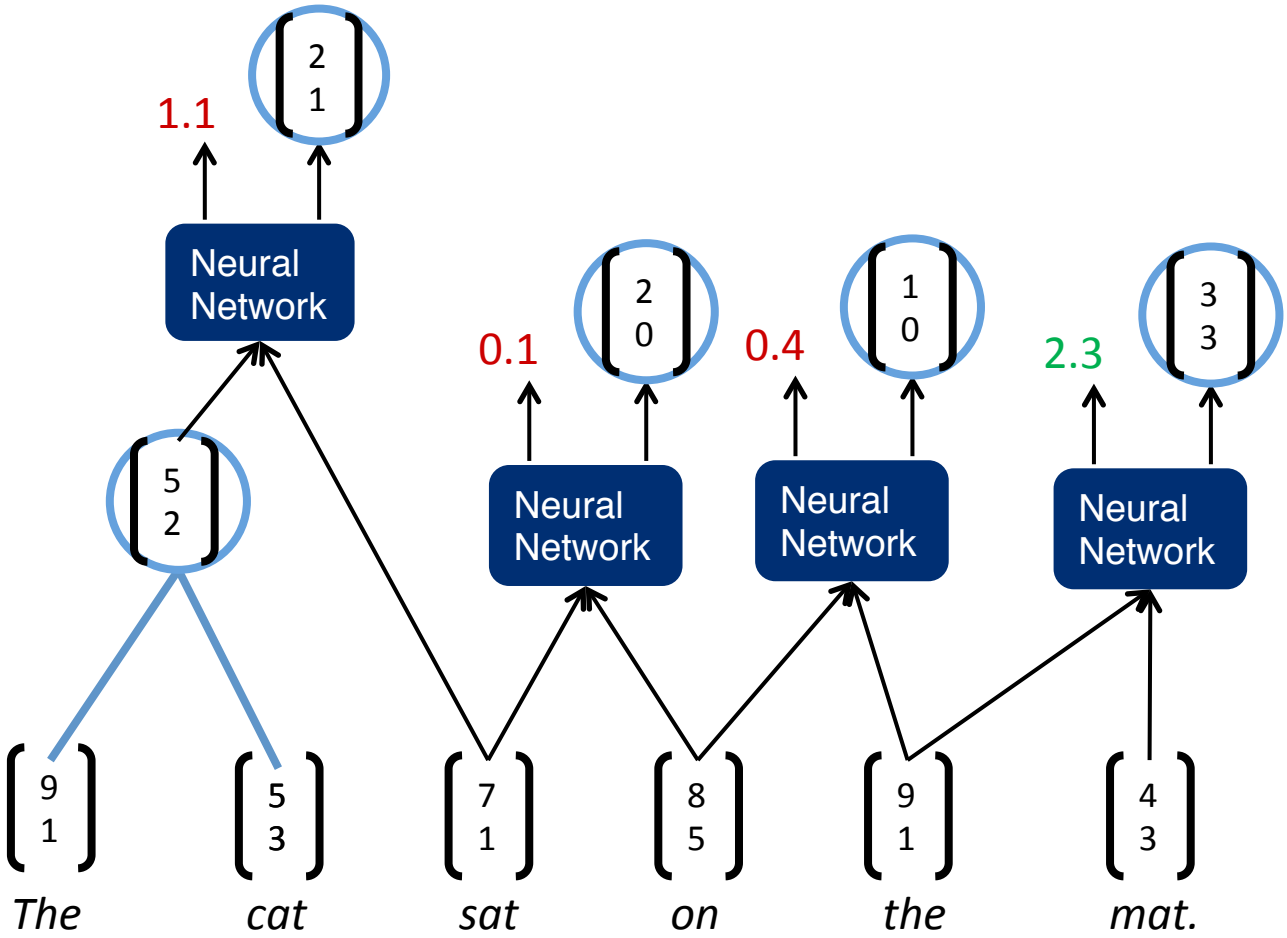
UNIVERSITEIT STELLENBOSCH UNIVERSITY

mih media lab.

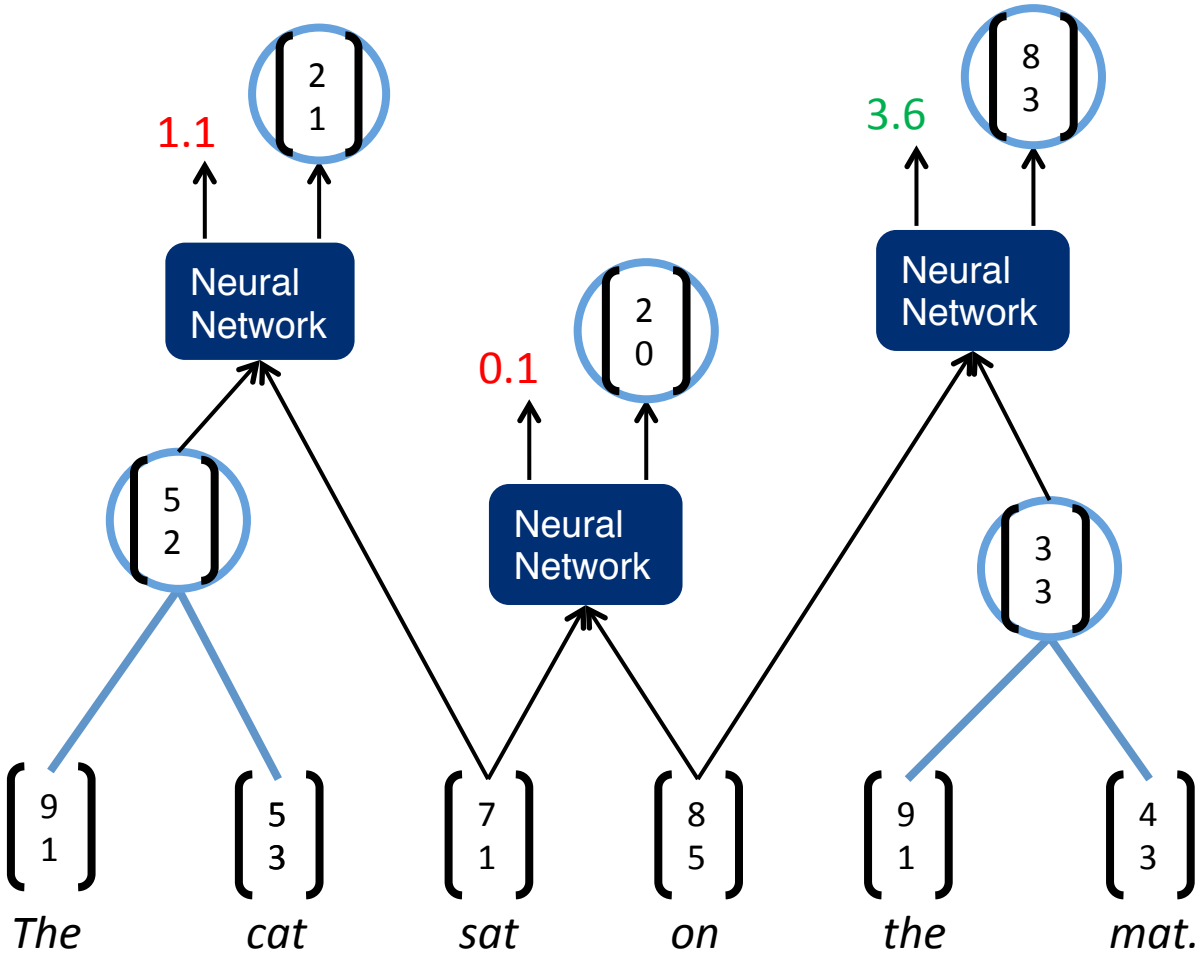
Parsing a sentence



Parsing a sentence



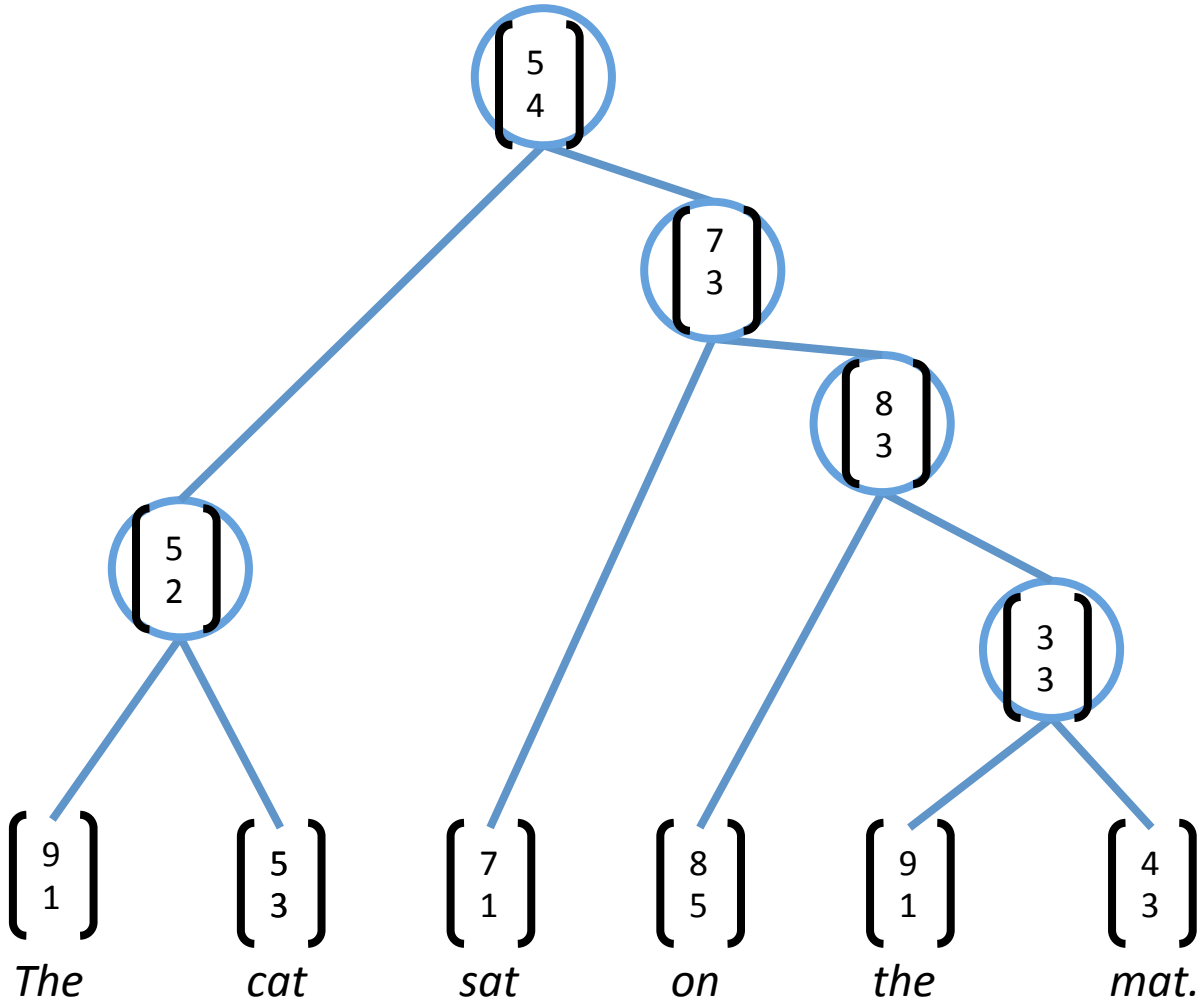
Parsing a sentence



UNIVERSITEIT STELLENBOSCH UNIVERSITY

mih media lab.

Parsing a sentence

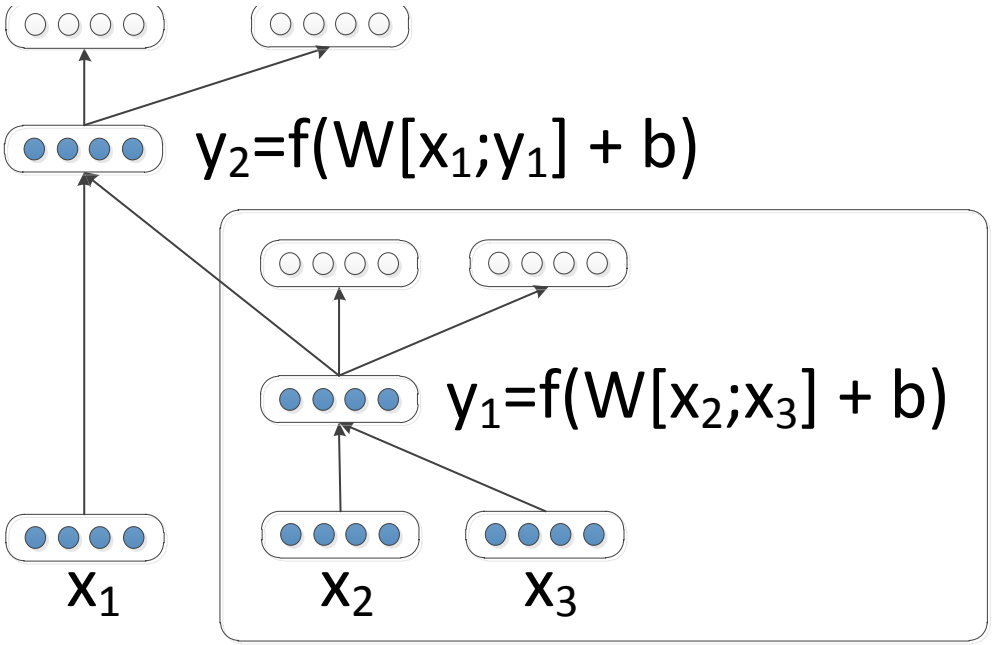


Recursive Auto-encoders (RAEs)

[Socher et al. 2011]

- Similar to a Recursive Neural Net, but instead of a **supervised** score we compute an unsupervised **reconstruction error** at each node:

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



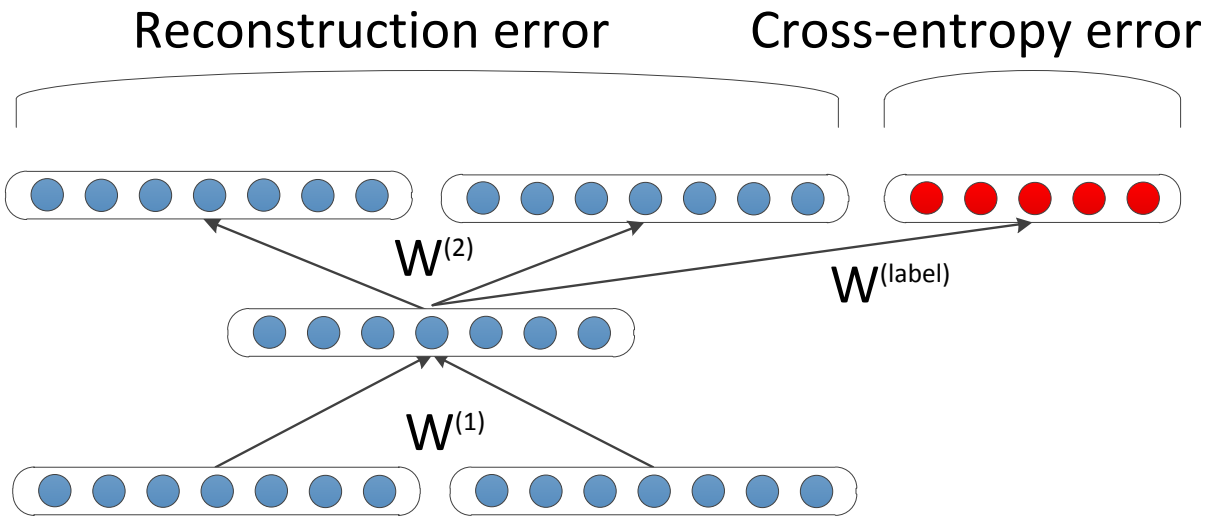
UNIVERSITEIT STELLENBOSCH UNIVERSITY

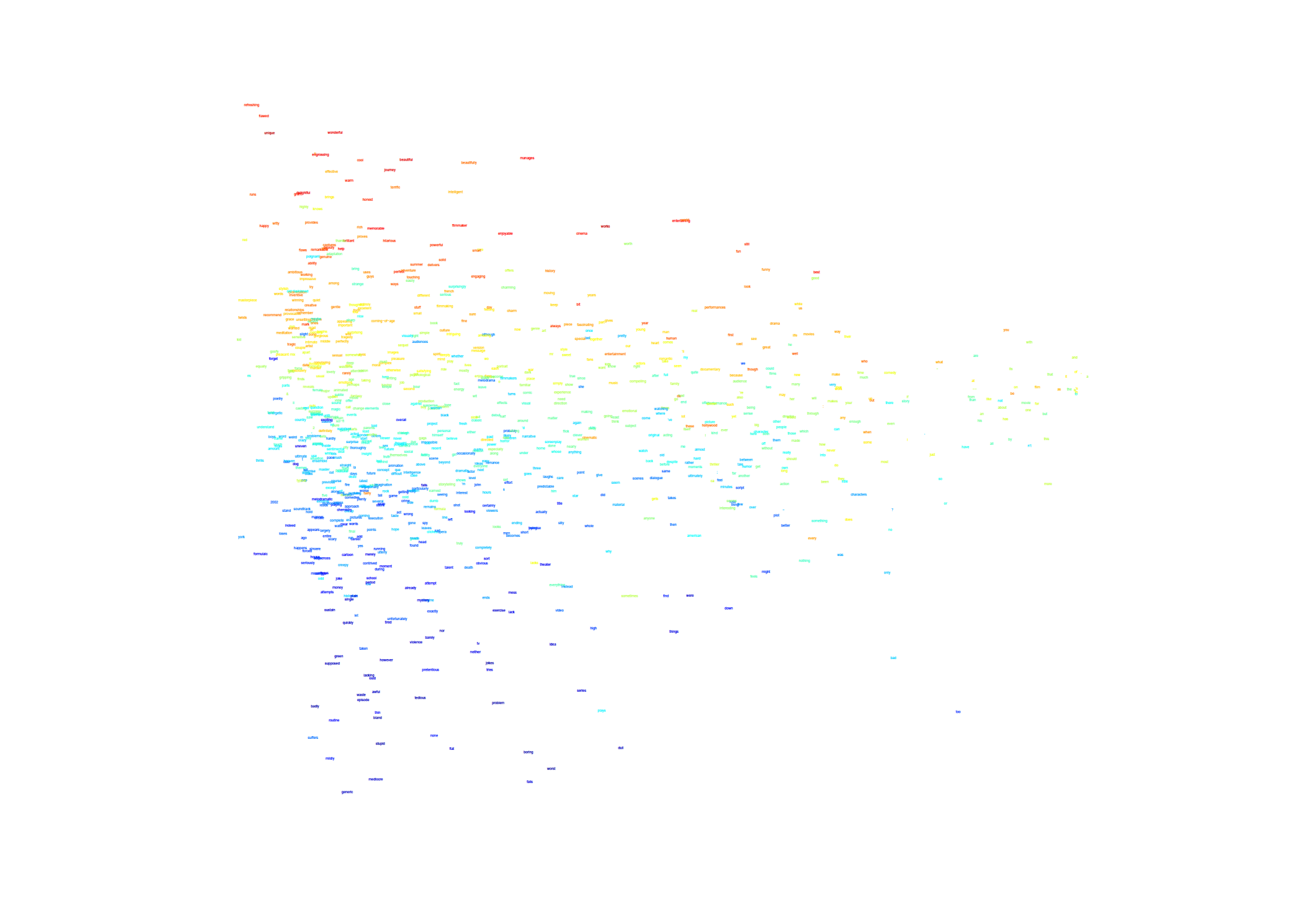
mih media lab.

Semi-supervised RAE

[Socher et al 2011]

- In order for representations to capture sentiment, *we train a softmax classifier on the learned representations*
- Error is a weighted combination of reconstruction error and cross-entropy (distribution likelihood)



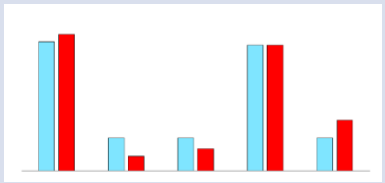
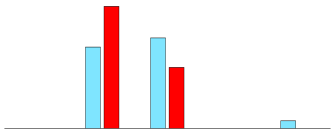
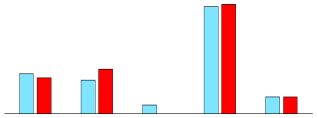

refreshing
flawed
unique
wonderful
effortlessly
cool
beautiful
beautifully
manages
journey
effective
warm
terrific
intelligent
runs
honest
highly
knows
brings
happy
witty
provides
rich
memorable
filmmaker
works
entertaining
enjoyable
cinema
proves
brilliant
hilarious
powerful
worth
still
fun
smart
solid
delivers
summer
offers
history
funny
best
good
engaging
touching
look
surprisingly
charming
moving
years
performances
real
white
as
drama
gay
filmmaking
keep
bit
fascinating
gives
always
piece
year
once
pretty
young
man
human
heart
comes
first
movies
way
their
cast
see
great
he
entertainment
sweet
style
mr
well
who
what
we
though
could
films
now
make
time
much
comedy
because
audience
documentary
quite
actors
night
seen
romantic
my
fans
ward
show
music
compelling
after
full
family
two
many
very
makes
your
but
there
story
may
her
will
also
boring
sense
through
any
enough
even
big
other
people
those
which
can
when
some
how
them
made
off
without
almost
really
should
between
rather
moments
thriller
humor
girl
own
most
just
info
never
have
all
try
this
more
characters
script
minutes
american
every
was
only
might
nothing
why
sometimes
find
were
down
things
high
idea
bad
too
series
plays
problem
none
dull
boring
worst
tails
mediocre
generic
stupid
flat
mindy
suffers
routine
bland
thin
badly
awful
waste
episode
tedious
lacking
lame
however
supposed
green
pretentious
jokes
tries
neither
tv
violence
barely
nor
taken
unfortunately
tired
quickly
sustain
let
exactly
exercise
lack
video
mess
ends
mystery
single
already
attempt
during
school
period
joke
money
attempts
odd
talent
death
obvious
sort
theater
looks
moment
continues
creepy
cartoon
barely
utterly
seriously
audiences
happens
ignore
forced
formulaic
york
indeed
appears
largely
loses
ago
entire
scary
flaws
remarkably
help
poignant
genuine
ability
ambitious
working
impressive
among
strange
ways
bring
uses
guys
portrayal
adventure
touching
easily
different
serious
surprisingly
french
charm
sure
small
book
fine
coming-of-age
rise
relationships
provocative
member
recommend
grace
unsettling
notable
mark
shot
filled
eight
gorgeous
meditation
tragedy
fragile
couple
intimate
middle
perfectly
sequel
audiences
visually
simple
culture
intriguing
although
version
message
spirit
whether
mind
play
lives
each
quirk
mostly
ride
otherwise
rarely
pedestrian
gripping
finds
parts
reveals
female
animated
subtle
opening
perhaps
talking
second
tour
fact
energy
leave
melodrama
filmmakers
place
familiar
simply
she
turns
comic
effects
visual
experience
need
direction
making
emotional
watching
where
end
other
performances
such
picture
hollywood
these
itself
kind
ever
me
almost
watch
old
back
people
behind
hard
moments
thriller
humor
girl
between
take
tone
another
ultimately
ca
feel
minutes
script
action
been
characters
takes
material
over
interesting
pilot
anyone
then
better
american
every
was
only
might
nothing
why
sometimes
find
were
down
things
high
idea
bad
too
series
plays
problem
none
dull
boring
worst
tails
mediocre
generic
stupid
flat
mindy
suffers
routine
bland
thin
badly
awful
waste
episode
tedious
lacking
lame
however
supposed
green
pretentious
jokes
tries
neither
tv
violence
barely
nor
taken
unfortunately
tired
quickly
sustain
let
exactly
exercise
lack
video
mess
ends
mystery
single
already
attempt
during
school
period
joke
money
attempts
odd
talent
death
obvious
sort
theater
looks
moment
continues
creepy
cartoon
barely
utterly
seriously
audiences
happens
ignore
forced
formulaic
york
indeed
appears
largely
loses
ago
entire
scary
flaws
remarkably
help
poignant
genuine
ability
ambitious
working
impressive
among
strange
ways
bring
uses
guys
portrayal
adventure
touching
easily
different
serious
surprisingly
french
charm
sure
small
book
fine
coming-of-age
rise
relationships
provocative
member
recommend
grace
unsettling
notable
mark
shot
filled
eight
gorgeous
meditation
tragedy
fragile
couple
intimate
middle
perfectly
sequel
audiences
visually
simple
culture
intriguing
although
version
message
spirit
whether
mind
play
lives
each
quirk
mostly
ride
otherwise
rarely
pedestrian
gripping
finds
parts
reveals
female
animated
subtle
opening
perhaps
talking
second
tour
fact
energy
leave
melodrama
filmmakers
place
familiar
simply
she
turns
comic
effects
visual
experience
need
direction
making
emotional
watching
where
end
other
performances
such
picture
hollywood
these
itself
kind
ever
me
almost
watch
old
back
people
behind
hard
moments
thriller
humor
girl
between
take
tone
another
ultimately
ca
feel
minutes
script
action
been
characters
takes
material
over
interesting
pilot
anyone
then
better
american
every
was
only
might
nothing
why
sometimes
find
were
down
things
high
idea
bad
too
series
plays
problem
none
dull
boring
worst
tails
mediocre
generic

UNIVERSITEIT STELLENBOSCH UNIVERSITY

mih media lab.

Sentiment distributions

- 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. ...

UNIVERSITEIT STELLENBOSCH UNIVERSITY

mih media lab.

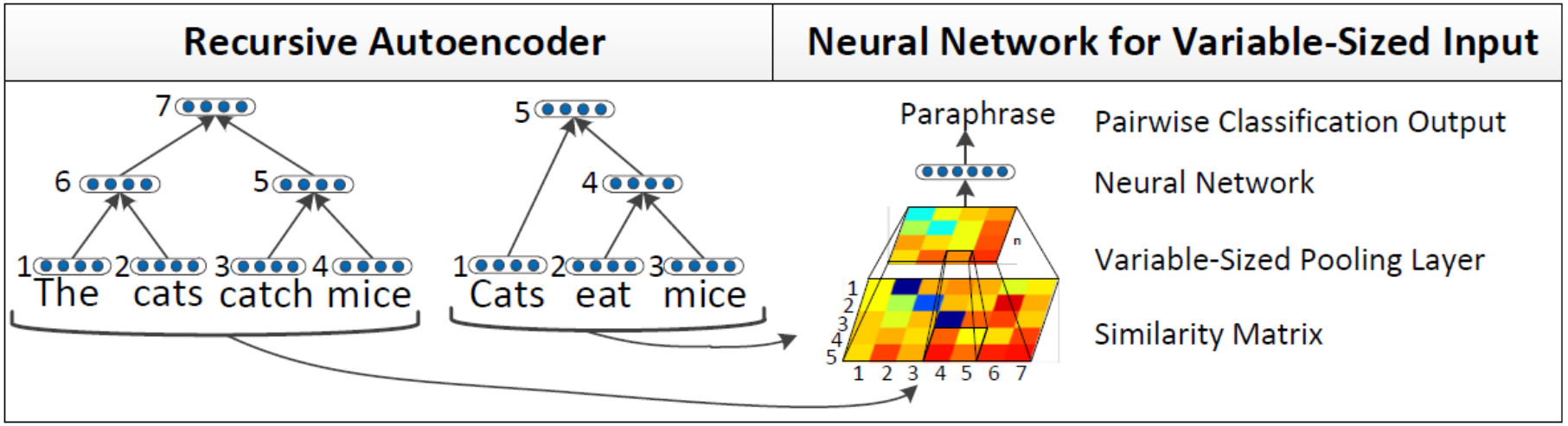
1. Motivation
2. Background Theory
3. Neural Language Models
4. Other NLP Tasks
    1. POS, Chunking, NER, SRL
    2. Sentiment Analysis
    3. Paraphrase Detection
    4. Semantic Parsing
5. My Work
6. End

UNIVERSITEIT STELLENBOSCH UNIVERSITY

Recursive Auto-encoders for Full Sentence Paraphrase Detection

[Socher et al. 2011]

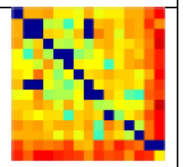
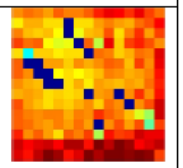
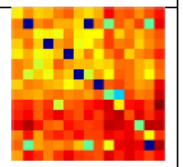
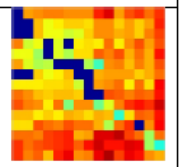
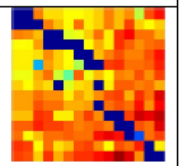
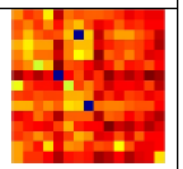
Unsupervised RAE and a pair-wise sentence comparison of nodes in parse trees



UNIVERSITEIT STELLENBOSCH UNIVERSITY

Recursive Auto-encoders for Full Sentence Paraphrase Detection

mih media lab.

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	

UNIVERSITEIT STELLENBOSCH UNIVERSITY

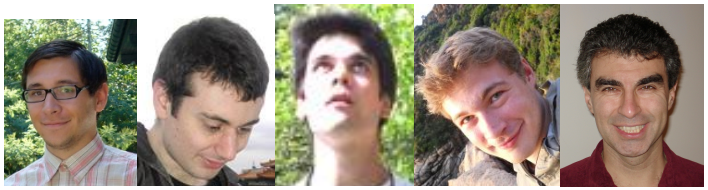
mih media lab.

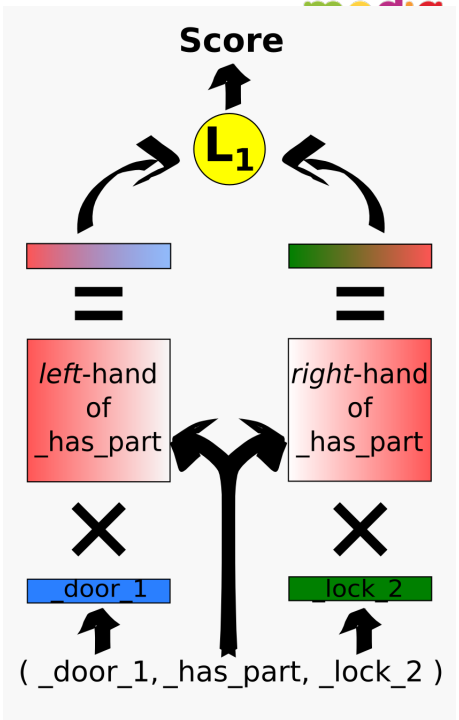
1. Motivation
2. Background Theory
3. Neural Language Models
4. Other NLP Tasks
    1. POS, Chunking, NER, SRL
    2. Sentiment Analysis
    3. Paraphrase Detection
    4. Semantic Parsing
5. My Work
6. End

UNIVERSITEIT STELLENBOSCH UNIVERSITY

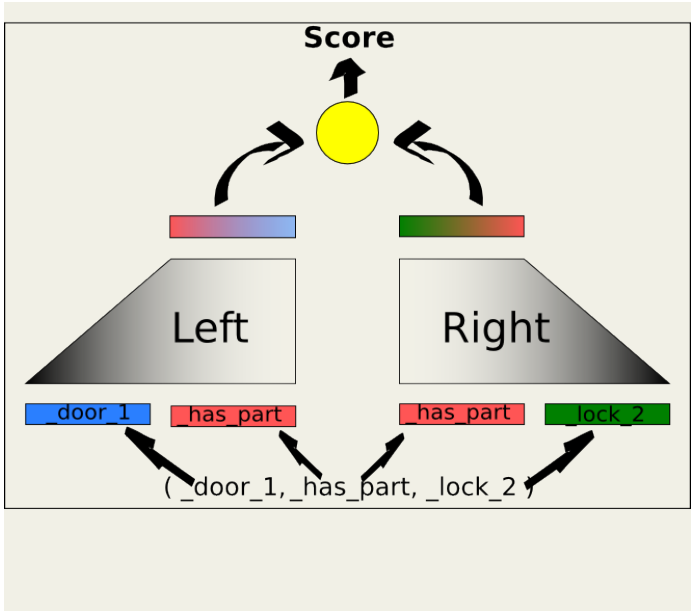
Modeling Semantics

Learning Structured Embeddings of Knowledge Bases, [Bordes, Weston, Collobert & Bengio, AAAI 2011]





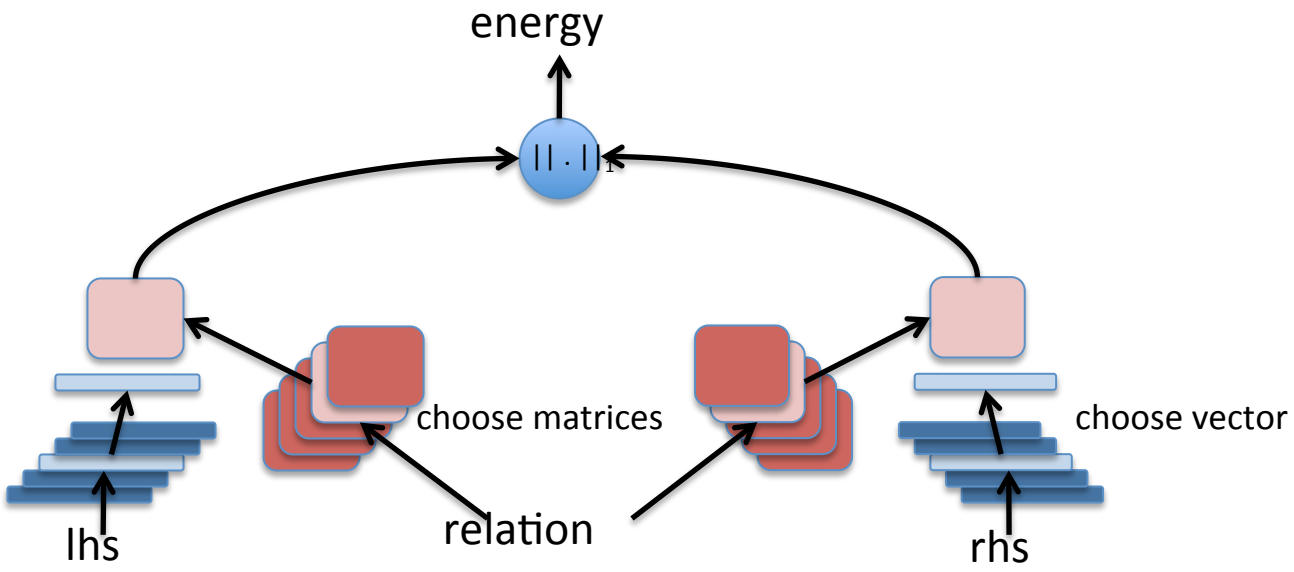
Joint Learning of Words and Meaning Representations for Open-Text Semantic Parsing, [Bordes, Glorot, Weston & Bengio, AISTATS 2012]



UNIVERSITEIT STELLENBOSCH UNIVERSITY

mih media lab.

Modeling Relations with Matrices



Model (lhs, relation, rhs)

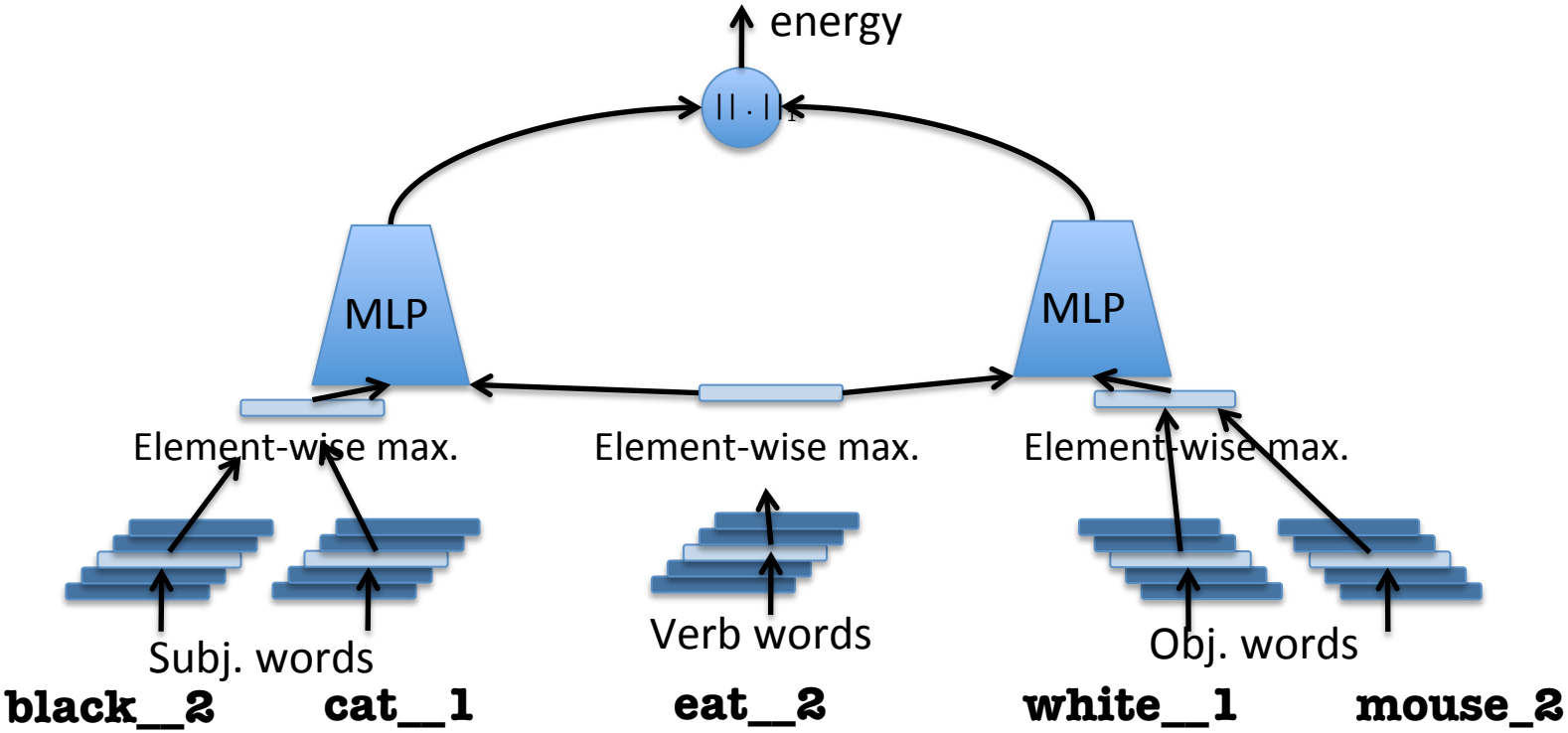
Each concept = 1 embedding vector

Each relation = 2 matrices. **Matrix acts like an operator.**

Ranking criterion

Energy = low for training examples, high o/w

Training on Full Sentences

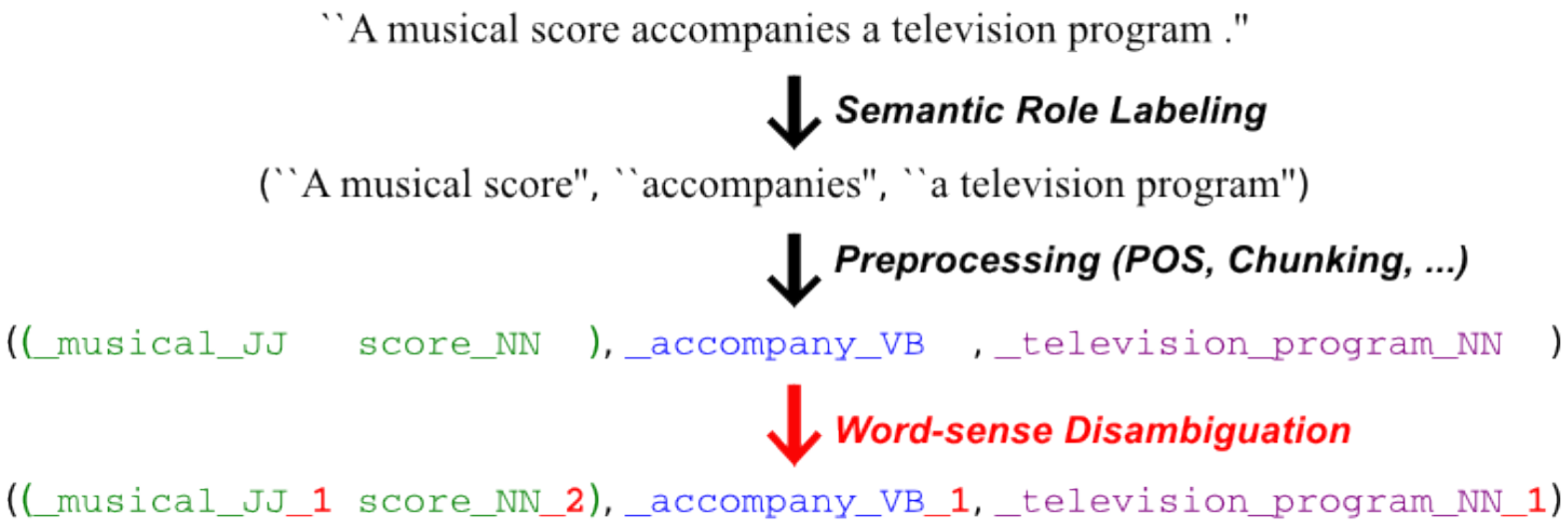


Use SENNA (Collobert 2010) = embedding-based NLP tagger for Semantic Role Labeling, breaks sentence into (subject part, verb part, object part)

→ Use max-pooling to aggregate embeddings of words inside each part

Open-Text Semantic Parsing

- 3 steps:



- Last formula defines the Meaning Representation (MR).

UNIVERSITEIT STELLENBOSCH UNIVERSITY

mih media lab.

Training Criterion

- Intuition: If an entity of a triplet was missing, we would like our model to predict it correctly i.e. to give it the lowest energy.

- For example, this would allow us to answer questions like “what is part of a car?”

- Hence, for any training triplet $x_i = (lhs_i, rel_i, rhs_i)$ we would like:
  (1) $E(lhs_i, rel_i, rhs_i) < E(lhs_j, rel_i, rhs_i)$,
  (2) $E(lhs_i, rel_i, rhs_i) < E(lhs_i, rel_j, rhs_i)$,
  (3) $E(lhs_i, rel_i, rhs_i) < E(lhs_i, rel_i, rhs_j)$,
- That is: the energy function E is trained to rank training samples below all other triplets.

UNIVERSITEIT STELLENBOSCH UNIVERSITY

mih media lab.

Training Algorithm

Pseudo-likelihood + Uniform sampling of negative variants

Train by stochastic gradient descent (SGD):

1. Randomly select a positive training triplet $x_i = (\text{lhs}_i, \text{rel}_i, \text{rhs}_i)$.
2. Randomly select constraint (1), (2) or (3) and an entity $\tilde{e}$:
   - If constraint (1), construct negative triplet $\tilde{x} = (\tilde{e}, \text{rel}_i, \text{rhs}_i)$.
   - Else if constraint (2), construct $\tilde{x} = (\text{lhs}_i, \tilde{e}, \text{rhs}_i)$.
   - Else, construct $\tilde{x} = (\text{lhs}_i, \text{rel}_i, \tilde{e})$.
3. If $E(x_i) > E(\tilde{x}) - 1$ make a gradient step to minimize:

   $\max(0, 1 - E(\tilde{x}) + E(x_i))$.
4. Constrain embedding vectors to norm 1

Question Answering: Implicitly adding new relations to WN or FB

	Model (All)	*TextRunner*
lhs	**_army_NN_1**	*army*
rel	**_attack_VB_1**	*attacked*
top ranked *rhs*	_troop_NN_4	*Israel*
	_armed_service_NN_1	*the village*
	_ship_NN_1	*another army*
	_territory_NN_1	*the city*
	_military_unit_NN_1	*the fort*
top ranked *lhs*	_business_firm_NN_1	*People*
	_person_NN_1	*Players*
	_family_NN_1	*one*
	_payoff_NN_3	*Students*
	_card_game_NN_1	*business*
rel	**_earn_VB_1**	*earn*
rhs	**_money_NN_1**	*money*

MRs inferred from text define triplets between WordNet synsets.

Model captures knowledge about relations between nouns and verbs.

→ Implicit addition of new relations to WordNet

→ Generalize Freebase

Embedding Near Neighbours of Words & Senses

_mark_NN	_mark_NN_1	_mark_NN_2
_indication_NN	_score_NN_1	_marking_NN_1
_print_NN_3	_number_NN_2	_symbolizing_NN_1
_print_NN	_gradation_NN	_naming_NN_1
_roll_NN	_evaluation_NN_1	_marking_NN
_pointer_NN	_tier_NN_1	_punctuation_NN_3

_take_VB	_canary_NN	_different_JJ_1
_bring_VB	_sea_mew_NN_1	_eccentric_NN
_put_VB	_yellowbird_NN_2	_dissimilar_JJ
_ask_VB	_canary_bird_NN_1	_same_JJ_2
_hold_VB	_larus_marinus_NN_1	_similarity_NN_1
_provide_VB	_mew_NN	_common_JJ_1

UNIVERSITEIT STELLENBOSCH UNIVERSITY

mih media lab.

1. Motivation
2. Background Theory
3. Neural Language Models
4. Other NLP Tasks
5. My Work
6. End

UNIVERSITEIT STELLENBOSCH UNIVERSITY

mih media lab.

The real world is noisy

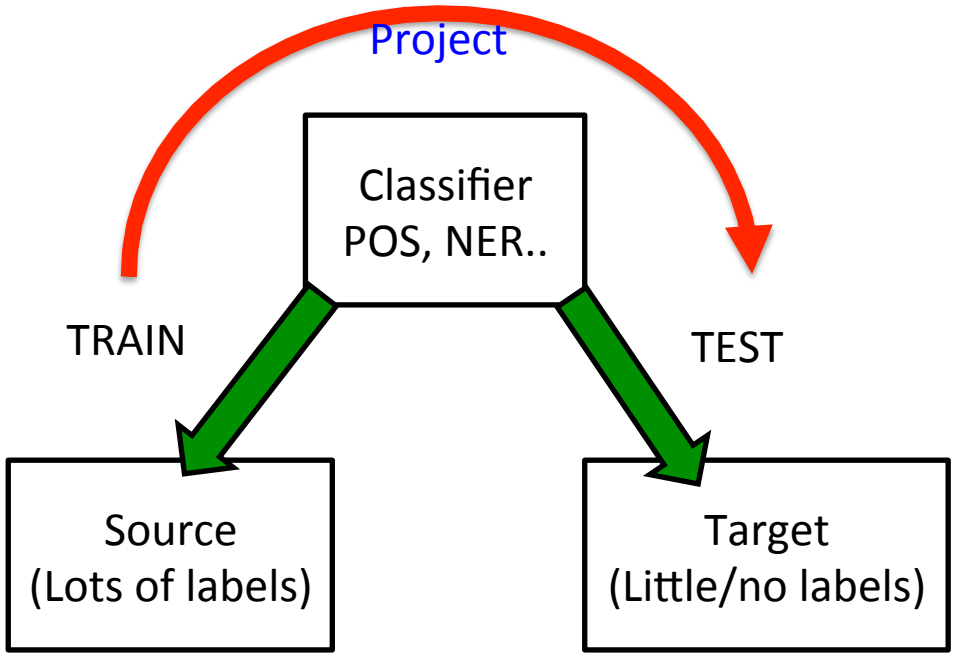
- Supervised methods have high accuracy when P(test) ≈ P(train)
- But the real world is noisy, systems based on carefully engineered features in one domain might not easily transfer to another

UNIVERSITEIT STELLENBOSCH UNIVERSITY

mih media lab.

Domain adaptation

- How can we train a system on a source distribution to perform well on a target distribution?

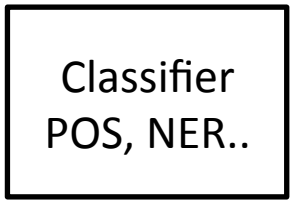


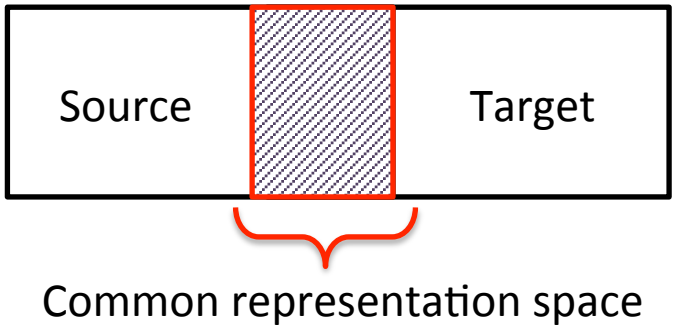
UNIVERSITEIT STELLENBOSCH UNIVERSITY

mih media lab.

Domain adaptation

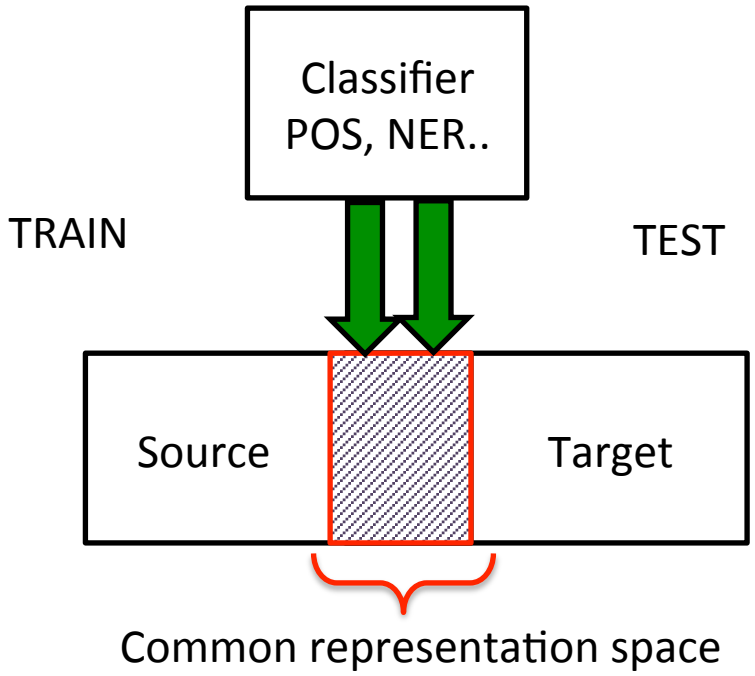
- How can we train a system on a source distribution to perform well on a target distribution?





Domain adaptation

- How can we train a system on a source distribution to perform well on a target distribution?



UNIVERSITEIT STELLENBOSCH UNIVERSITY

mih media lab.

Cross-domain transfer of linguistic features

- Transfer learning:

  Given a resource-rich source domain A, and a resource-poor target domain B, learn source and target models but constrain their embedded spaces to align as best possible

- Then use task classifiers trained on source feature representation for performing task in target domain

UNIVERSITEIT STELLENBOSCH UNIVERSITY

mih media lab.

Part-of-Speech Tagging

English Language (WSJ)

DT	**NN**	**VBZ**	**DT**	**NN**	**IN**	**DT**	**JJ**	**NN**	**CC**	
The	clash	is	a	sign	of	a	new	toughness	and	
NN		**IN**	**NNP**	**POS**	**JJ**		**JJ**	**NNS**	**.**	
divisiveness		in	Japan	‘s	once-cozy		financial	circles	.	

SOURCE

Different Language (Afrikaans)

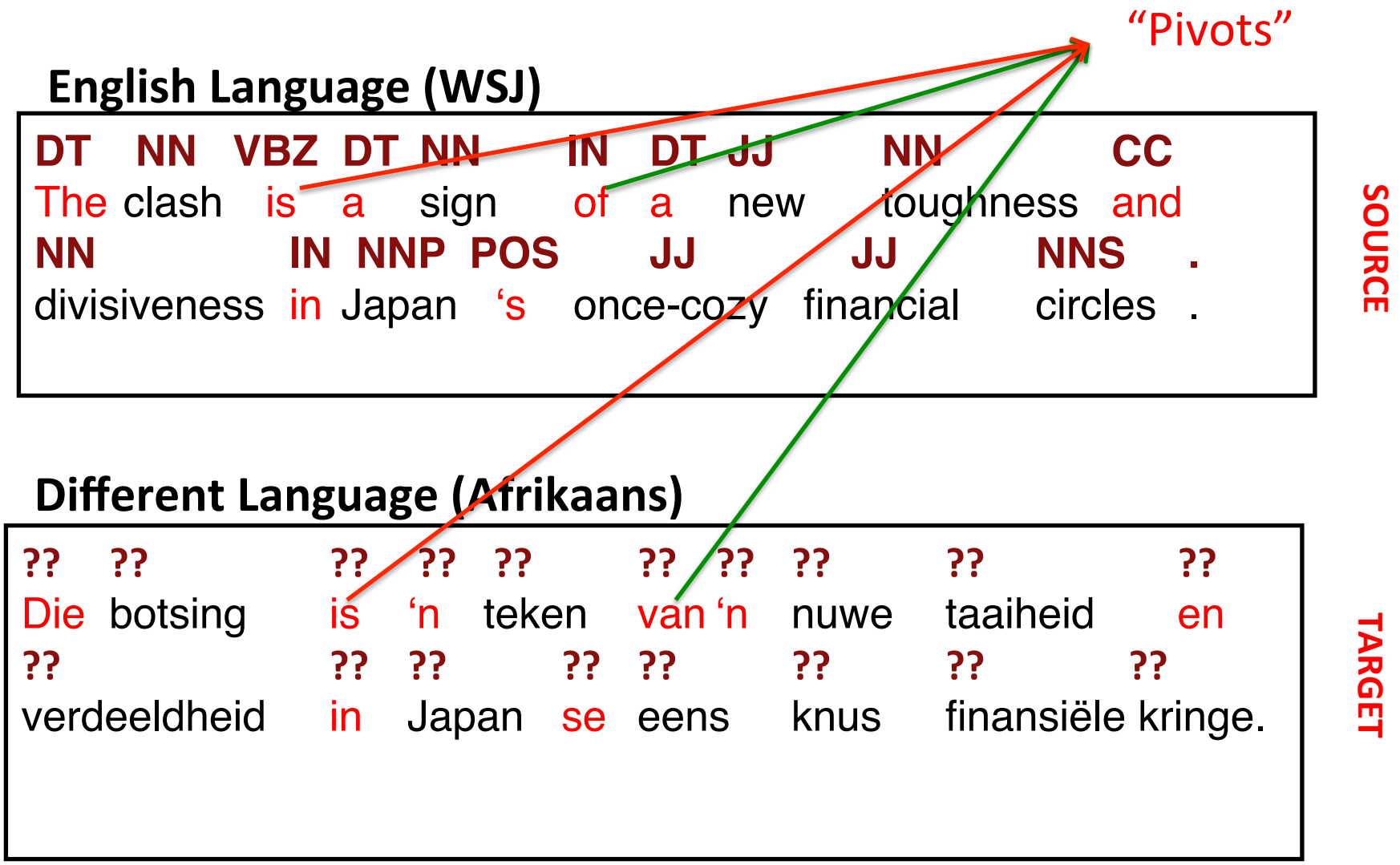
??	**??**	**??**	**??**	**??**	**??**	**??**	**??**	**??**	**??**	
Die	botsing	is	‘n	teken	van	‘n	nuwe	taaiheid	en	
??		**??**	**??**	**??**	**??**		**??**	**??**	**??**	
verdeeldheid		in	Japan	se	eens		knus	finansiële	kringe.	

TARGET

UNIVERSITEIT STELLENBOSCH UNIVERSITY

mih media lab.

Part-of-Speech Tagging



UNIVERSITEIT STELLENBOSCH UNIVERSITY

mih media lab.

Part-of-Speech Tagging

English Language (WSJ)

DT	NN	VBZ	DT	NN	IN	DT	JJ	NN	CC
The	clash	is	a	sign	of	a	new	toughness	and
NN		**IN**	**NNP**	**POS**	**JJ**		**JJ**	**NNS**	**.**
divisiveness		in	Japan	‘s	once-cozy		financial	circles	.

SOURCE

Different Language (Afrikaans)

??	??	??	??	??	??	??	??	??	??
Die	botsing	is	‘n	teken	van	‘n	nuwe	taaiheid	en
??		??	??	??	??		??	??	??
verdeeldheid		in	Japan	se	eens		knus	finansiële	kringe.

TARGET

UNIVERSITEIT STELLENBOSCH UNIVERSITY

mih media lab.

Part-of-Speech Tagging

English Language (WSJ)

DT	**NN**	**VBZ**	**DT**	**NN**	**IN**	**DT**	**JJ**	**NN**	**CC**	
The	clash	is	a	sign	of	a	new	toughness	and	
NN		**IN**	**NNP**	**POS**	**JJ**		**JJ**	**NNS**	**.**	
divisiveness		in	Japan	‘s	once-cozy		financial	circles	.	

SOURCE

Different Language (Afrikaans)

DT	**??**	**VBZ**	**DT**	**??**	**IN**	**DT**	**??**	**??**	**CC**	
Die	botsing	is	‘n	teken	van	‘n	nuwe	taaiheid	en	
??		**JJ**	**??**	**POS**	**??**		**??**	**??**	**??**	
verdeeldheid		in	Japan	se	eens		knus	finansiële	kringe.	

TARGET

UNIVERSITEIT STELLENBOSCH UNIVERSITY

mih media lab.

Part-of-Speech Tagging

English Language (WSJ)

DT	**NN**	**VBZ**	**DT**	**NN**	**IN**	**DT**	**JJ**	**NN**	**CC**	
The	clash	is	a	sign	of	a	new	toughness	and	SOURCE
NN		**IN**	**NNP**	**POS**	**JJ**		**JJ**	**NNS**	**.**	
divisiveness		in	Japan	‘s	once-cozy		financial	circles	.	

Different Language (Afrikaans)

DT	**??**	**VBZ**	**DT**	**??**	**IN**	**DT**	**??**	**??**	**CC**	
Die	botsing	is	‘n	teken	van	‘n	nuwe	taaiheid	en	TARGET
??		**JJ**	**??**		**POS**	**??**	**??**	**??**	**??**	
verdeeldheid		in	Japan		se	eens	knus	finansiële	kringe.	

UNIVERSITEIT STELLENBOSCH UNIVERSITY

mih media lab.

Part-of-Speech Tagging

English Language (WSJ)

DT	NN	VBZ	DT	NN	IN	DT	JJ	NN	CC
The	clash	is	a	sign	of	a	new	toughness	and
NN		**IN**	**NNP**	**POS**	**JJ**		**JJ**	**NNS**	**.**
divisiveness		in	Japan	‘s	once-cozy		financial	circles	.

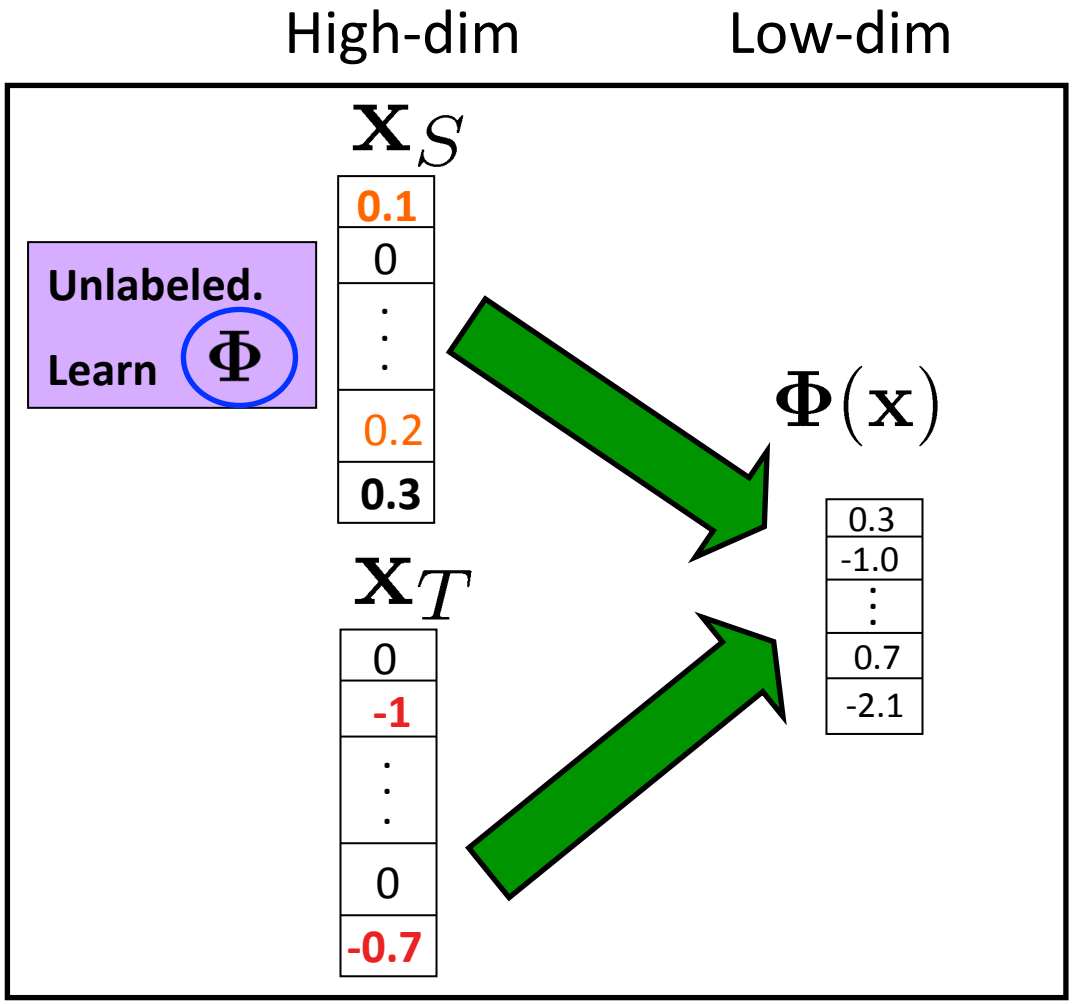
SOURCE

Different Language (Afrikaans)

DT	NN	VBZ	DT	NN	IN	DT	JJ	NN	CC
Die	botsing	is	‘n	teken	van	‘n	nuwe	taaiheid	en
NN		**JJ**	**NNP**	**POS**	**JJ**		**JJ**	**JJ**	**NNS**
verdeeldheid		in	Japan	se	eens		knus	finansiële	kringe.

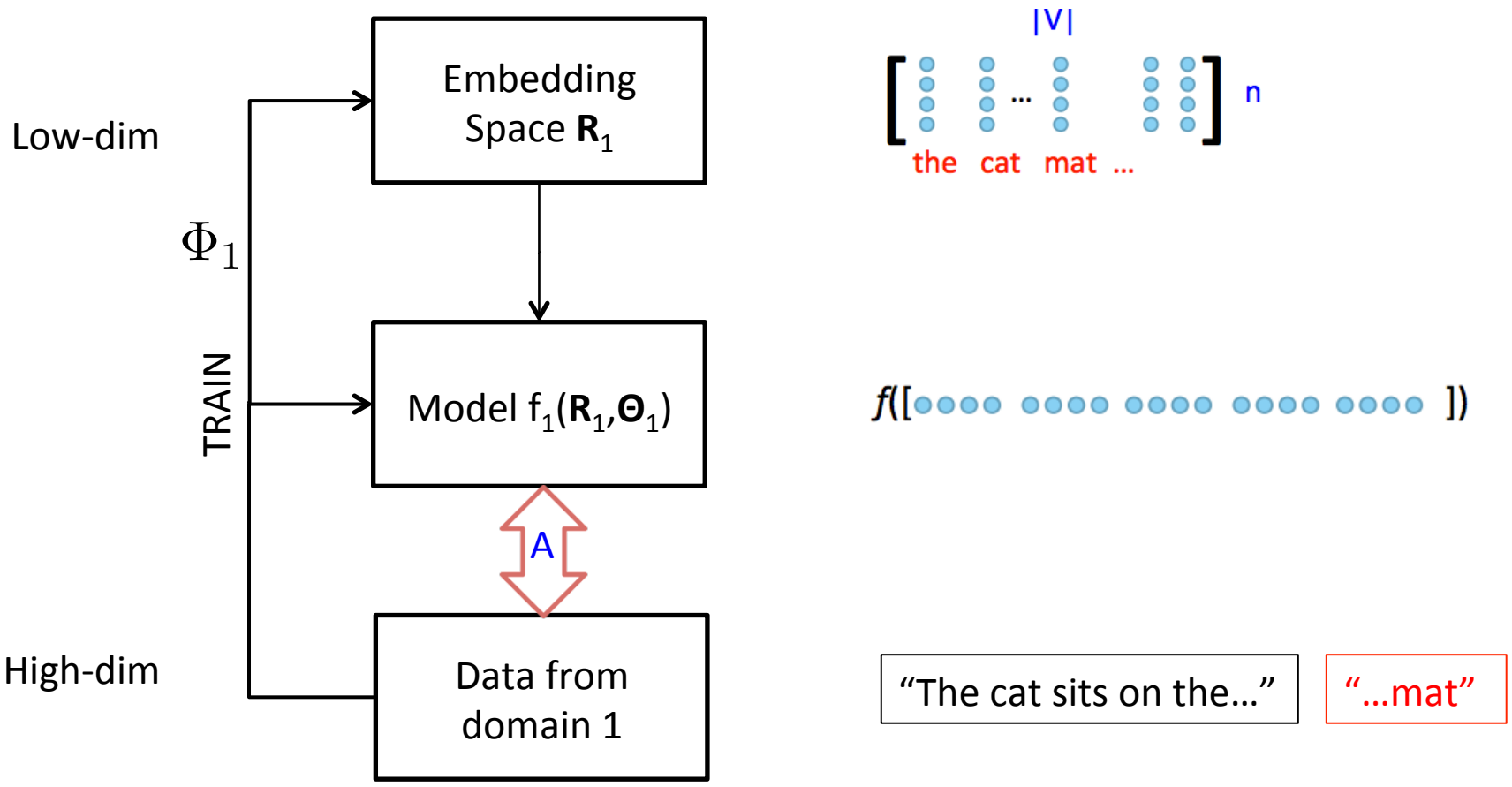
TARGET

Structural Correspondence Learning [Blitzer et al. 2006]



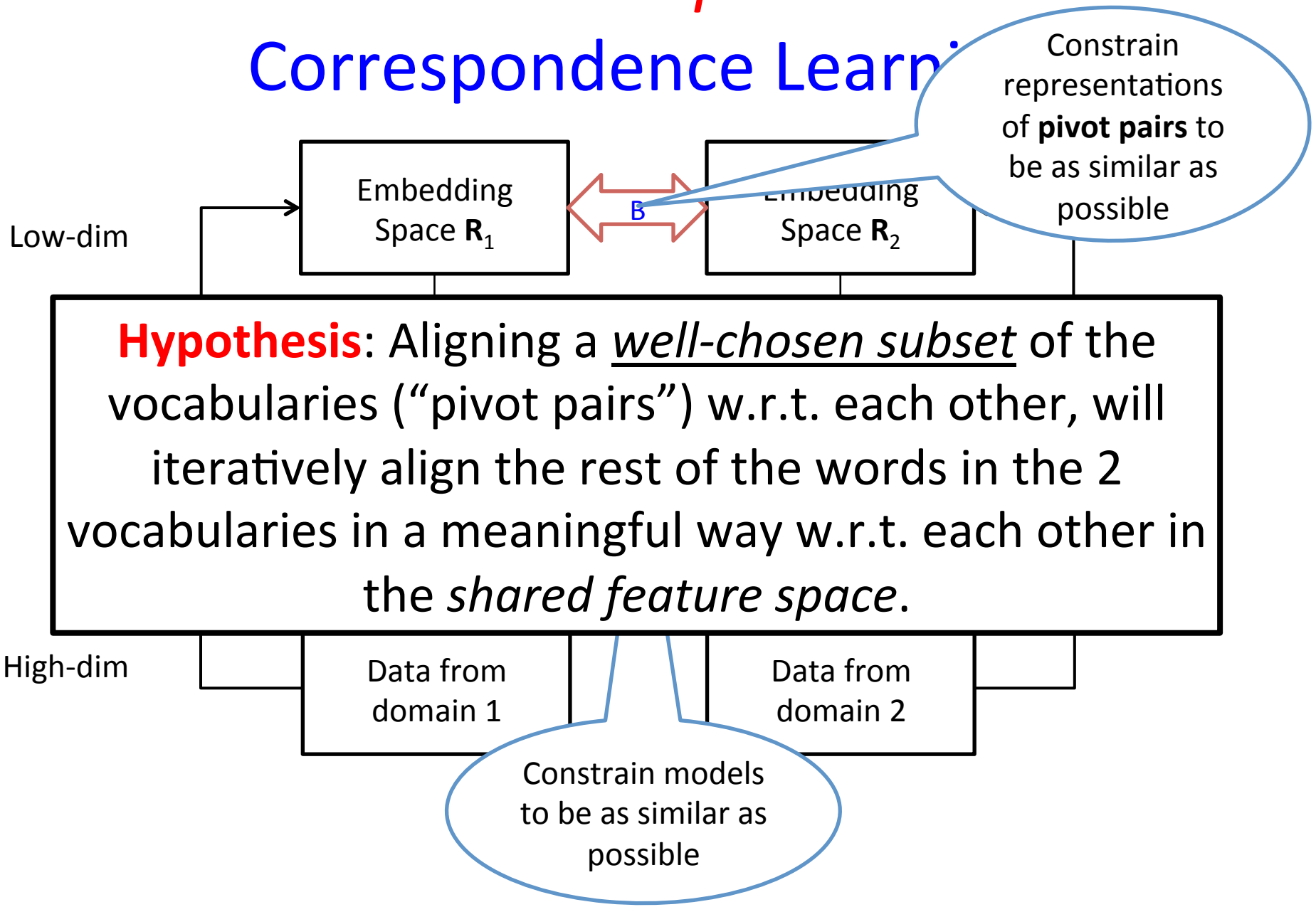
- Φ should make the domains look as similar as possible
- But should also allow us to classify well
- Φ defines a ***linear*** mapping between **feature spaces** of both domains

This Work: Deep Structural Correspondence Learning



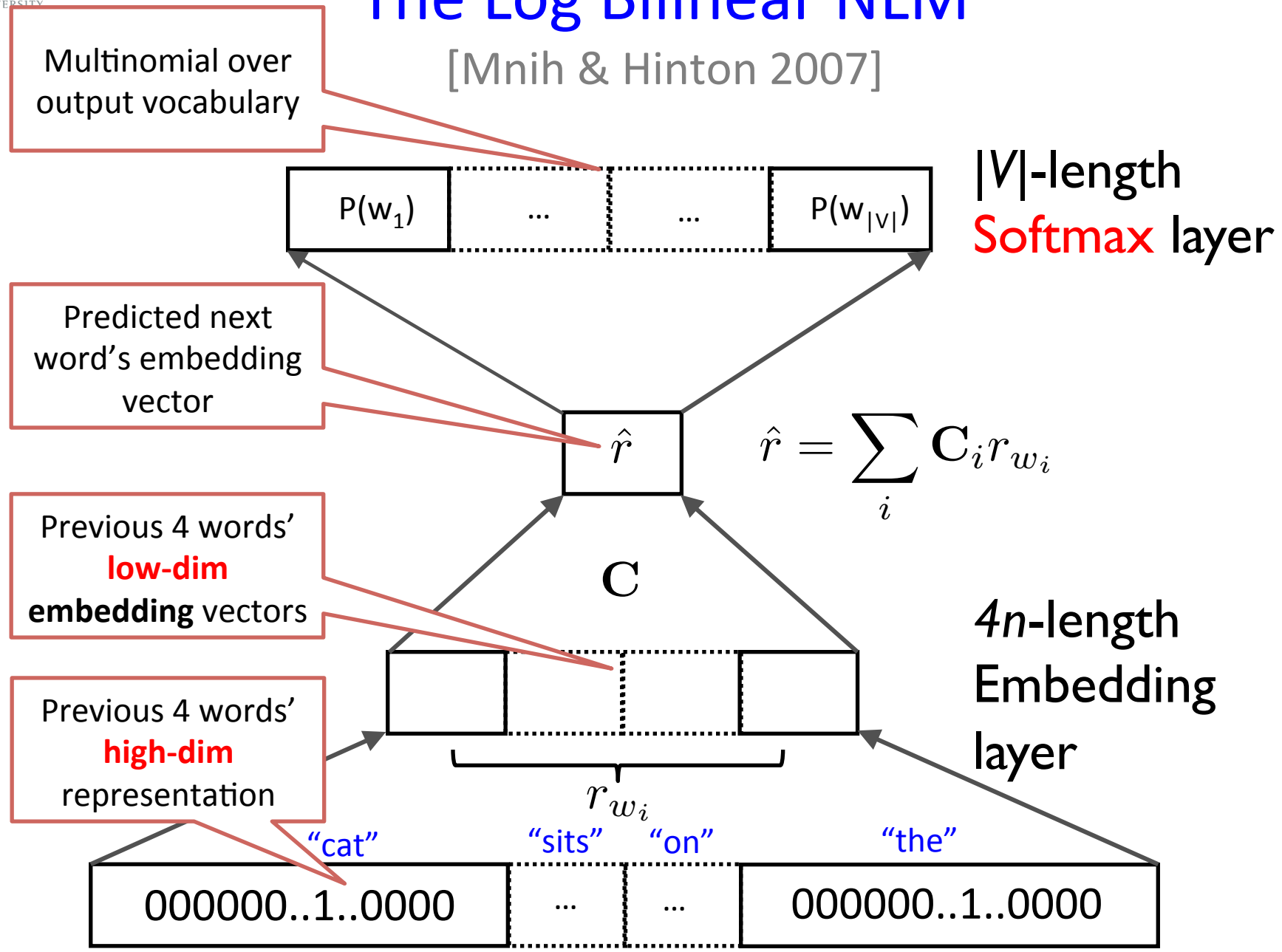
We use **Neural Language Models** [Bengio et al., 2003]

This Work: Deep Structural Correspondence Learni



The Log Bilinear NLM

[Mnih & Hinton 2007]



UNIVERSITEIT STELLENBOSCH UNIVERSITY

mih media lab.

Augmented cost function

$$J^{TOT} = \underbrace{J_1^{nll} + J_2^{nll}}$$

Fit the data well

UNIVERSITEIT STELLENBOSCH UNIVERSITY

mih media lab.

Augmented cost function

$$J^{TOT} = \underbrace{J_1^{nll} + J_2^{nll}}_{\text{Fit the data well}} + \underbrace{\alpha J^R(\theta_1, \theta_2) + \beta J^f(\theta_1, \theta_1)}_{\text{Learn structural similarities}}$$

UNIVERSITEIT STELLENBOSCH UNIVERSITY

mih media lab.

Augmented cost function

$$J^{TOT} = J_1^{nll} + J_2^{nll} + \alpha J^R(\theta_1, \theta_2) + \beta J^f(\theta_1, \theta_1)$$

Fit the data well

Learn structural similarities

UNIVERSITEIT STELLENBOSCH UNIVERSITY

mih media lab.

Constraining the Embeddings

$$J^R(\theta_1, \theta_2) = \sum_{\vec{r_i}, \vec{r_j} \in \text{Pivots}} \lambda_{ij} ||\vec{r_i} - \vec{r_j}||^2$$

Push known pivot pairs $(\vec{r_i}, \vec{r_j}) \in V_1 \times V_2$ closer together by minimizing the weighted sum of squared distances between them.

UNIVERSITEIT STELLENBOSCH UNIVERSITY

mih media lab.

Augmented cost function

$$J^{TOT} = J_1^{nll} + J_2^{nll} + \alpha J^R(\theta_1, \theta_2) + \beta J^f(\theta_1, \theta_1)$$

Fit the data well

Learn structural similarities

UNIVERSITEIT STELLENBOSCH UNIVERSITY

mih media lab.

Constraining the learned Functions

- Intuitively: In the cases where it is appropriate, we constrain the learned NLM functions *f* to be as similar as possible while modelling the data as accurately as possible
  - Reduces degrees of freedom
- [Erhan et al. 2010]

UNIVERSITEIT STELLENBOSCH UNIVERSITY

mih media lab.

Experiments I: Synthetic Data

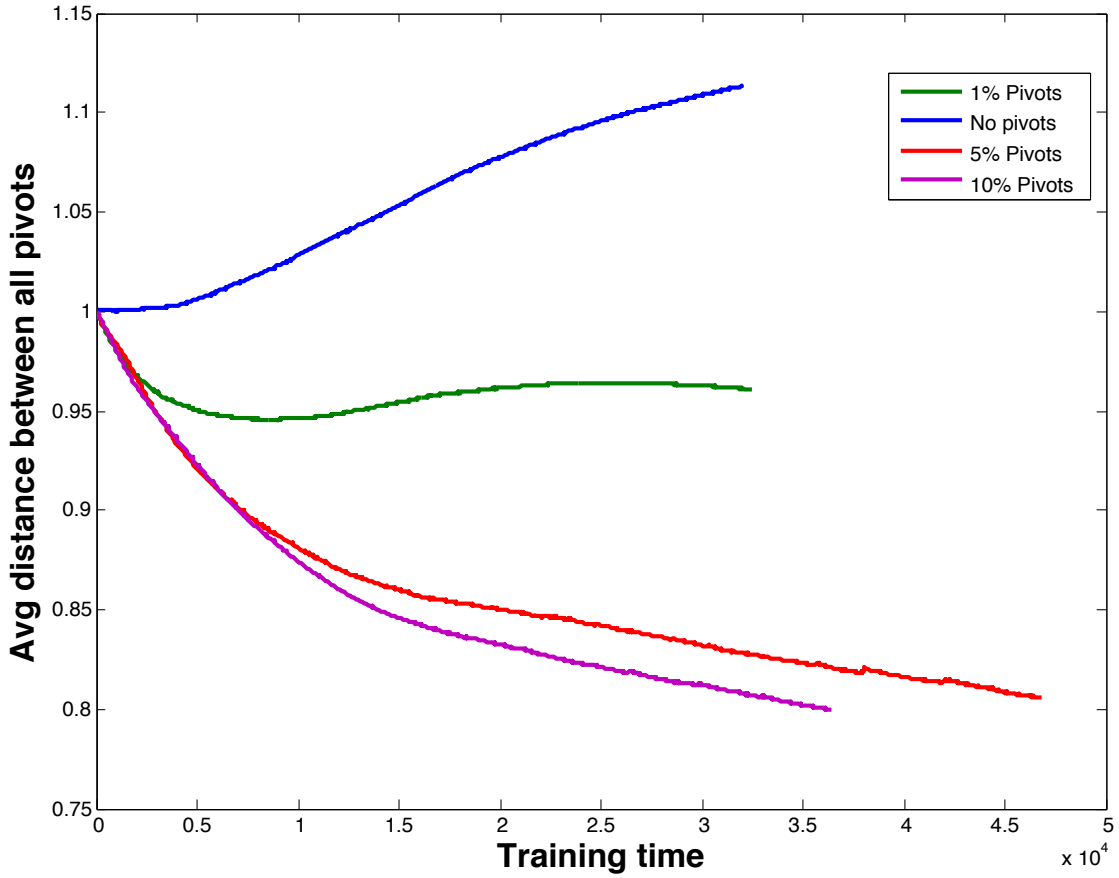
- Sampled two datasets each **without** replacement from LA Times, encoded each in a *different vocabulary*
  - "the_1 president_1 of_1 the_1 united_1 …"
  - "between_2 the_2 hours_2 of_2 midnight_2…"
- Give the networks *k* % of pivot pairs (chose most frequent words)
  - (president_1, president_2), …

EVALUATE: We measure the similarity of vectors for all translation pairs

UNIVERSITEIT STELLENBOSCH UNIVERSITY

mih media lab.

Experiments I: Synthetic Data



UNIVERSITEIT STELLENBOSCH UNIVERSITY

mih media lab.

Experiments II: English-French

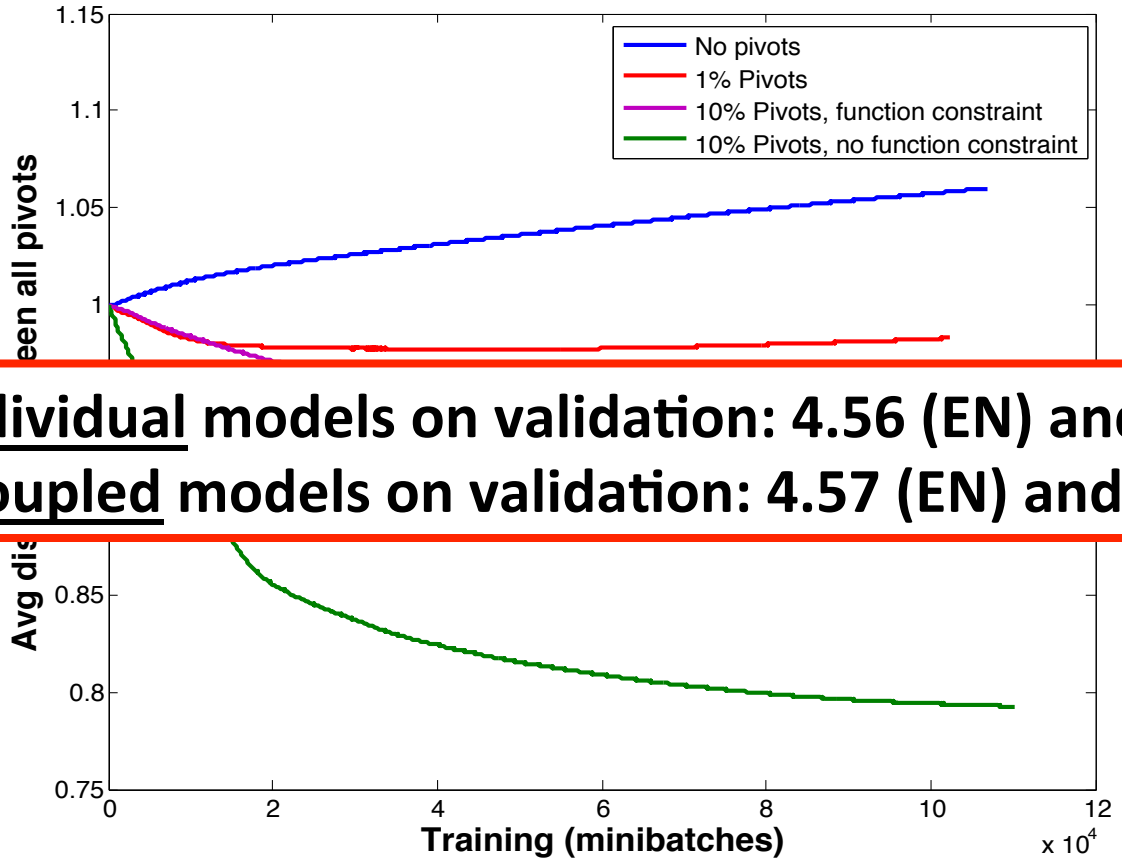
- Sampled two non-parallel datasets from English and French newswire text
- Provided models with list of *k* % translation pairs

EVALUATE: We measure the similarity of vectors for all translation pairs

UNIVERSITEIT STELLENBOSCH UNIVERSITY

mih media lab.

Experiments II: English-French



NLL of individual models on validation: 4.56 (EN) and 4.09 (FR)
NLL of coupled models on validation: 4.57 (EN) and 4.10 (FR)

UNIVERSITEIT STELLENBOSCH UNIVERSITY

mih media lab.

Future Work

- Our results indicate that even 5% pivot pairs are sufficient to start convergence.
- Previous work has shown these features to be useful for POS-tagging and NER in the *single*-domain setting [Turian et al, 2010]
- Future work will evaluate these learned features in *multi*-domain sequence-tagging tasks.

UNIVERSITEIT STELLENBOSCH UNIVERSITY

mih media lab.

1. Motivation
2. Background Theory
3. Neural Language Models
4. Other NLP Tasks
5. My Work
6. End

UNIVERSITEIT STELLENBOSCH UNIVERSITY

mih media lab.

Conclusion

- Deep learning in NLP can be seen as an adaptive framework for learning informative linguistic representations w.r.t. some task
- Great potential for incorporating stronger linguistic insight into these models
- Deep learning can be used to learn features that are invariant across domains for domain adaptation

UNIVERSITEIT STELLENBOSCH UNIVERSITY

mih media lab.

Software

- Theano (Python CPU/GPU mathematical and deep learning library
  - http://deeplearning.net/software/theano/
- Senna by (Collobert et al 2011, JMLR)
  - POS, Chunking, NER, SRL
  - State-of-the-art performance on many tasks
  - http://ronan.collobert.com/senna/
  - 3500 lines of C, extremely fast and using very little memory

- Recurrent Neural Network Language Model
  - http://www.fit.vutbr.cz/~imikolov/rnnlm/
- Recursive Neural Net and RAE models for paraphrase detection, sentiment analysis, relation classification
  - www.socher.org

UNIVERSITEIT STELLENBOSCH UNIVERSITY

mih media lab.