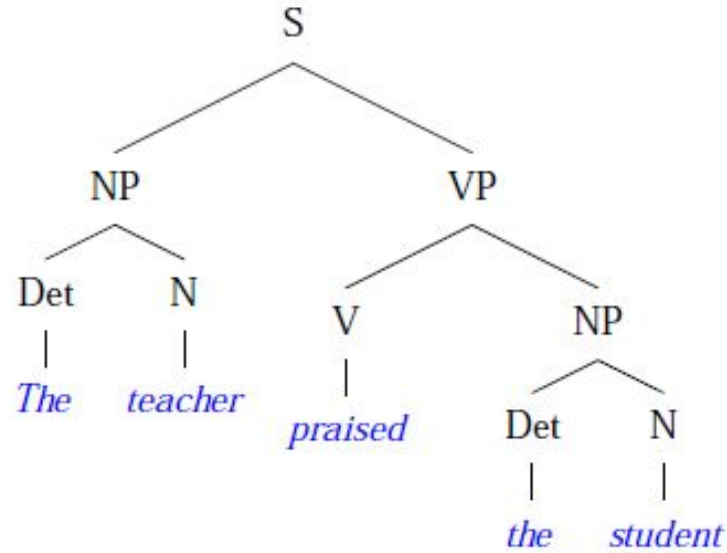# Learning Continuous Phrase Representations and Syntactic Parsing with Recursive Neural Networks
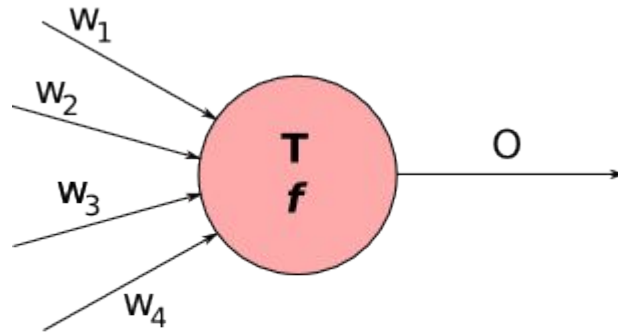
with Mattias Appelgren

The problem

Neural Networks!
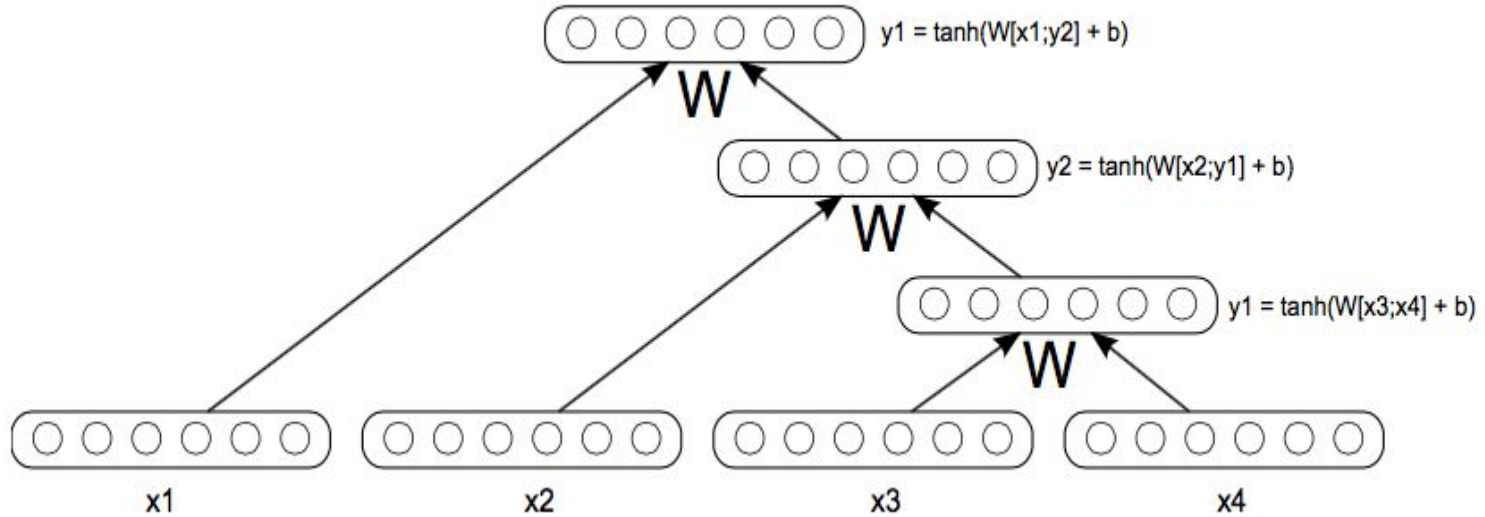

Perceptron

w_1
w_2
w_3
w_4

T
f

O

w_{1-4} Weights
T   Threshold
f   Output function
O   Output

Recursive Neural Network

$y1 = tanh(W[x1;y2] + b)$

$y2 = tanh(W[x2;y1] + b)$

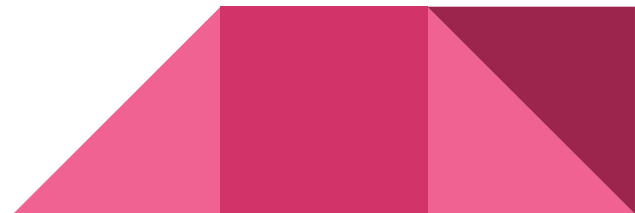$y1 = tanh(W[x3;x4] + b)$

W

W

W

x1

x2

x3

x4

Model 1: Greedy RNN

Input: x_1, x_2, x_3, x_4

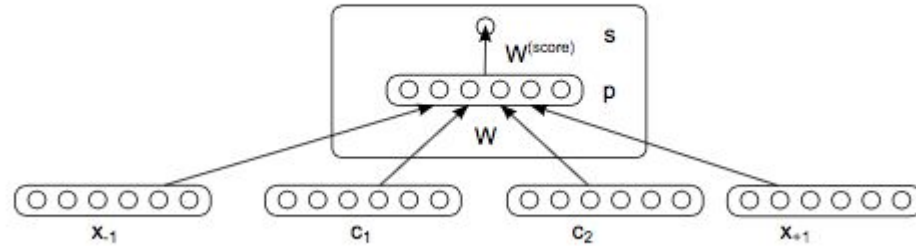$(c_1, c_2) := x_1, x_2 \rightarrow p = \tanh(W [c_1; c_2] + b)$

$s_{1,2} = W^{score} p$

Next layer: $x_1, x_2, p_{(3,4)}$

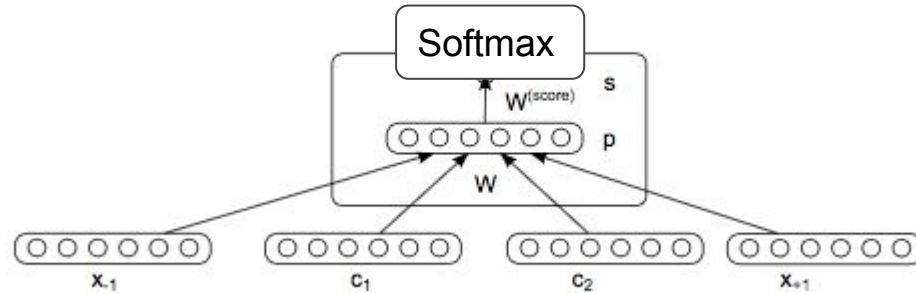Final layer: $p_{(1,(2(3,4)))} \; p_{((1,2),(3,4))}$

Model 2: Context Dependent Greedy RNN

$$s = W^{score}p \qquad (4)$$
$$p = \tanh(W[x_{-1}; c_1; c_2; x_{+1};] + b^{(1)})$$

Model 3: Greedy CRNN and Category Classifier

$$s = W^{score}p \qquad (4)$$
$$p = \tanh(W[x_{-1}; c_1; c_2; x_{+1};] + b^{(1)})$$

Model 4: Max-Margin Framework with Beam-Search

Training data: (sentence, tree) -> (x_i, y_i)

A(x_i): set of all possible trees created from a sentence

$$J = \sum_i s(x_i, y_i) - \max_{y \in A(x_i)} \left(s(x_i, y) + \Delta(y, y_i) \right)$$

$$s(x_i, y_i) = \sum_{d \in T(y_i)} s_d(c_1, c_2).$$

$$\Delta(y, y_i) = \sum_{d \in T(y)} \lambda \mathbf{1}\{d \notin T(y_i)\}.$$

Training

Look-up table trained on 611 million unlabeled words

Use words with more than 2 occurrences -> vocabulary of 15.942

Make all numbers 2

100 dimensional word representation

Results: Unlabeled bracketing on WSJ

Method	F1
Model 1 (Greedy RNN)	76.55
Model 2 (Greedy, context-aware RNN)	83.36
Model 3 (Greedy, context-aware RNN + category classifier)	87.05
Model 4 (Beam, context-aware RNN + category classifier)	92.06
Left Corner PCFG, [MC97]	90.64
Current Implementation of the Stanford Parser, [KM03]	**93.98**

More Results

POS tagging: 93.86% accuracy

Category (non-terminal) classification: 82%

On a 2.6GHz laptop our current matlab implementation needs 72 seconds to parse 421 sentences of length less than 15.
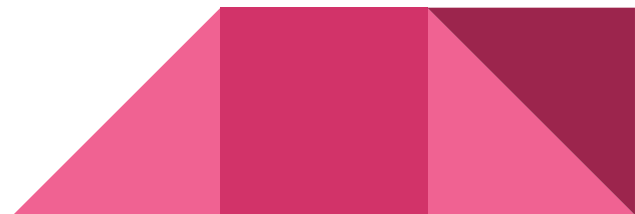
Results: Phrase Nearest Neighbours

(C) A DPC spokesman declined to elaborate on the group's new plan

1. Those two offers were private and the spokesman refused to identify the bidding companies.
2. Among the firms were Merrill Lynch & Co. and Dean Witter Reynolds Inc.
3. The real key is to have the economy working and interest rates down.
4. The market upheaval apparently has n't triggered any cash crunch yet.

(E) Columbia, S.C

1. Greenville, Miss
2. UNK, Md
3. UNK, Miss
4. UNK, Calif

(F) Fujisawa gained 22 to 2,222

1. Mochida advanced 22 to 2,222.
2. Commerzbank gained 2 to 222.2.
3. Paris loved her at first sight.
4. Profits improved across Hess 's businesses.

Any Questions?