Unsupervised Neural Hidden Markov Models

Ke Tran[1], **Yonatan Bisk**, Ashish Vaswani[2],
Daniel Marcu and Kevin Knight
USC Information Sciences Institute
[1]*Univ of Amsterdam*, [2]*Google Brain*

I am not Ke Tran

https://github.com/ketranm/neuralHMM

Bayesian Models

- HMMs, CFGs, … have been standard workhorses of the NLP community **+**
- Generative models lend themselves to unsupervised estimation **+**
- Bayesian models have elegant, but often very parametrically expensive smoothing approaches **-**

Why Neuralize Bayesian Models?

- Unsupervised structure learning +
- Simple modular extensions +
- Embeddings and vector representations have been shown to generalize well. +

This is a nice direction

Relevant EMNLP 2016 Papers:

Online Segment to Segment Neural Transduction.
Lei Yu, Jan Buys, and Phil Blunsom.

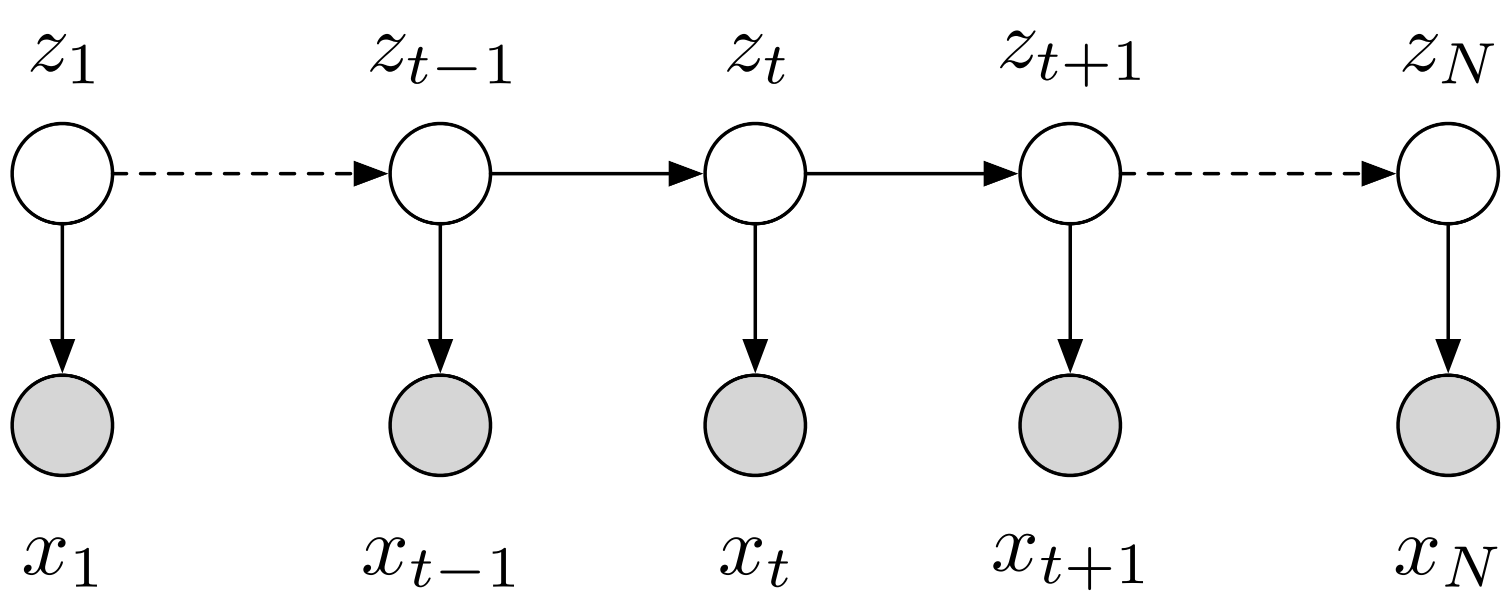
Unsupervised Neural Dependency Parsing.
Yong Jiang, Wenjuan Han, and Kewei Tu.

Hidden Markov Models

Given an observed sequence of text: x

Probability of a given token: $p(x_t|z_t) \times P(z_t|z_{t-1})$

$$p(\mathbf{x}, \mathbf{z}) = \prod_{t=1}^{n+1} p(z_t \mid z_{t-1}) \prod_{t=1}^{n} p(x_t \mid z_t)$$



Supervised POS Tagging

The	orange	man	will	lose	the	election
DT	JJ	NN	MD	VB	DT	NN

Goal: Predict the correct class for each word in the sentence

Solution: Count and divide

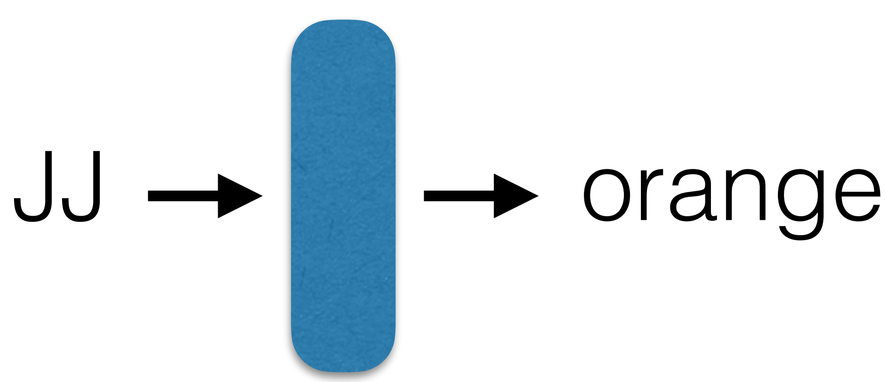
$$p(\text{orange}|\texttt{JJ}) = \frac{|\text{orange}, \texttt{JJ}|}{|\texttt{JJ}|} \qquad p(\texttt{JJ}|\texttt{DT}) = \frac{|\texttt{DT}, \texttt{JJ}|}{|\texttt{DT}|}$$

Parameters: $V \times K$ $K \times K$

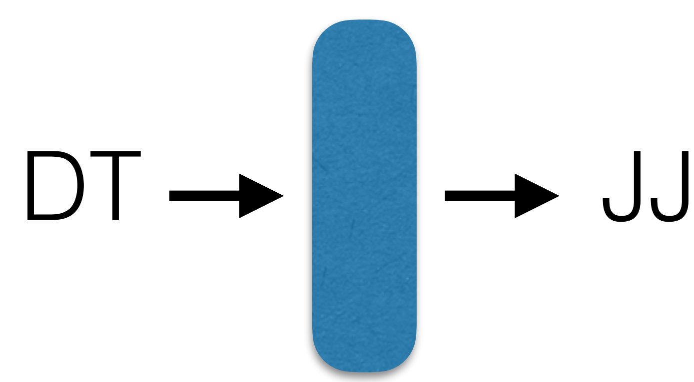
Simple Supervised Neural HMM

The	orange	man	will	lose	the	election
DT	JJ	NN	MD	VB	DT	NN

Replace parameter matrices with NNs + Softmax
Train with Cross Entropy



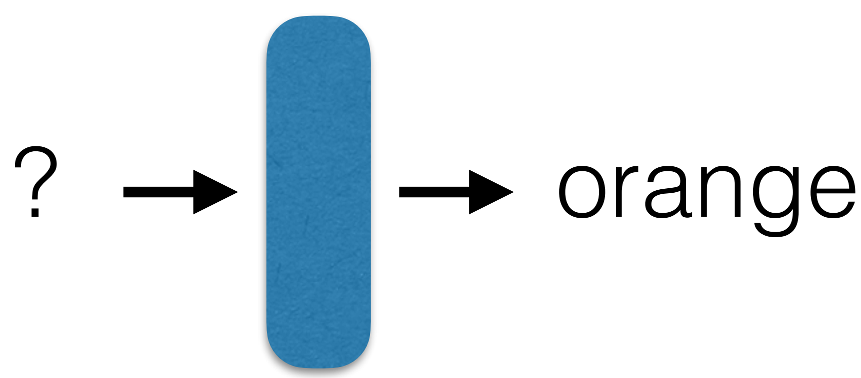
Emission Network



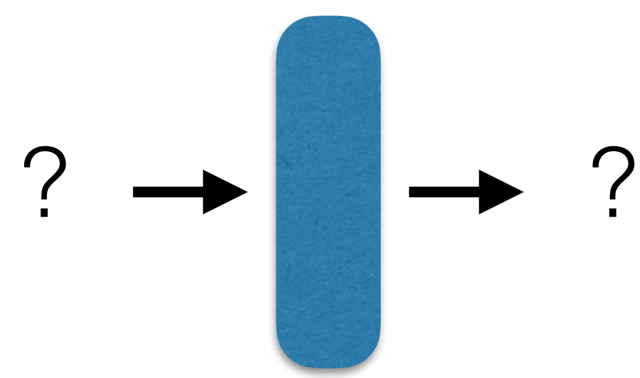
Transition Network

Unsupervised Neural HMM

The	orange	man	will	lose	the	election
?	?	?	?	?	?	?



Emission Network



Transition Network

Bayesian POS Tag Induction

The	orange	man	will	lose	the	election
C_1	C_2	C_4	C_{14}	C_{12}	C_1	C_4

Goal: Discover the set of classes which best model the observed data.

Solution: Baum-Welch

Posteriors

Probability of a specific cluster assignment

$$p(z_t = i|\mathbf{x})$$

Probability of a specific cluster transition

$$p(z_t = i, z_{t+1} = j|\mathbf{x})$$

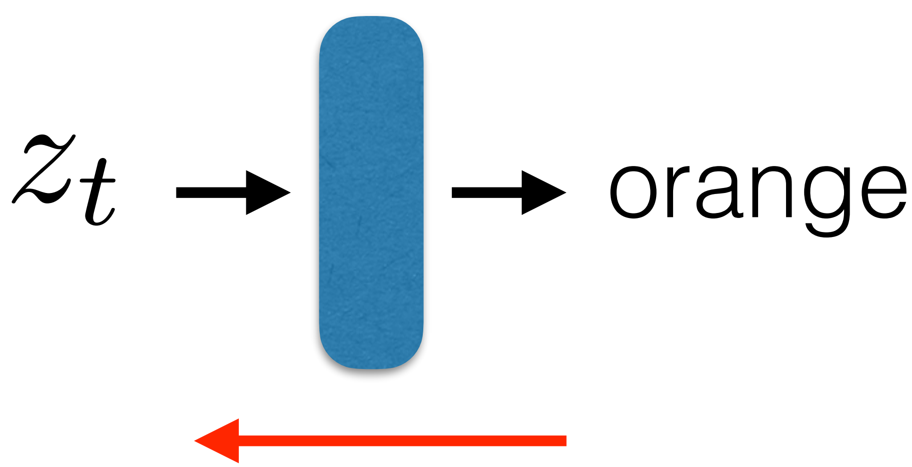
Bayesian update: Count and Divide

Count and Divide

Initialize	Compute Posteriors	Normalize
0.3	50	0.55
0.1	2	0.02
0.2	4	0.04
0.4	35	0.38
$p(w_i \mid C_j)$	$\sum_{\text{corpus}} p(w_i, C_j)$	$\hat{p}(w_i \mid C_j)$

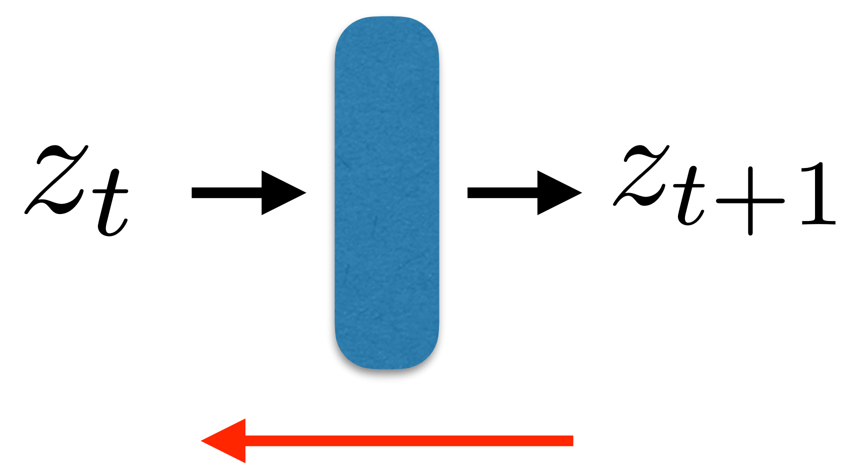
Unsupervised Neural HMM

The	orange	man	will	lose	the	election
?	?	?	?	?	?	?



$$p(z_t = i|\mathbf{x})$$

Emission Network



$$p(z_t = i, z_{t+1} = j|\mathbf{x})$$

Transition Network

Generalized EM

$$\ln p(\mathbf{x}|\theta) =$$

$$\mathbb{E}_{q(\mathbf{z})}[\ln p(\mathbf{x}, \mathbf{z}|\theta)] + \mathrm{H}[q(\mathbf{z})] + \mathrm{KL}[q(\mathbf{z})||p(\mathbf{z}|\mathbf{x}, \theta)]$$

E-Step Compute Surrogate q

M-Step Maximize Expectation

What is the gradient?

Set $q(\mathbf{z}) = p(\mathbf{z}|\mathbf{x}, \theta)$

$$\mathbb{E}_{q(\mathbf{z})}[\ln\, p(\mathbf{x}, \mathbf{z}|\theta)] + \mathrm{H}[q(\mathbf{z})] + \mathrm{KL}[q(\mathbf{z})||p(\mathbf{z}|\mathbf{x}, \theta)]$$

0

Take Derivative w.r.t. θ

$$\mathbb{E}_{q(\mathbf{z})}[\ln\, p(\mathbf{x}, \mathbf{z}|\theta)] + \mathrm{H}[q(\mathbf{z})]$$

0

$$J(\theta) = \sum_{\mathbf{z}} \underline{p(\mathbf{z}|\mathbf{x})} \frac{\partial \mathrm{ln} p(\mathbf{x}, \mathbf{z}|\theta)}{\partial \theta}$$

Jason Eisner probably
has something to say here

Initial Evaluation

Induction Metrics

- 1-1: Bijection between induced and gold classes
- M-1: Map induced class to its closest gold class
- V-M: Harmonic mean of H(c,g) and H(g,c)

Higher numbers are better

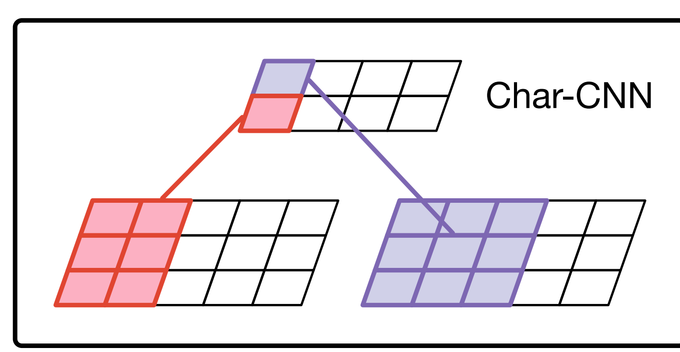
Evaluation

	1-1	M-1	V-M
HMM	41.4	**62.5**	53.3
Neural HMM	**45.7**	59.8	**54.2**

The neural model has access to no additional information

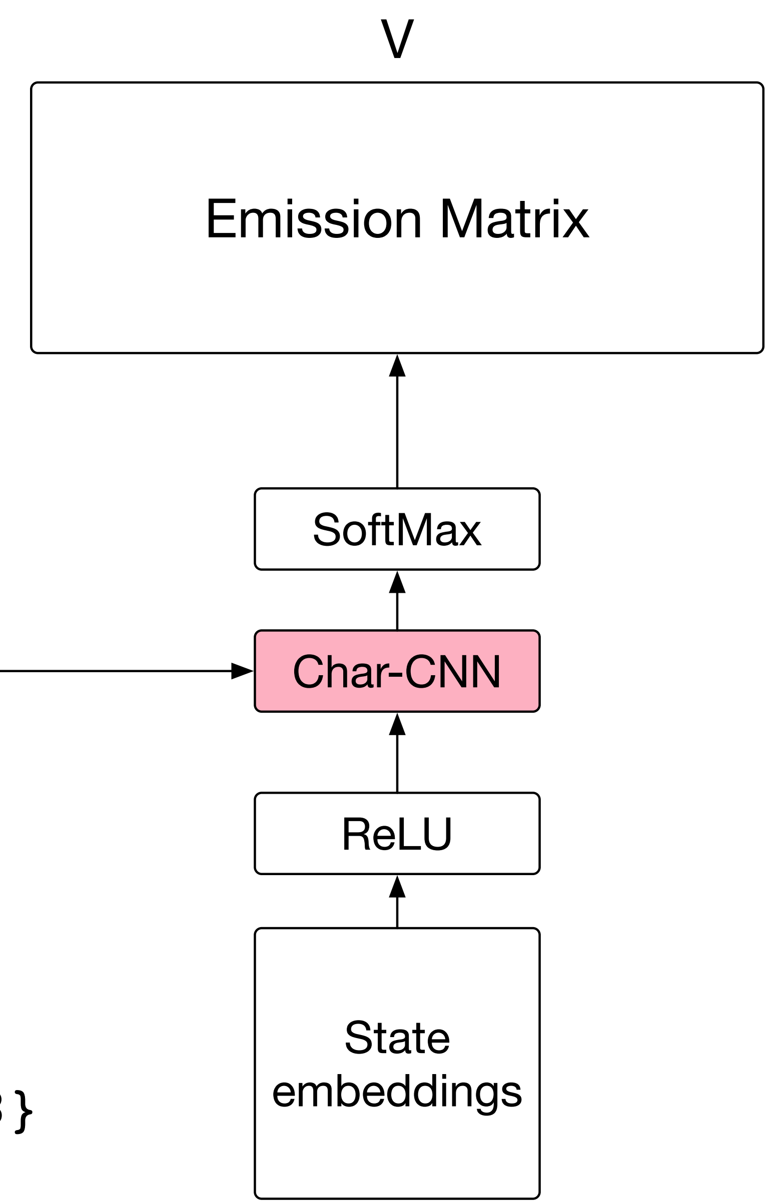
Morphology

CNN based embeddings provide morphological information




```
kernels = {1,2,3,4,5,6,7}

feature_maps =
{50, 100, 128, 128, 128, 128, 128}
```



Evaluation

	1-1	M-1	V-M
HMM	41.4	62.5	53.3
Neural HMM	45.7	59.8	54.2
+ Conv	**48.3**	**74.1**	**66.1**

Extended Context

Traditional:

Bi-gram transition	$p(z_t \mid z_{t-1})$	K^2
Tri-gram transition	$p(z_t \mid z_{t-1}, z_{t-2})$	K^3
N-gram transition	$p(z_t \mid z_{t-1}, z_{t-2}, ..., z_{t-n})$	K^{n+1}

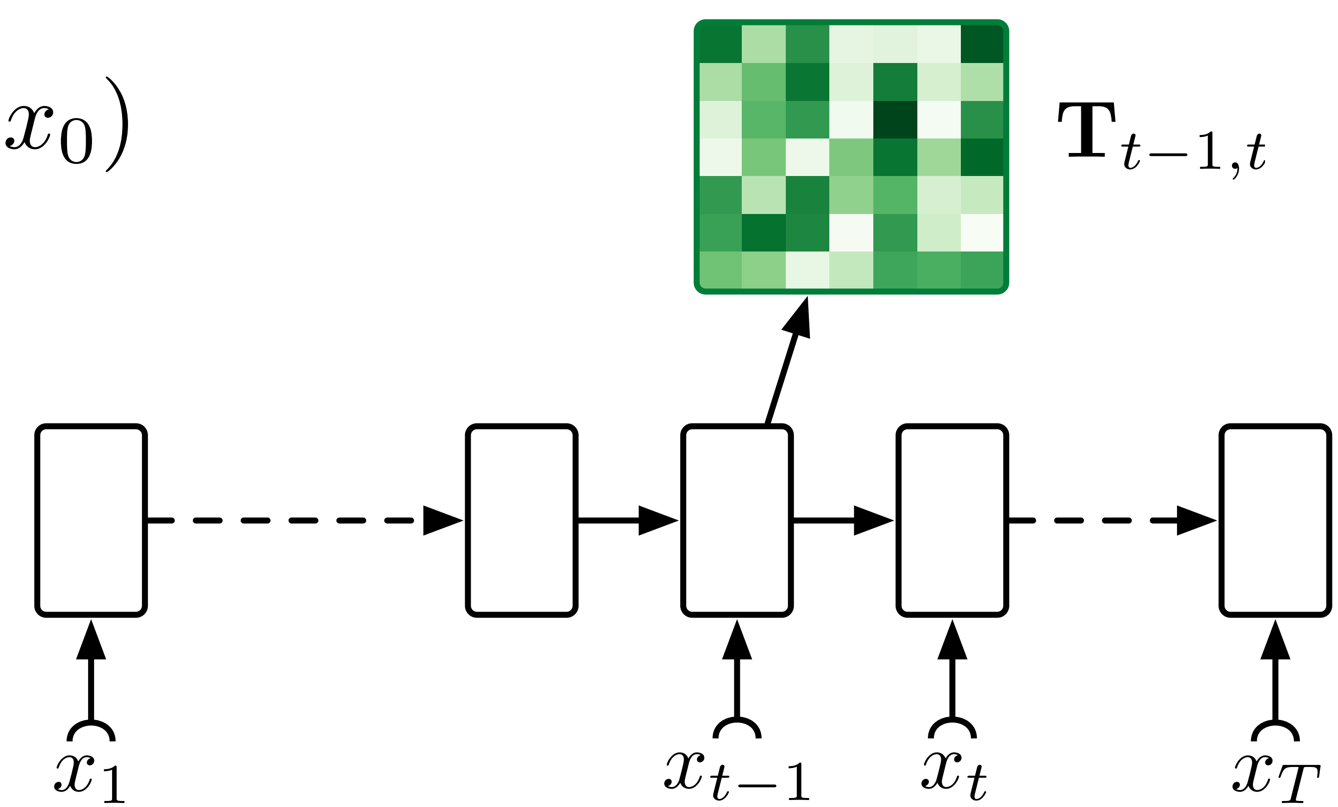
Alternative:

Previous tag and word	$p(z_t \mid z_{t-1}, x_{t-1})$	$V \times K^2$
Previous tag and sentence	$p(z_t \mid z_{t-1}, x_{t-1}, ..., x_0)$	$V^t \times K^2$

LSTM Context

LSTM consumes the sentence
and produces a transition matrix

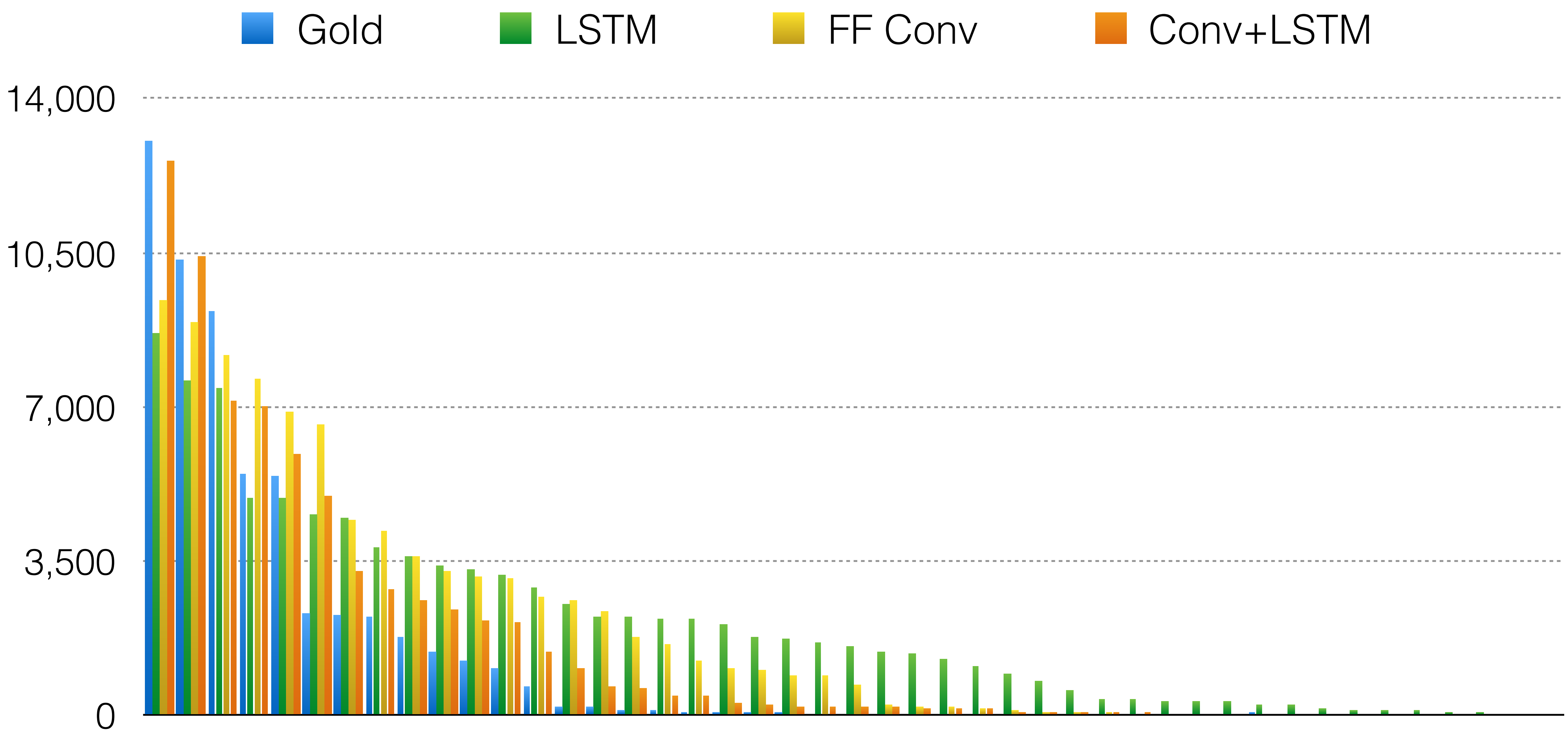
$$p(z_t | z_{t-1}, x_{t-1}, ..., x_0)$$



Evaluation

	1-1	M-1	V-M
HMM	41.4	62.5	53.3
Neural HMM	45.7	59.8	54.2
+ Conv	48.3	74.1	66.1
+ LSTM	52.4	65.1	60.4
+ Conv & LSTM	**60.7**	**79.1**	**71.7**
Blunsom 2011		77.4	69.8
Yatbaz 2012		80.2	72.1

Types / Cluster



Clusterings

Largest Cluster

LSTM	*Conv*
of	years
in	trading
to	sales
for	president
on	companies
from	prices

Numbers

LSTM	*Conv*
%	million
million	billion
year	cents
share	points
cents	point
1/2	trillion

What’s a good clustering?

	C_{15}	C_{25}
NNP	American	Corp.
	British	Inc.
	National	Co.
	Congress	Board
	Japan	Group
	San	Bank
	Federal	Inc
	West	Bush
	Dow	Department

Future Work

- Harnessing Extra Data
- Modifying the objective function
- Multilingual experiments
- Using this approach with other generative models

Thanks!

https://github.com/ketranm/neuralHMM

Parameter Initialization, Tricks, Ablation
in paper and in Github README