Reified Context Models

Jacob Steinhardt Percy Liang

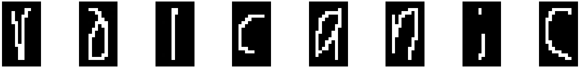
Stanford University

{jsteinhardt,pliang}@cs.stanford.edu

July 8, 2015

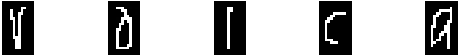
Structured Prediction Task

input x:



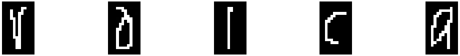
output y: v o l c a n i c

Contexts Are Key



v o l c a

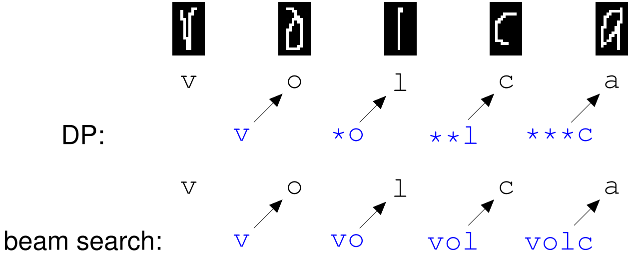
Contexts Are Key



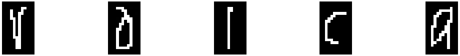
v o l c a

DP: v *o **l ***c

Contexts Are Key



Contexts Are Key



v o l c a

DP: v *o **l ***c

v o l c a

beam search: v vo vol volc

Key idea: contexts!

$$\star\mathtt{o} \stackrel{\text{def}}{=} \left\{ \begin{array}{c} \mathtt{ao} \\ \mathtt{bo} \\ \mathtt{co} \\ \vdots \end{array} \right\}$$

Desiderata

```
r   *o   **l   ***c
v   *a   **i   ***r
```

- coverage (short contexts)
  - better uncertainty estimates (precision)
  - stabler partially supervised learning updates

Desiderata

r ⋆o ⋆⋆l ⋆⋆⋆c
v ⋆a ⋆⋆i ⋆⋆⋆r

- coverage (short contexts)
  - better uncertainty estimates (precision)
  - stabler partially supervised learning updates

Desiderata

```
r  *o  **l  ***c
v  *a  **i  ***r
```

- coverage (short contexts)
  - better uncertainty estimates (precision)
  - stabler partially supervised learning updates

Desiderata

```
r   *o   **l   ***c
v   *a   **i   ***r
```

- coverage (short contexts)
  - better uncertainty estimates (precision)
  - stabler partially supervised learning updates

```
r   ro   rol   rolc
v   ra   ral   ralc
```

- expressivity (long contexts)
  - capture complex dependencies

Desiderata

```
r  ⋆o  ⋆⋆l  ⋆⋆⋆c
v  ⋆a  ⋆⋆i  ⋆⋆⋆r
```

- coverage (short contexts)
  - better uncertainty estimates (precision)
  - stabler partially supervised learning updates

```
r  ro  rol  rolc
v  ra  ral  ralc
```

- expressivity (long contexts)
  - capture complex dependencies

Desiderata

```
r  ⋆o  ⋆⋆l  ⋆⋆⋆c
v  ⋆a  ⋆⋆i  ⋆⋆⋆r
```

- coverage (short contexts)
  - better uncertainty estimates (precision)
  - stabler partially supervised learning updates

```
r  ro  rol  rolc
v  ra  ral  ralc
```

- expressivity (long contexts)
  - capture complex dependencies

Desiderata

r ⋆o ⋆⋆l ⋆⋆⋆c
v ⋆a ⋆⋆i ⋆⋆⋆r

- coverage (short contexts)
  - better uncertainty estimates (precision)
  - stabler partially supervised learning updates

r ro rol rolc
v ra ral ralc

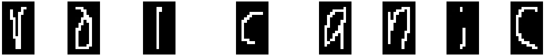
- expressivity (long contexts)
  - capture complex dependencies

r ro rol ⋆olc
v ra ral ⋆⋆⋆c
y ⋆o ⋆ol ⋆⋆⋆r
⋆ ⋆⋆ ⋆⋆⋆ ⋆⋆⋆⋆

← best of both worlds

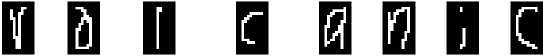
Reifying Contexts

input x:



Reifying Contexts

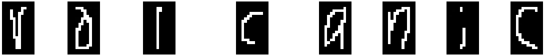
input x:



output y: v o l c a n i c

Reifying Contexts

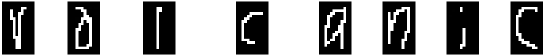
input x:



output y:	v	o	l	c	a	n	i	c
context c:	v	$\star$o	$\star$ol	$\star$olc		······		

Reifying Contexts

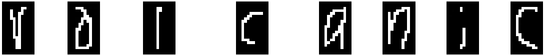
input x:



output y:	v	o	l	c	a	n	i	c
context c:	v	$\star$o	$\star$ol	$\star$olc		$\cdots\cdots$		
	r	ro	rol	$\star$olc				
	v	ra	ral	$\star\star\star$c				
	y	$\star$o	$\star$ol	$\star\star\star$r				
	$\star$	$\star\star$	$\star\star\star$	$\star\star\star\star$				

Reifying Contexts

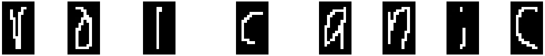
input x:



output y:	v	o	l	c	a	n	i	c	
context c:	v	⋆o	⋆ol	⋆olc	······				
	r	ro	rol	⋆olc					
	v	ra	ral	⋆⋆⋆c	←**“context sets”**				
	y	⋆o	⋆ol	⋆⋆⋆r					
	⋆	⋆⋆	⋆⋆⋆	⋆⋆⋆⋆					
	$\mathcal{C}_1$	$\mathcal{C}_2$	$\mathcal{C}_3$	$\mathcal{C}_4$					

Reifying Contexts

input x:

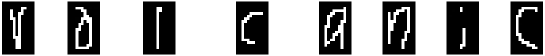


output y:	v	o	l	c	a	n	i	c	
context c:	v	⋆o	⋆ol	⋆olc		······			
	r	ro	rol	⋆olc					
	v	ra	ral	⋆⋆⋆c					←**"context sets"**
	y	⋆o	⋆ol	⋆⋆⋆r					
	⋆	⋆⋆	⋆⋆⋆	⋆⋆⋆⋆					
	$\mathcal{C}_1$	$\mathcal{C}_2$	$\mathcal{C}_3$	$\mathcal{C}_4$					

Challenge: how to trade off contexts of different lengths?

Reifying Contexts

input x:



output y:	v	o	l	c	a	n	i	c
context c:	v	⋆o	⋆ol	⋆olc	······			
	r	ro	rol	⋆olc				
	v	ra	ral	⋆⋆⋆c	←**"context sets"**			
	y	⋆o	⋆ol	⋆⋆⋆r				
	⋆	⋆⋆	⋆⋆⋆	⋆⋆⋆⋆				
	$\mathcal{C}_1$	$\mathcal{C}_2$	$\mathcal{C}_3$	$\mathcal{C}_4$				

Challenge: how to trade off contexts of different lengths?

$\Longrightarrow$ *Reify* contexts as part of model!

Reified Context Models

Given:

- context sets $\mathcal{C}_1, \ldots, \mathcal{C}_L$

Reified Context Models

Given:

- context sets $\mathcal{C}_1, \ldots, \mathcal{C}_L$
- features $\phi_i(c_{i-1}, y_i)$

Reified Context Models

Given:

- context sets $\mathcal{C}_1, \ldots, \mathcal{C}_L$
- features $\phi_i(c_{i-1}, y_i)$

Define the model

$$p_\theta(y_{1:L}, c_{1:L-1}) \propto \exp\left(\sum_{i=1}^{L} \theta^\top \phi_i(c_{i-1}, y_i)\right) \cdot \underbrace{\kappa(y, c)}_{\text{consistency}}$$

Reified Context Models

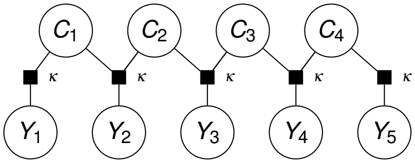
Given:

- context sets $\mathcal{C}_1, \dots, \mathcal{C}_L$
- features $\phi_i(c_{i-1}, y_i)$

Define the model

$$p_\theta(y_{1:L}, c_{1:L-1}) \propto \exp\left(\sum_{i=1}^{L} \theta^\top \phi_i(c_{i-1}, y_i)\right) \cdot \underbrace{\kappa(y, c)}_{\text{consistency}}$$

Graphical model structure:



Reified Context Models

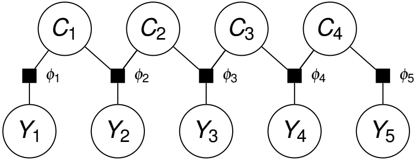
Given:

- context sets $\mathcal{C}_1, \dots, \mathcal{C}_L$
- features $\phi_i(c_{i-1}, y_i)$

Define the model

$$p_\theta(y_{1:L}, c_{1:L-1}) \propto \exp\left(\sum_{i=1}^{L} \theta^\top \phi_i(c_{i-1}, y_i)\right) \cdot \underbrace{\kappa(y, c)}_{\text{consistency}}$$

Graphical model structure:



Reified Context Models

Given:

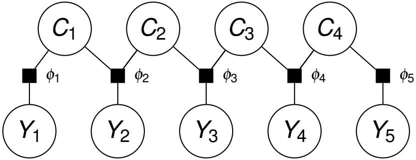
- context sets $\mathcal{C}_1, \dots, \mathcal{C}_L$
- features $\phi_i(c_{i-1}, y_i)$

Define the model

$$p_\theta(y_{1:L}, c_{1:L-1}) \propto \exp\left(\sum_{i=1}^{L} \theta^\top \phi_i(c_{i-1}, y_i)\right) \cdot \underbrace{\kappa(y, c)}_{\text{consistency}}$$

Graphical model structure:

inference via forward-backward!



Adaptive Context Selection

- Select context sets $\mathcal{C}_i$ during forward pass of inference

Adaptive Context Selection

- Select context sets $\mathcal{C}_i$ during forward pass of inference
- Greedily select contexts with largest mass

Adaptive Context Selection

- Select context sets $\mathcal{C}_i$ during forward pass of inference
- Greedily select contexts with largest mass

a
b
c
d
e
⋮

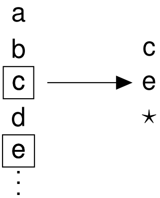
Adaptive Context Selection

- Select context sets $\mathcal{C}_i$ during forward pass of inference
- Greedily select contexts with largest mass

$$\begin{array}{c} \text{a} \\ \text{b} \\ \boxed{\text{c}} \\ \text{d} \\ \boxed{\text{e}} \\ \vdots \end{array}$$

Adaptive Context Selection

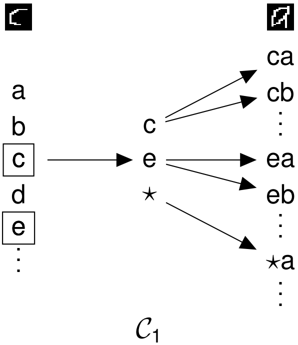
- Select context sets $\mathcal{C}_i$ during forward pass of inference
- Greedily select contexts with largest mass



$\mathcal{C}_1$

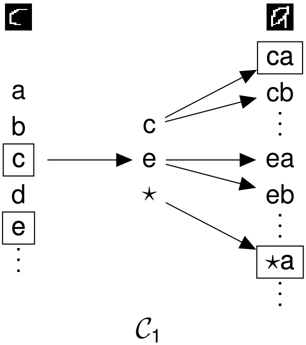
Adaptive Context Selection

- Select context sets $\mathcal{C}_i$ during forward pass of inference
- Greedily select contexts with largest mass



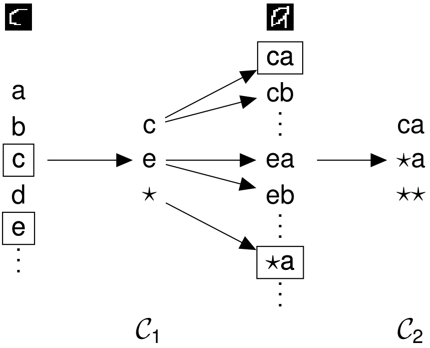
Adaptive Context Selection

- Select context sets $\mathcal{C}_i$ during forward pass of inference
- Greedily select contexts with largest mass



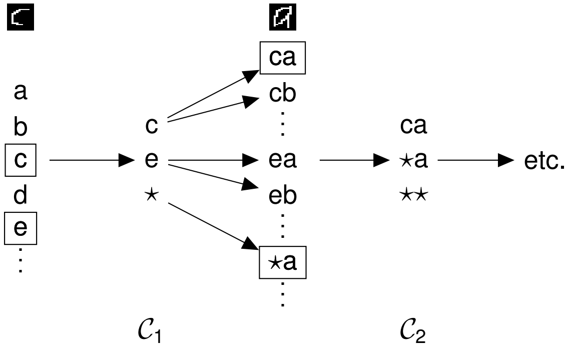
Adaptive Context Selection

- Select context sets $\mathcal{C}_i$ during forward pass of inference
- Greedily select contexts with largest mass



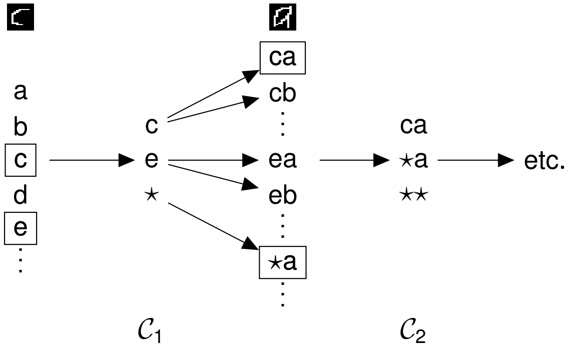
Adaptive Context Selection

- Select context sets $\mathcal{C}_i$ during forward pass of inference
- Greedily select contexts with largest mass



Adaptive Context Selection

- Select context sets $\mathcal{C}_i$ during forward pass of inference
- Greedily select contexts with largest mass



Biases towards short contexts unless there is high confidence.

Precision

input x: 

output y: v o l c a n i c

Precision

input x: 

output y: v o l c a n i c

Model assigns probability to each prediction, so can predict on most confident subset.

Precision

input x: 

output y: `v o l c a n i c`

Model assigns probability to each prediction, so can predict on most confident subset.

Measure precision (# of correct words) vs. recall (# of words predicted).

Precision

input x: 

output y: `v o l c a n i c`

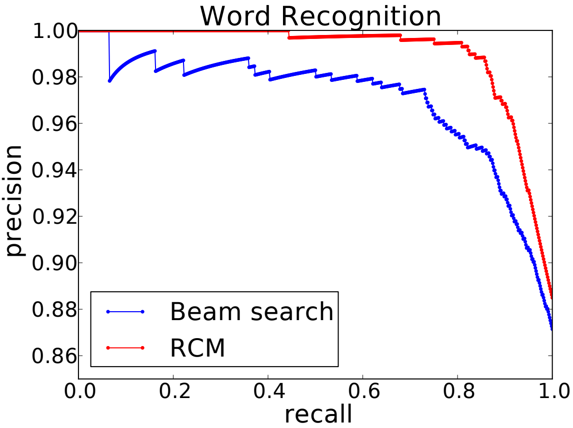
Model assigns probability to each prediction, so can predict on most confident subset.

Measure precision (# of correct words) vs. recall (# of words predicted).

- comparison: beam search

Precision

Measure precision (# of correct words) vs. recall (# of words predicted).



Partially Supervised Learning

Decipherment task:

cipher am $\mapsto$ 5, I $\mapsto$ 13, what $\mapsto$ 54, ...

Partially Supervised Learning

Decipherment task:

cipher	am $\mapsto$ 5, I $\mapsto$ 13, what $\mapsto$ 54, ...
latent z	I am what I am

Partially Supervised Learning

Decipherment task:

cipher	am $\mapsto$ 5, I $\mapsto$ 13, what $\mapsto$ 54, ...				
latent z	I	am	what	I	am
output y	13	5	54	13	5

Partially Supervised Learning

Decipherment task:

cipher	am $\mapsto$ 5, I $\mapsto$ 13, what $\mapsto$ 54, ...				
latent z	I	am	what	I	am
output y	13	5	54	13	5

Goal: determine cipher

Partially Supervised Learning

Decipherment task:

cipher	am $\mapsto$ 5, I $\mapsto$ 13, what $\mapsto$ 54, ...				
latent z	I	am	what	I	am
output y	13	5	54	13	5

Goal: determine cipher

Fit 2nd-order HMM with EM, using RCMs for approximate E-step.

Partially Supervised Learning

Decipherment task:

cipher	am $\mapsto$ 5, I $\mapsto$ 13, what $\mapsto$ 54, ...				
latent z	I	am	what	I	am
output y	13	5	54	13	5

Goal: determine cipher

Fit 2nd-order HMM with EM, using RCMs for approximate E-step.

- use learned emissions to determine cipher.

Partially Supervised Learning

Decipherment task:

cipher	am $\mapsto$ 5, I $\mapsto$ 13, what $\mapsto$ 54, ...				
latent z	I	am	what	I	am
output y	13	5	54	13	5

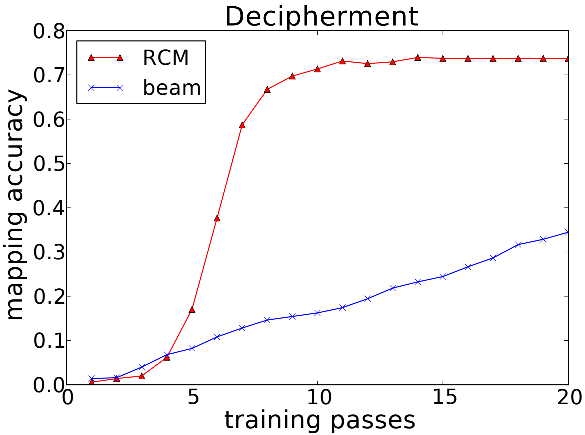
Goal: determine cipher

Fit 2nd-order HMM with EM, using RCMs for approximate E-step.

- use learned emissions to determine cipher.
- again compare to beam search (Nuhn et al., 2013)

Partially Supervised Learning

Fraction of correctly mapped words:

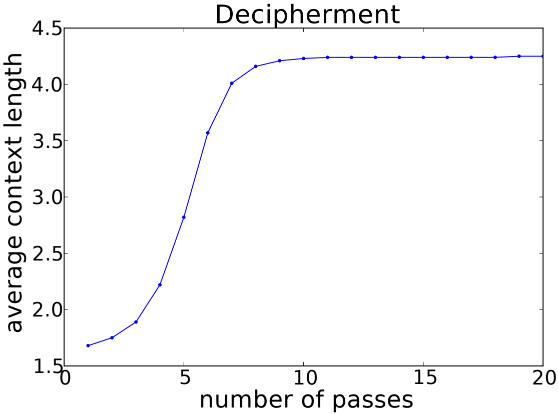


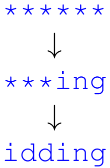
Contexts During Training

Context lengths increase smoothly during training:

Contexts During Training

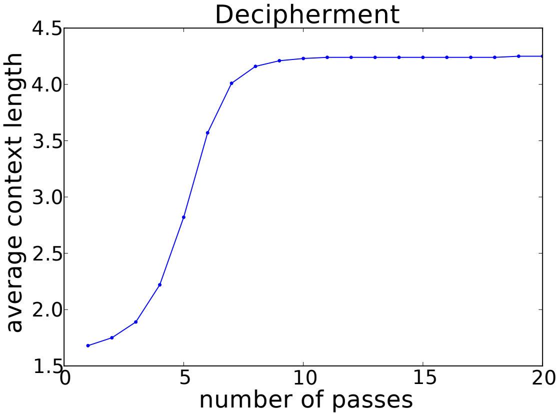
Context lengths increase smoothly during training:

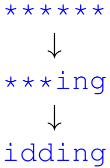




Contexts During Training

Context lengths increase smoothly during training:

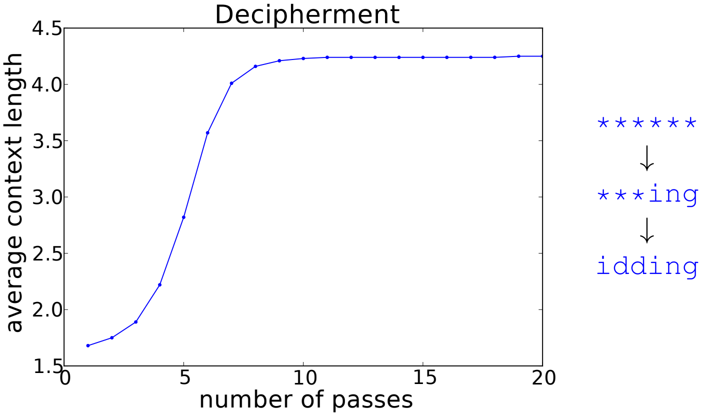




Start of training: little information, short contexts.

Contexts During Training

Context lengths increase smoothly during training:



Start of training: little information, short contexts.

End of training: lots of information, long contexts.

Discussion

RCMs provide both expressivity and coverage, which enable:

Discussion

RCMs provide both expressivity and coverage, which enable:

- More accurate uncertainty estimates (precision)

Discussion

RCMs provide both expressivity and coverage, which enable:

- More accurate uncertainty estimates (precision)
- Better partially supervised learning updates

Discussion

RCMs provide both expressivity and coverage, which enable:

- More accurate uncertainty estimates (precision)
- Better partially supervised learning updates

Related work:

- Coarse-to-fine inference (Petrov et al., 2006; Weiss et al., 2010)

Discussion

RCMs provide both expressivity and coverage, which enable:

- More accurate uncertainty estimates (precision)
- Better partially supervised learning updates

Related work:

- Coarse-to-fine inference (Petrov et al., 2006; Weiss et al., 2010)
- Certificates of optimality (Sontag, 2010)

Discussion

RCMs provide both expressivity and coverage, which enable:

- More accurate uncertainty estimates (precision)
- Better partially supervised learning updates

Related work:

- Coarse-to-fine inference (Petrov et al., 2006; Weiss et al., 2010)
- Certificates of optimality (Sontag, 2010)
- Tractable models (Poon & Domingos, 2011; Niepert & Domingos, 2014; Li & Zemel, 2014; S. & Liang, 2015)

Discussion

RCMs provide both expressivity and coverage, which enable:

- More accurate uncertainty estimates (precision)
- Better partially supervised learning updates

Related work:

- Coarse-to-fine inference (Petrov et al., 2006; Weiss et al., 2010)
- Certificates of optimality (Sontag, 2010)
- Tractable models (Poon & Domingos, 2011; Niepert & Domingos, 2014; Li & Zemel, 2014; S. & Liang, 2015)

Reproducible experiments on Codalab: `codalab.org/worksheets`