Towards dynamic computation graphs via sparse latent structure

Vlad Niculae Instituto de Telecomunicações

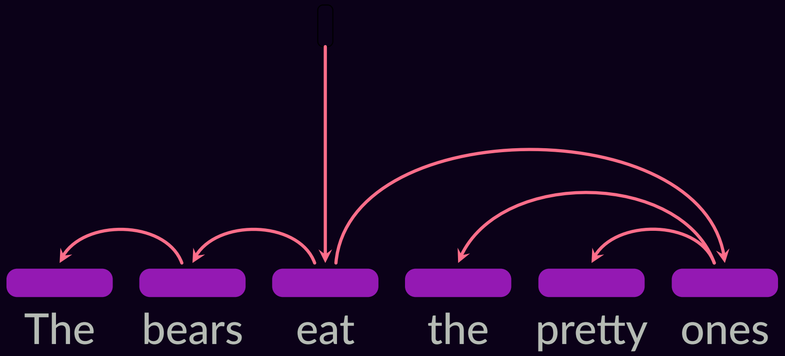
André Martins IT & Unbabel

Claire Cardie Cornell University

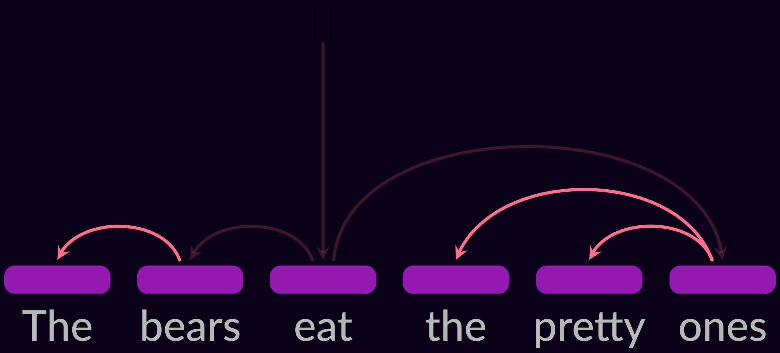
github.com/vene/sparsemap @vnfrombucharest

Dependency TreeLSTM

Dependency TreeLSTM

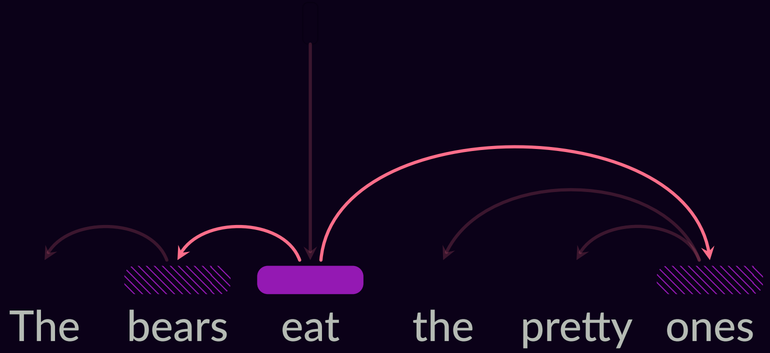


Dependency TreeLSTM

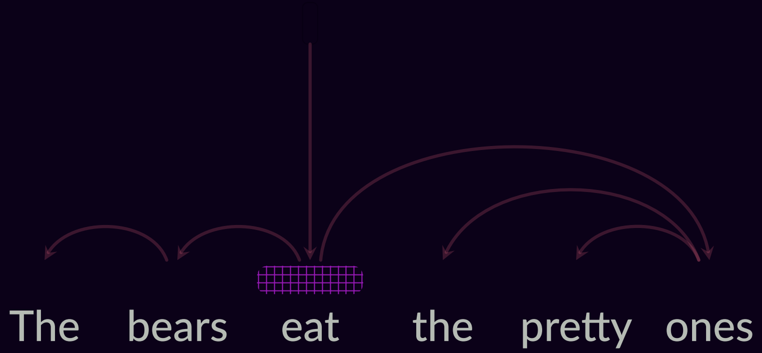


Dependency TreeLSTM

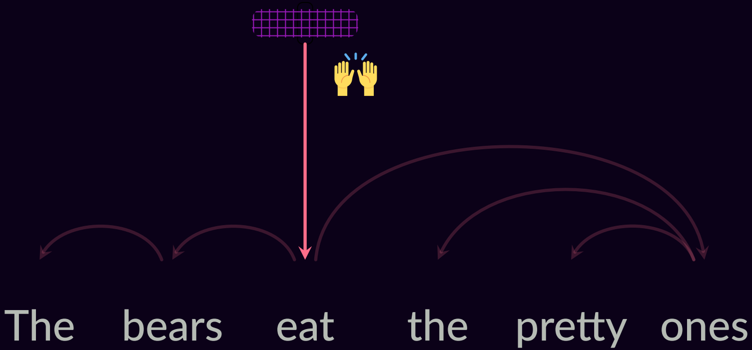
Dependency TreeLSTM



Dependency TreeLSTM



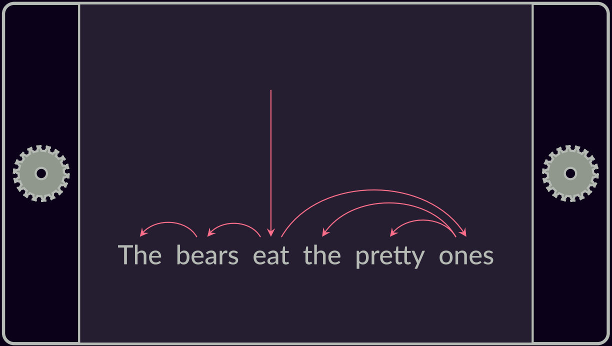
Dependency TreeLSTM



Latent Dependency TreeLSTM

input

x

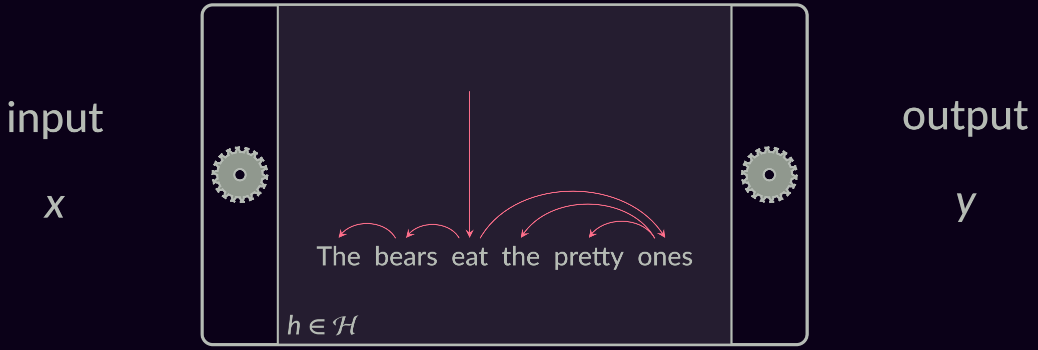


output

y

Latent Dependency TreeLSTM

$$p(y|x) = \sum_{h \in \mathcal{H}} p(y \mid h, x)\, p(h \mid x)$$



Structured Latent Variable Models

$$p(y \mid x) = \sum_{h \in \mathcal{H}} p \quad (y \mid h, x)\, p \quad (h \mid x)$$

Structured Latent Variable Models

$$p(y \mid x) = \sum_{h \in \mathcal{H}} p_{\boldsymbol{\phi}}(y \mid h, x)\, p_{\boldsymbol{\pi}}(h \mid x)$$

Structured Latent Variable Models

e.g., a TreeLSTM defined by h

$$p(y \mid x) = \sum_{h \in \mathcal{H}} p_{\boldsymbol{\phi}}(y \mid h, x)\, p_{\boldsymbol{\pi}}(h \mid x)$$

Structured Latent Variable Models

e.g., a TreeLSTM defined by h

$$p(y \mid x) = \sum_{h \in \mathcal{H}} p_{\boldsymbol{\phi}}(y \mid h, x)\, p_{\boldsymbol{\pi}}(h \mid x)$$

parsing model,
using some $\mathrm{score}_{\boldsymbol{\pi}}(h; x)$

Structured Latent Variable Models

sum over
all possible trees

e.g., a TreeLSTM defined by h

$$p(y \mid x) = \sum_{h \in \mathcal{H}} p_{\boldsymbol{\phi}}(y \mid h, x)\, p_{\boldsymbol{\pi}}(h \mid x)$$

parsing model,
using some $\text{score}_{\boldsymbol{\pi}}(h; x)$

Exponentially large sum!

Structured Latent Variable Models

sum over
all possible trees

e.g., a TreeLSTM defined by h

$$p(y \mid x) = \sum_{h \in \mathcal{H}} p_{\boldsymbol{\phi}}(y \mid h, x)\, p_{\boldsymbol{\pi}}(h \mid x)$$

parsing model,
using some $\text{score}_{\boldsymbol{\pi}}(h; x)$

How to define $\boldsymbol{p_{\pi}}$?

idea 1

idea 2

idea 3

Structured Latent Variable Models

sum over
all possible trees

e.g., a TreeLSTM defined by h

$$p(y \mid x) = \sum_{h \in \mathcal{H}} p_{\boldsymbol{\phi}}(y \mid h, x)\, p_{\boldsymbol{\pi}}(h \mid x)$$

parsing model,
using some $\text{score}_{\boldsymbol{\pi}}(h; x)$

How to define $\boldsymbol{p_{\pi}}$?

$$\sum_{h \in \mathcal{H}}$$

idea 1

idea 2

idea 3

Structured Latent Variable Models

sum over
all possible trees

e.g., a TreeLSTM defined by h

$$p(y \mid x) = \sum_{h \in \mathcal{H}} p_{\boldsymbol{\phi}}(y \mid h, x)\, p_{\boldsymbol{\pi}}(h \mid x)$$

parsing model,
using some $\text{score}_{\boldsymbol{\pi}}(h; x)$

How to define p_{π}?

$$\sum_{h \in \mathcal{H}} \frac{\partial p(y \mid x)}{\partial \boldsymbol{\pi}}$$

idea 1

idea 2

idea 3

Structured Latent Variable Models

sum over
all possible trees

e.g., a TreeLSTM defined by h

$$p(y \mid x) = \sum_{h \in \mathcal{H}} p_{\boldsymbol{\phi}}(y \mid h, x)\, p_{\boldsymbol{\pi}}(h \mid x)$$

parsing model,
using some $\text{score}_{\boldsymbol{\pi}}(h; x)$

How to define $p_{\boldsymbol{\pi}}$?

$$\sum_{h \in \mathcal{H}} \frac{\partial p(y \mid x)}{\partial \boldsymbol{\pi}}$$

idea 1 $p_{\boldsymbol{\pi}}(h \mid x) = 1$ if $h = h^\star$ else 0 argmax

idea 2

idea 3

Structured Latent Variable Models

sum over
all possible trees

e.g., a TreeLSTM defined by h

$$p(y \mid x) = \sum_{h \in \mathcal{H}} p_{\boldsymbol{\phi}}(y \mid h, x)\, p_{\boldsymbol{\pi}}(h \mid x)$$

parsing model,
using some $\text{score}_{\boldsymbol{\pi}}(h; x)$

How to define $p_{\boldsymbol{\pi}}$?

$$\sum_{h \in \mathcal{H}} \quad \frac{\partial p(y \mid x)}{\partial \boldsymbol{\pi}}$$

😊

idea 1 $\quad p_{\boldsymbol{\pi}}(h \mid x) = 1 \text{ if } h = h^\star \text{ else } 0$ $\quad$ argmax

idea 2

idea 3

Structured Latent Variable Models

sum over
all possible trees

e.g., a TreeLSTM defined by h

$$p(y \mid x) = \sum_{h \in \mathcal{H}} p_{\boldsymbol{\phi}}(y \mid h, x)\, p_{\boldsymbol{\pi}}(h \mid x)$$

parsing model,
using some $\text{score}_{\boldsymbol{\pi}}(h; x)$

How to define $p_{\boldsymbol{\pi}}$?

$$\sum_{h \in \mathcal{H}} \qquad \frac{\partial p(y \mid x)}{\partial \boldsymbol{\pi}}$$

😊 😟

idea 1 $p_{\boldsymbol{\pi}}(h \mid x) = 1$ if $h = h^\star$ else 0 argmax

idea 2

idea 3

Structured Latent Variable Models

sum over
all possible trees

e.g., a TreeLSTM defined by h

$$p(y \mid x) = \sum_{h \in \mathcal{H}} p_{\boldsymbol{\phi}}(y \mid h, x)\, p_{\boldsymbol{\pi}}(h \mid x)$$

parsing model,
using some $\text{score}_{\boldsymbol{\pi}}(h; x)$

How to define $p_{\boldsymbol{\pi}}$?

$$\sum_{h \in \mathcal{H}} \qquad \frac{\partial p(y \mid x)}{\partial \boldsymbol{\pi}}$$

idea 1	$p_{\boldsymbol{\pi}}(h \mid x) = 1$ if $h = h^\star$ else 0	argmax		
idea 2	$p_{\boldsymbol{\pi}}(h \mid x) \propto \exp\big(\text{score}_{\boldsymbol{\pi}}(h; x)\big)$	softmax		
idea 3				

Structured Latent Variable Models

sum over
all possible trees

e.g., a TreeLSTM defined by h

$$p(y \mid x) = \sum_{h \in \mathcal{H}} p_{\boldsymbol{\phi}}(y \mid h, x)\, p_{\boldsymbol{\pi}}(h \mid x)$$

parsing model,
using some $\text{score}_{\boldsymbol{\pi}}(h; x)$

How to define $p_{\boldsymbol{\pi}}$?

			$\sum_{h \in \mathcal{H}}$	$\frac{\partial p(y \mid x)}{\partial \boldsymbol{\pi}}$
idea 1	$p_{\boldsymbol{\pi}}(h \mid x) = 1$ if $h = h^\star$ else 0	argmax	😊	😟
idea 2	$p_{\boldsymbol{\pi}}(h \mid x) \propto \exp\big(\text{score}_{\boldsymbol{\pi}}(h; x)\big)$	softmax		😊
idea 3				

Structured Latent Variable Models

sum over
all possible trees

e.g., a TreeLSTM defined by h

$$p(y \mid x) = \sum_{h \in \mathcal{H}} p_{\boldsymbol{\phi}}(y \mid h, x)\, p_{\boldsymbol{\pi}}(h \mid x)$$

parsing model,
using some $\text{score}_{\boldsymbol{\pi}}(h; x)$

How to define $p_{\boldsymbol{\pi}}$?

			$\sum_{h \in \mathcal{H}}$	$\frac{\partial p(y \mid x)}{\partial \boldsymbol{\pi}}$
idea 1	$p_{\boldsymbol{\pi}}(h \mid x) = 1$ if $h = h^\star$ else 0	argmax	😊	😟
idea 2	$p_{\boldsymbol{\pi}}(h \mid x) \propto \exp\left(\text{score}_{\boldsymbol{\pi}}(h; x)\right)$	softmax	😱	😊
idea 3				

Structured Latent Variable Models

sum over
all possible trees

e.g., a TreeLSTM defined by h

$$p(y \mid x) = \sum_{h \in \mathcal{H}} p_{\boldsymbol{\phi}}(y \mid h, x)\, p_{\boldsymbol{\pi}}(h \mid x)$$

parsing model,
using some $\text{score}_{\boldsymbol{\pi}}(h; x)$

How to define $p_{\boldsymbol{\pi}}$?

			$\sum_{h \in \mathcal{H}}$	$\frac{\partial p(y \mid x)}{\partial \boldsymbol{\pi}}$
idea 1	$p_{\boldsymbol{\pi}}(h \mid x) = 1$ if $h = h^\star$ else 0	argmax	😊	😟
idea 2	$p_{\boldsymbol{\pi}}(h \mid x) \propto \exp\left(\text{score}_{\boldsymbol{\pi}}(h; x)\right)$	softmax	😱	😊
idea 3		SparseMAP	😊	😊

SparseMAP Inference

(Niculae et al, ICML 2018)

 = .7  + .3 

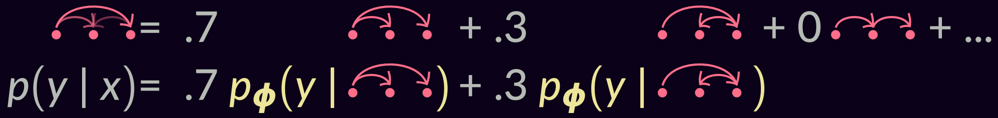
SparseMAP Inference

(Niculae et al, ICML 2018)

 $= .7$  $+ .3$  $+ 0$  $+ \ldots$

SparseMAP Inference

(Niculae et al, ICML 2018)



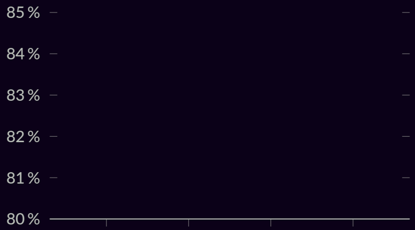
SparseMAP Inference

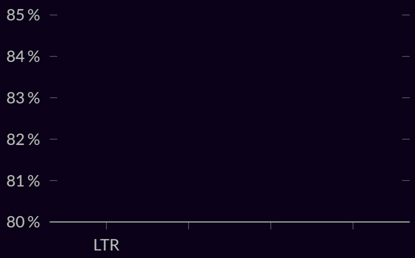
(Niculae et al, ICML 2018)

 $= .7$  $+ .3$  $+ 0$  $+ \dots$

$p(y \mid x) = .7\, p_{\phi}(y \mid$  $) + .3\, p_{\phi}(y \mid$  $)$

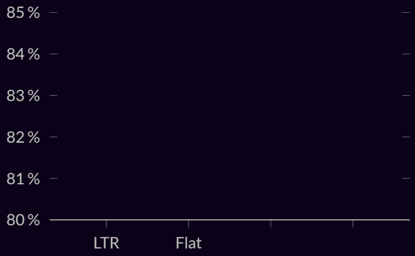
 is not a tree itself: $p(y \mid x) \neq p_{\phi}(y \mid$  $)$!

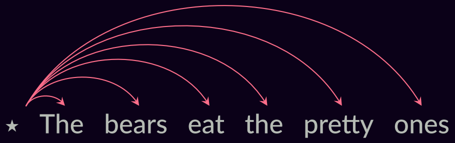

85 %
84 %
83 %
82 %
81 %
80 %



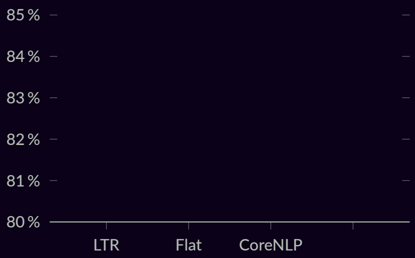
★ The bears eat the pretty ones

Left-to-right: regular LSTM

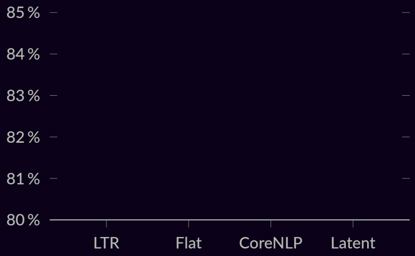


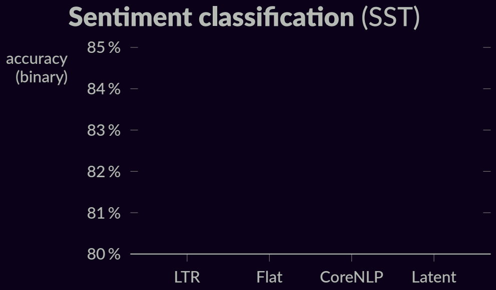


Flat: bag-of-words–like

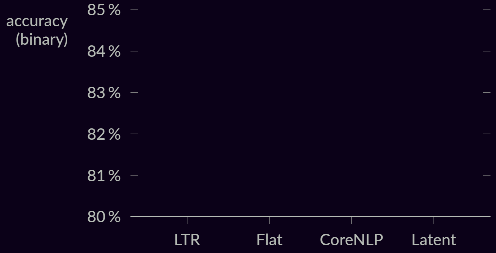


CoreNLP: off-line parser

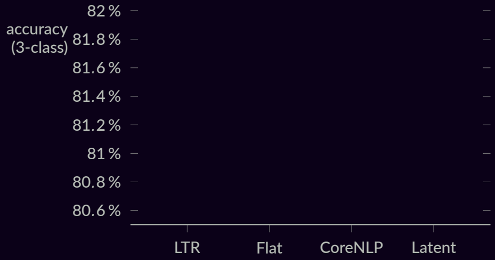

85 %
84 %
83 %
82 %
81 %
80 %
LTR
Flat
CoreNLP
Latent


Sentiment classification (SST)
accuracy
(binary)
85 %
84 %
83 %
82 %
81 %
80 %
LTR
Flat
CoreNLP
Latent

Sentiment classification (SST)



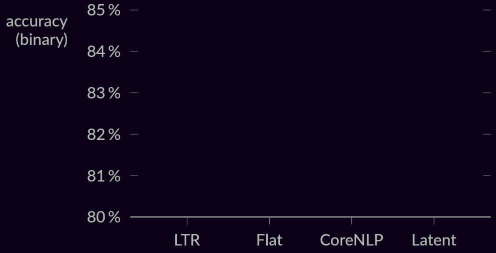
Natural Language Inference (SNLI)



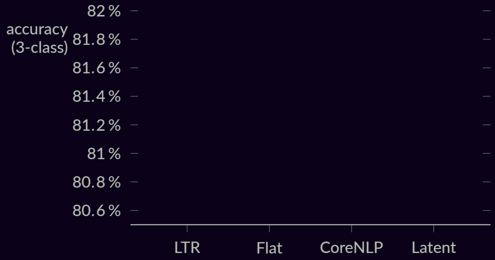
Sentence pair classification (P, H)

$$p(y \mid P, H) = \sum_{h_P \in \mathcal{H}(P)} \sum_{h_H \in \mathcal{H}(H)} p_{\boldsymbol{\phi}}(y \mid h_P, h_H)\, p_{\boldsymbol{\pi}}(h_P \mid P)\, p_{\boldsymbol{\pi}}(h_H \mid H)$$

Sentiment classification (SST)



Natural Language Inference (SNLI)

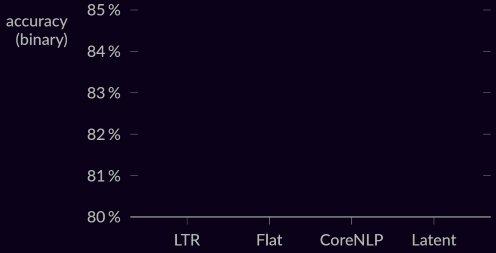


Reverse dictionary lookup

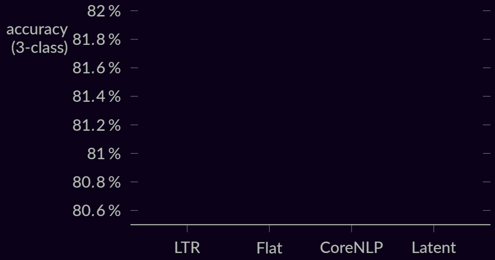
given word description, predict word embedding (Hill et al, 17)

instead of $p(y \mid x)$, we model $\mathbb{E}_{p_{\boldsymbol{\pi}}} \boldsymbol{g}(x) = \sum_{h \in \mathcal{H}} \boldsymbol{g}(x; h)\, p_{\boldsymbol{\pi}}(h \mid x)$

Sentiment classification (SST)

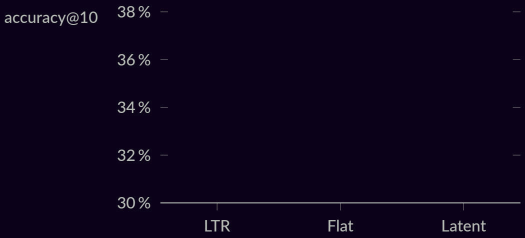


Natural Language Inference (SNLI)

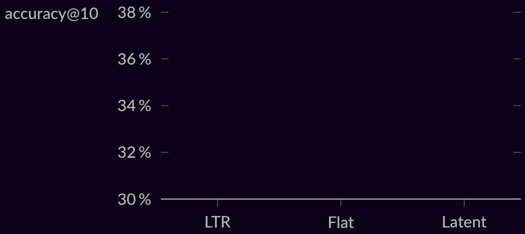


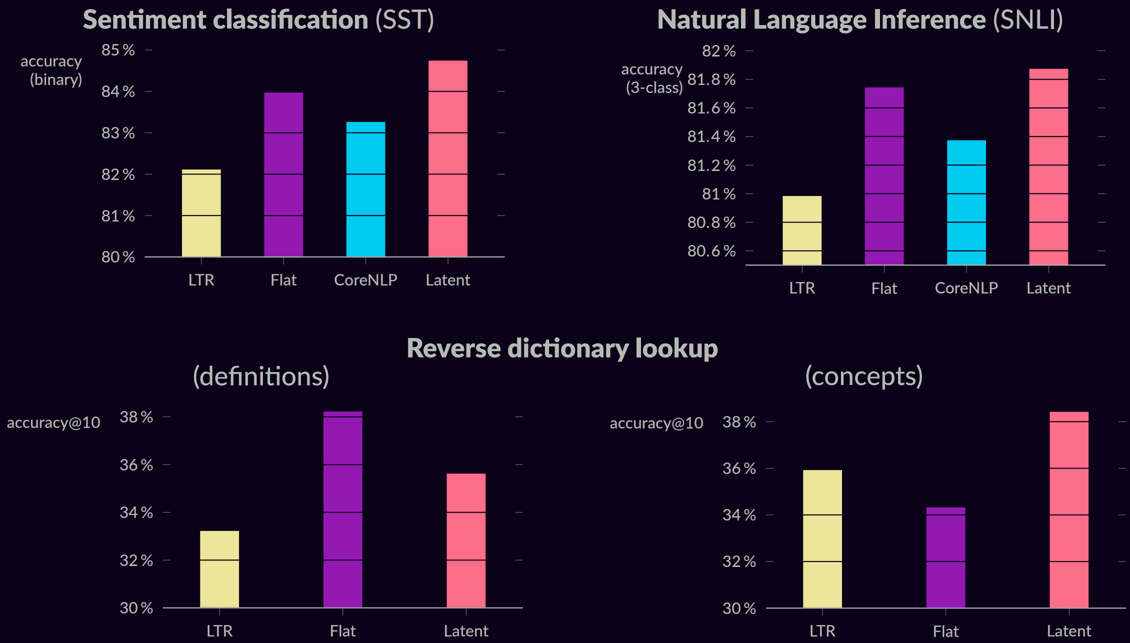
Reverse dictionary lookup

(definitions)



(concepts)




Sentiment classification (SST)
accuracy (binary)
85 %
84 %
83 %
82 %
81 %
80 %
LTR
Flat
CoreNLP
Latent
Natural Language Inference (SNLI)
accuracy (3-class)
82 %
81.8 %
81.6 %
81.4 %
81.2 %
81 %
80.8 %
80.6 %
LTR
Flat
CoreNLP
Latent
Reverse dictionary lookup
(definitions)
accuracy@10
38 %
36 %
34 %
32 %
30 %
LTR
Flat
Latent
(concepts)
accuracy@10
38 %
36 %
34 %
32 %
30 %
LTR
Flat
Latent

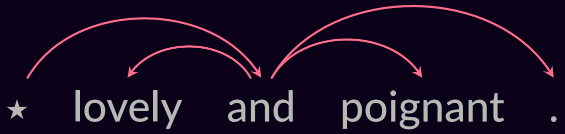
Syntax vs. Composition Order

CoreNLP parse, $p = 21.4\%$

★ lovely and poignant .

Syntax vs. Composition Order

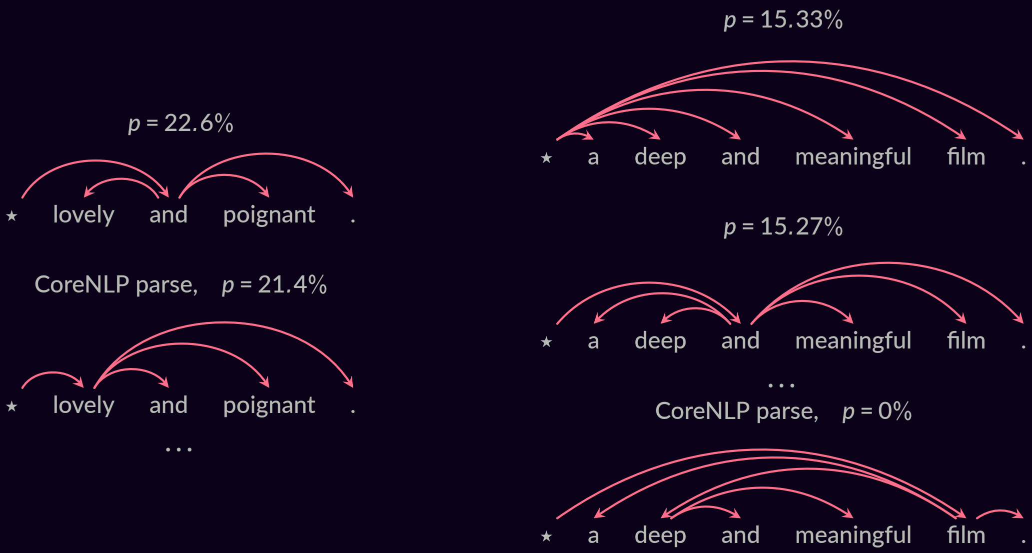
$p = 22.6\%$



CoreNLP parse, $p = 21.4\%$

. . .

Syntax vs. Composition Order

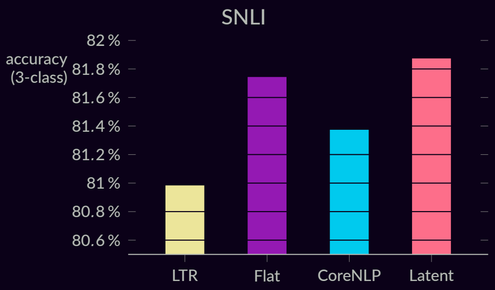


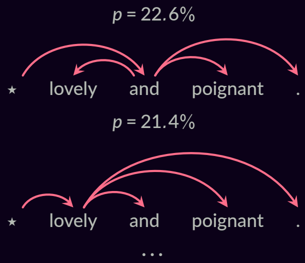
Conclusions

Latent structured variables for uncertainty & compositionality

Tractable marginalization via SparseMAP inference

Flexible model: arbitrary function of discrete latent structures





vlad@vene.ro
https://vene.ro
github.com/vene/sparsemap
@vnfrombucharest

Some icons by Dave Gandy and Freepik via flaticon.com.