




Towards **dynamic computation graphs** via **sparse latent structure**

Vlad Niculae Instituto de Telecomunicações

André Martins IT & Unbabel

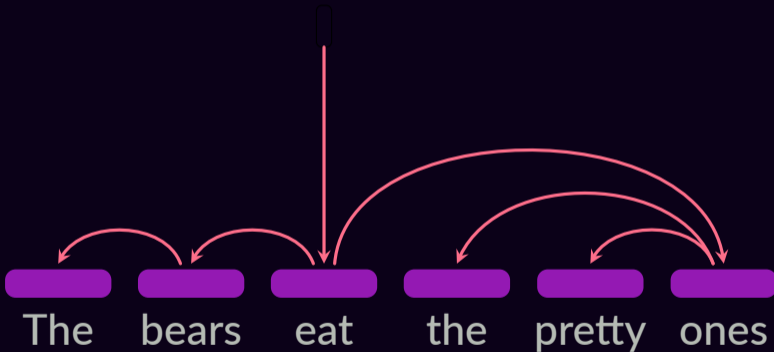
Claire Cardie Cornell University

Dependency TreeLSTM

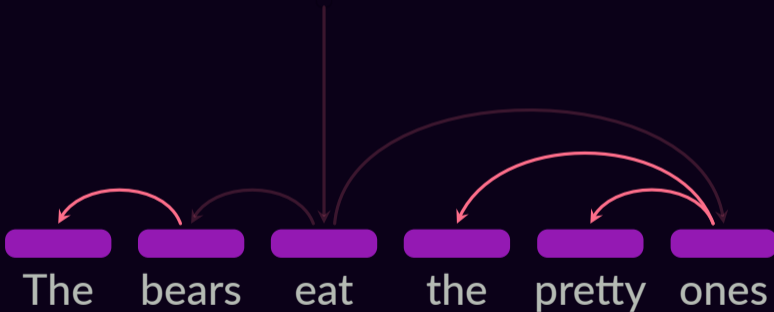


The bears eat the pretty ones

Dependency TreeLSTM



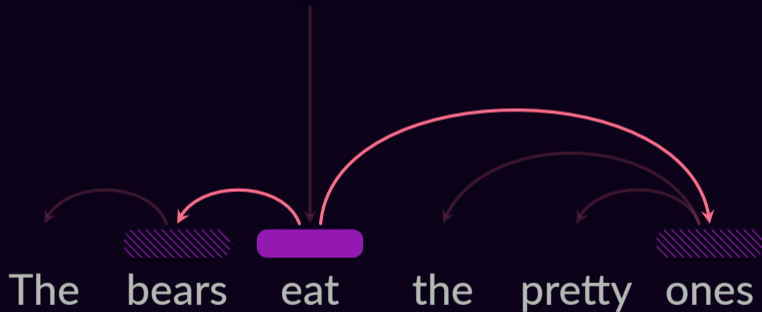
Dependency TreeLSTM



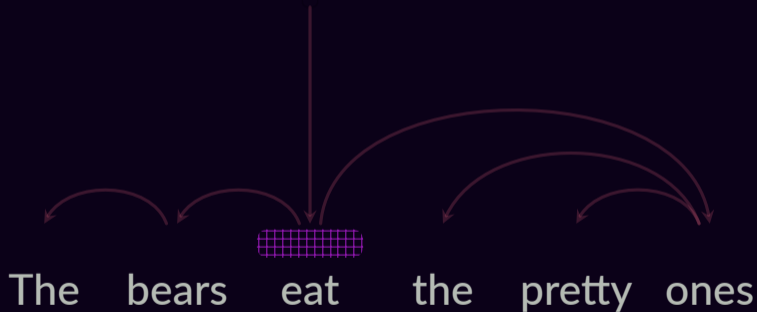
Dependency TreeLSTM



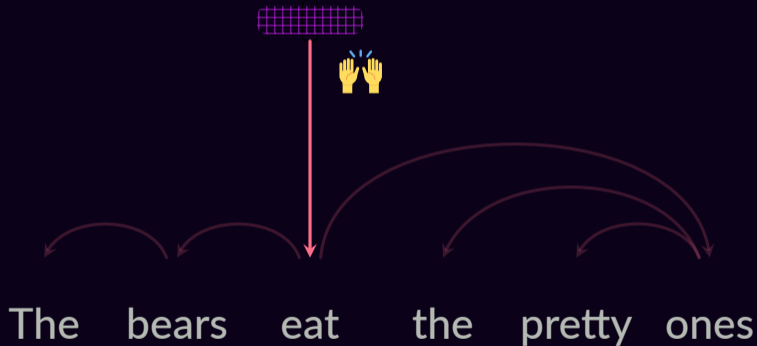
Dependency TreeLSTM



Dependency TreeLSTM



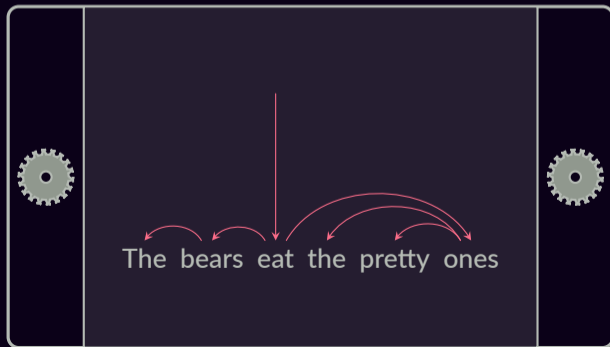
Dependency TreeLSTM



Latent Dependency TreeLSTM

input

x

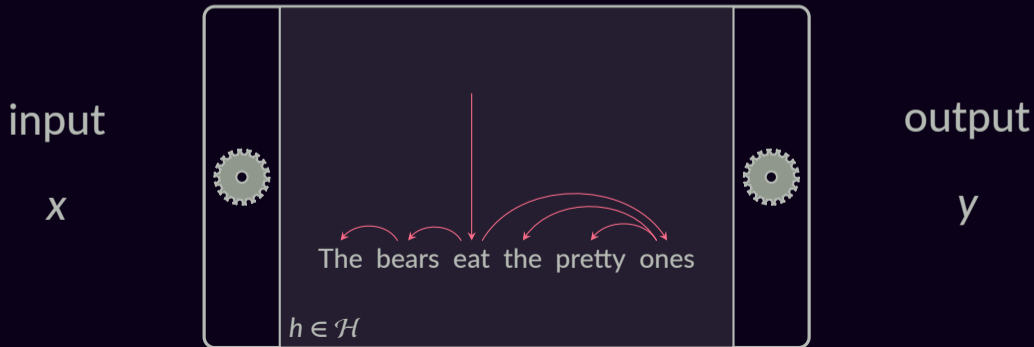


output

y

Latent Dependency TreeLSTM

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



Structured Latent Variable Models

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

Structured Latent Variable Models

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

Structured Latent Variable Models

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

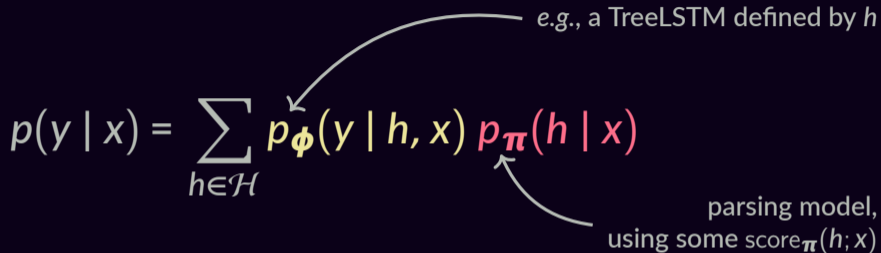
e.g., a TreeLSTM defined by h

Structured Latent Variable Models

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

e.g., a TreeLSTM defined by h

parsing model,
using some score $\pi(h; x)$

The diagram features the equation $p(y | x) = \sum_{h \in \mathcal{H}} p_{\phi}(y | h, x) p_{\pi}(h | x)$ centered on a black background. The term $p_{\phi}(y | h, x)$ is highlighted in yellow, and $p_{\pi}(h | x)$ is highlighted in red. A white arrow originates from the text "e.g., a TreeLSTM defined by h " and points to the ϕ parameter in the yellow term. Another white arrow originates from the text "parsing model, using some score $\pi(h; x)$ " and points to the π parameter in the red term.

Structured Latent Variable Models

sum over
all possible trees

e.g., a TreeLSTM defined by h

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

parsing model,
using some score $\pi(h; x)$

Exponentially large sum!

Structured Latent Variable Models

sum over
all possible trees

e.g., a TreeLSTM defined by h

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

parsing model,
using some score $\pi(h; x)$

How to define p_{π} ?

idea 1

idea 2

idea 3

Structured Latent Variable Models

sum over
all possible trees

e.g., a TreeLSTM defined by h

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

parsing model,
using some score $\pi(h; x)$

How to define p_{π} ?

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

idea 1

idea 2

idea 3

Structured Latent Variable Models

sum over
all possible trees

e.g., a TreeLSTM defined by h

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

parsing model,
using some score $\pi(h; x)$

How to define p_{π} ?

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

idea 1

idea 2

idea 3

Structured Latent Variable Models

sum over
all possible trees

e.g., a TreeLSTM defined by h

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

parsing model,
using some score $\pi(h; x)$

How to define p_{π} ?

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

idea 1 $p_{\pi}(h | x) = 1$ if $h = h^*$ else 0 argmax

idea 2

idea 3

Structured Latent Variable Models

sum over
all possible trees

e.g., a TreeLSTM defined by h

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

parsing model,
using some score $\pi(h; x)$

How to define p_{π} ?

idea 1 $p_{\pi}(h | x) = 1$ if $h = h^*$ else 0

argmax

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



idea 2

idea 3

Structured Latent Variable Models

sum over
all possible trees

e.g., a TreeLSTM defined by h

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

parsing model,
using some score $\pi(h; x)$

How to define p_{π} ?

idea 1 $p_{\pi}(h | x) = 1$ if $h = h^*$ else 0

argmax

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



idea 2

idea 3

Structured Latent Variable Models

sum over
all possible trees

e.g., a TreeLSTM defined by h

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

parsing model,
using some score $\pi(h; x)$

How to define p_{π} ?

idea 1 $p_{\pi}(h | x) = 1$ if $h = h^*$ else 0

argmax

idea 2 $p_{\pi}(h | x) \propto \exp(\text{score}_{\pi}(h; x))$

softmax

idea 3

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

😊 ☹️

Structured Latent Variable Models

sum over
all possible trees

e.g., a TreeLSTM defined by h

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

parsing model,
using some score $\pi(h; x)$

How to define p_{π} ?

idea 1 $p_{\pi}(h | x) = 1$ if $h = h^*$ else 0

argmax

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



idea 2 $p_{\pi}(h | x) \propto \exp(\text{score}_{\pi}(h; x))$

softmax



idea 3

Structured Latent Variable Models

sum over
all possible trees

e.g., a TreeLSTM defined by h

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

parsing model,
using some score $\pi(h; x)$

How to define p_{π} ?

idea 1 $p_{\pi}(h | x) = 1$ if $h = h^*$ else 0

argmax

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



$\frac{\partial p(y | x)}{\partial \pi}$

$\partial \pi$



idea 2 $p_{\pi}(h | x) \propto \exp(\text{score}_{\pi}(h; x))$

softmax



idea 3

Structured Latent Variable Models

sum over
all possible trees

e.g., a TreeLSTM defined by h

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

parsing model,
using some score $\pi(h; x)$

How to define p_{π} ?

idea 1 $p_{\pi}(h | x) = 1$ if $h = h^*$ else 0

argmax

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



$\frac{\partial p(y | x)}{\partial \pi}$



idea 2 $p_{\pi}(h | x) \propto \exp(\text{score}_{\pi}(h; x))$

softmax



idea 3

SparseMAP



SparseMAP Inference

(Niculae et al, ICML 2018)



SparseMAP Inference

(Niculae et al, ICML 2018)

$$\text{Diagram} = .7$$

$$\text{Diagram} + .3$$

$$\text{Diagram} + 0 \cdot \text{Diagram} + \dots$$

SparseMAP Inference

(Niculae et al, ICML 2018)

$$\begin{matrix} \text{•} & \text{•} & \text{•} \\ \text{•} & \text{•} & \text{•} \end{matrix} = .7 \quad \begin{matrix} \text{•} & \text{•} & \text{•} \\ \text{•} & \text{•} & \text{•} \end{matrix} + .3 \quad \begin{matrix} \text{•} & \text{•} & \text{•} \\ \text{•} & \text{•} & \text{•} \end{matrix} + 0 \begin{matrix} \text{•} & \text{•} & \text{•} \\ \text{•} & \text{•} & \text{•} \end{matrix} + \dots$$

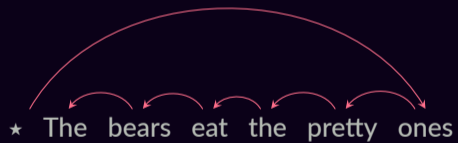
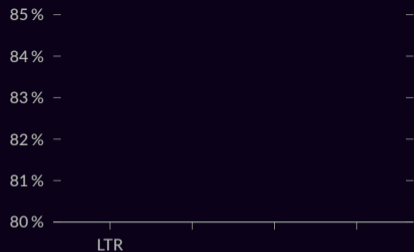
$$p(y | x) = .7 p_{\phi}(y | \begin{matrix} \text{•} & \text{•} & \text{•} \\ \text{•} & \text{•} & \text{•} \end{matrix}) + .3 p_{\phi}(y | \begin{matrix} \text{•} & \text{•} & \text{•} \\ \text{•} & \text{•} & \text{•} \end{matrix})$$

SparseMAP Inference

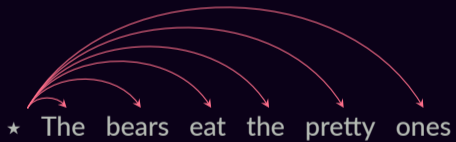
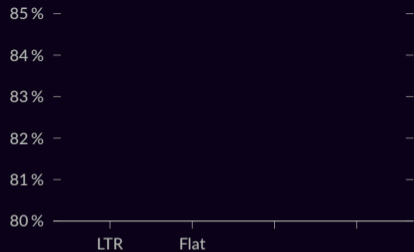
(Niculae et al, ICML 2018)

$$\begin{aligned} & \text{Diagram 1} = .7 \quad \text{Diagram 2} + .3 \quad \text{Diagram 3} + 0 \cdot \text{Diagram 4} + \dots \\ p(y | x) = & .7 p_{\phi}(y | \text{Diagram 1}) + .3 p_{\phi}(y | \text{Diagram 2}) \end{aligned}$$

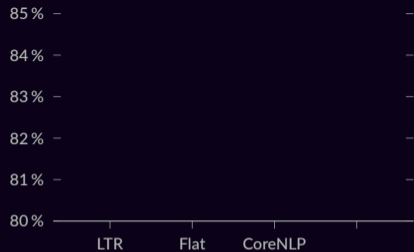
Diagram 1 is not a tree itself: $p(y | x) \neq p_{\phi}(y | \text{Diagram 1})!$



Left-to-right: regular LSTM



Flat: bag-of-words-like

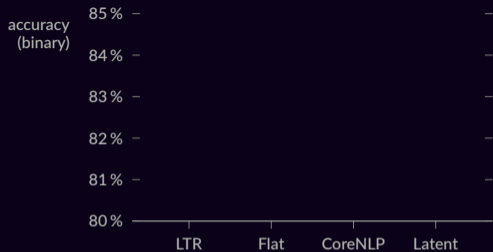


★ The bears eat the pretty ones

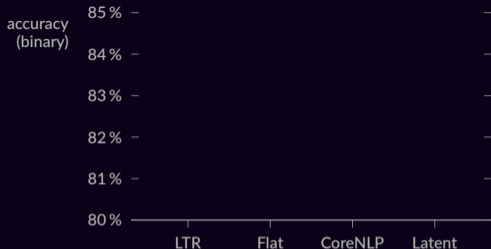
Diagram illustrating the dependency arcs for the sentence "The bears eat the pretty ones". The arcs are shown as red curved arrows connecting the words. The arcs are: "The" to "bears", "bears" to "eat", "eat" to "the", "eat" to "pretty", "eat" to "ones", "the" to "pretty", and "the" to "ones".

CoreNLP: off-line parser

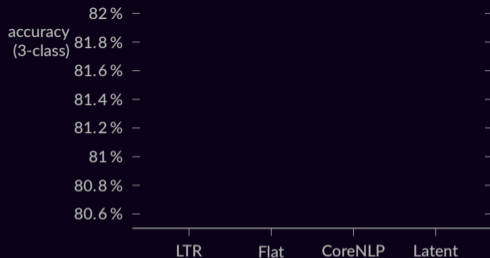
Sentiment classification (SST)



Sentiment classification (SST)



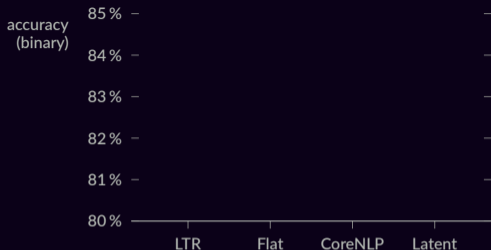
Natural Language Inference (SNLI)



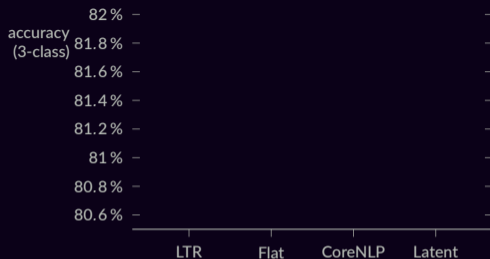
Sentence pair classification (P, H)

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

Sentiment classification (SST)



Natural Language Inference (SNLI)

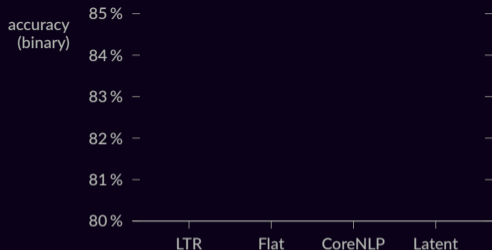


Reverse dictionary lookup

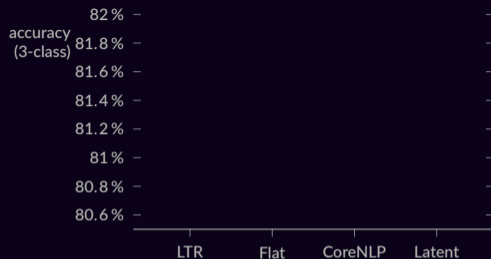
given word description, predict word embedding (Hill et al, 17)

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

Sentiment classification (SST)

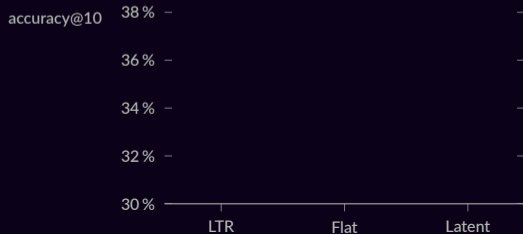


Natural Language Inference (SNLI)

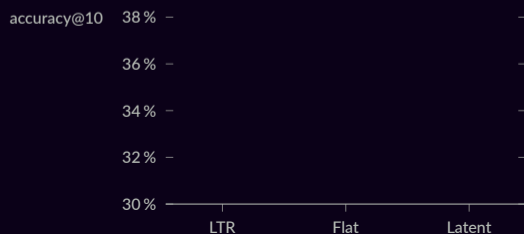


Reverse dictionary lookup

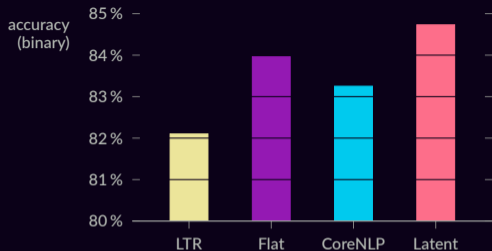
(definitions)



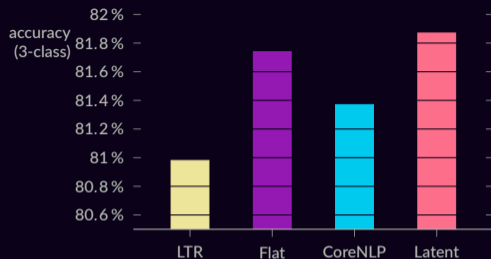
(concepts)



Sentiment classification (SST)

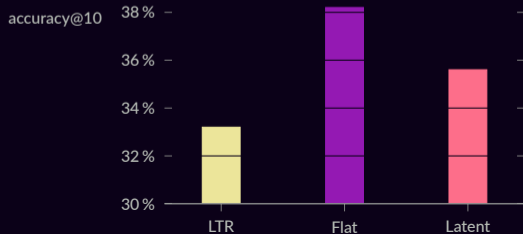


Natural Language Inference (SNLI)

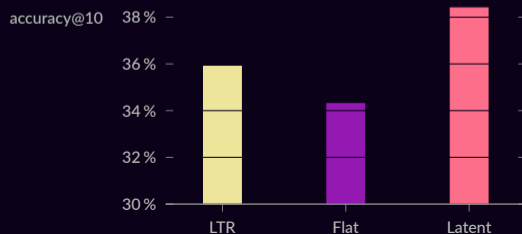


Reverse dictionary lookup

(definitions)

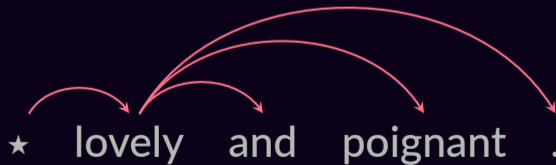


(concepts)



Syntax vs. Composition Order

CoreNLP parse, $p = 21.4\%$

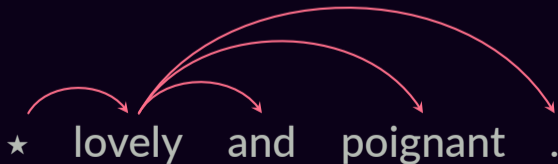


Syntax vs. Composition Order

$p = 22.6\%$

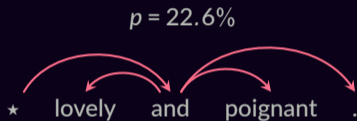


CoreNLP parse, $p = 21.4\%$

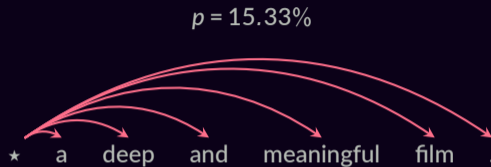


...

Syntax vs. Composition Order



CoreNLP parse, $p = 21.4\%$



$p = 15.27\%$



CoreNLP parse, $p = 0\%$

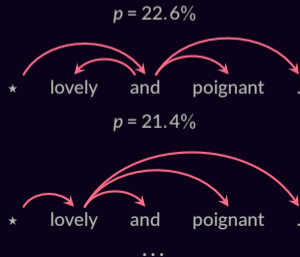
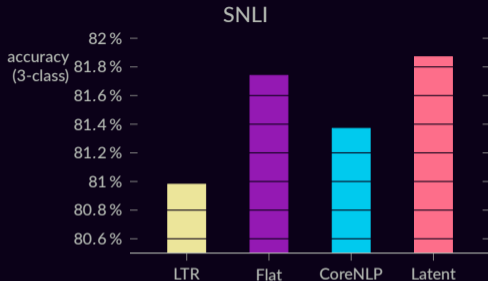


Conclusions

Latent structured variables for uncertainty & compositionality

Tractable marginalization via SparseMAP inference

Flexible model: arbitrary function of discrete latent structures



✉ vladavene.ro

🏠 <https://vene.ro>

🐙 github.com/vene/sparsemap

🐦 avnfrombucharest

Acknowledgements



This work was supported by the European Research Council (ERC StG DeepSPIN 758969) and by the Fundação para a Ciência e Tecnologia through contract UID/EEA/50008/2013.

Some icons by Dave Gandy and Freepik via [flaticon.com](https://www.flaticon.com/).