Interpretable Structure Induction Via Sparse Attention

Ben Peters Instituto de Telecomunicações

→ **Vlad Niculae** IT

André Martins IT & Unbabel

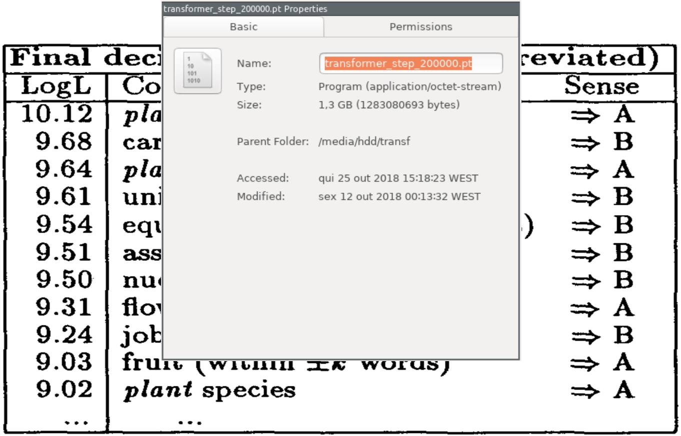
github.com/vene/{sparseattn, sparsemap} @vnfrombucharest

Sparse linear models are more interpretable...

Final decision list for *plant* (abbreviated)		
LogL	Collocation	Sense
10.12	*plant* growth	⇒ A
9.68	car (within $\pm k$ words)	⇒ B
9.64	*plant* height	⇒ A
9.61	union (within $\pm k$ words)	⇒ B
9.54	equipment (within $\pm k$ words)	⇒ B
9.51	assembly *plant*	⇒ B
9.50	nuclear *plant*	⇒ B
9.31	flower (within $\pm k$ words)	⇒ A
9.24	job (within $\pm k$ words)	⇒ B
9.03	fruit (within $\pm k$ words)	⇒ A
9.02	*plant* species	⇒ A
...	...	

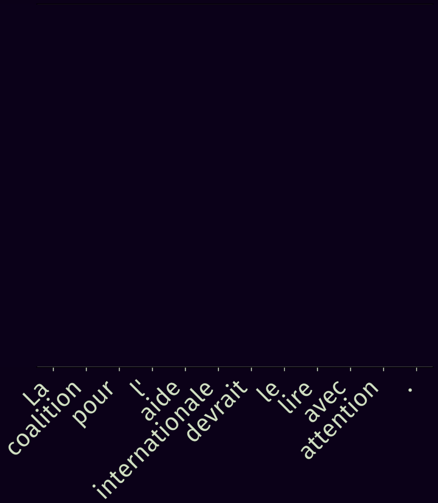
(Decision list from Yarowsky (1995)

Sparse linear models are more interpretable... but we use bigger models today!

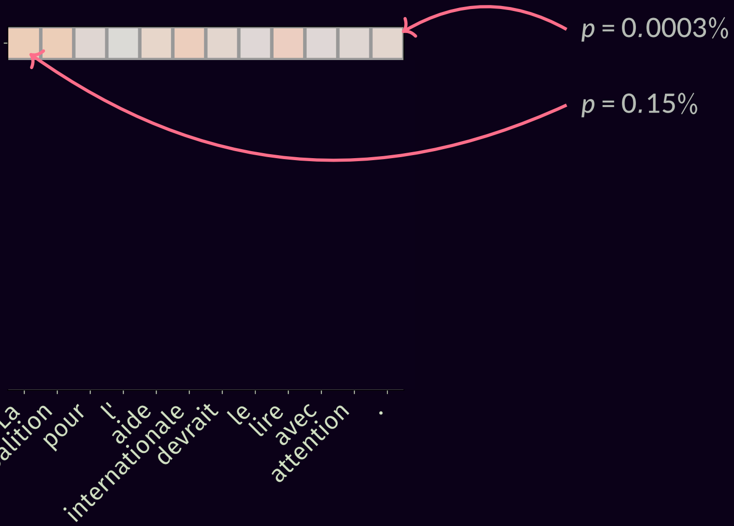


(Decision list from Yarowsky (1995))

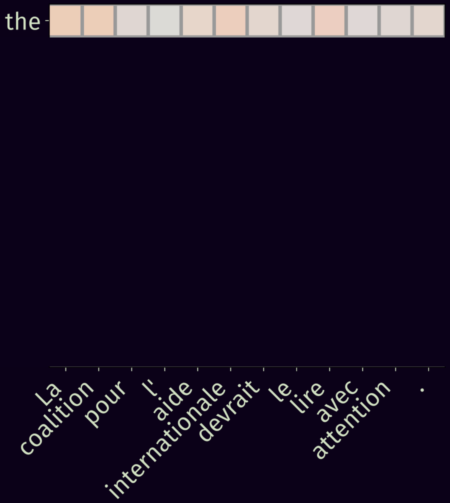
Neural Attention Mechanisms



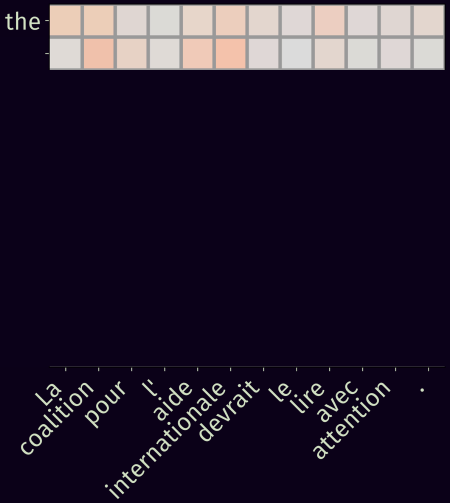
Neural Attention Mechanisms



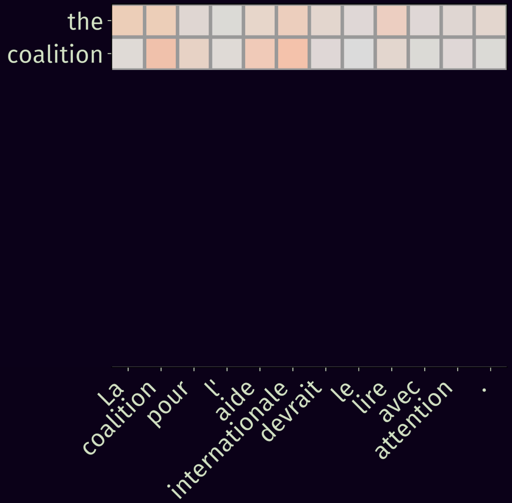
Neural Attention Mechanisms



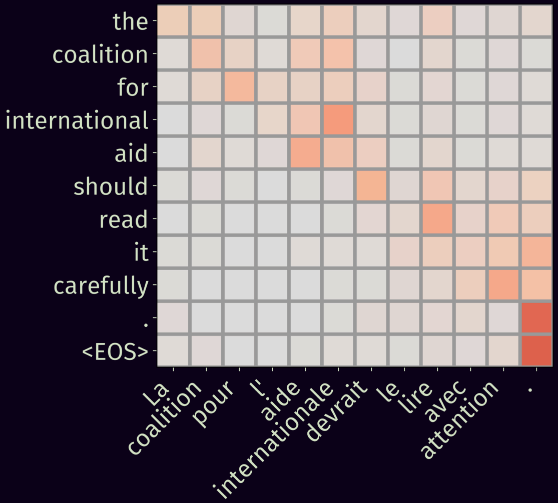
Neural Attention Mechanisms



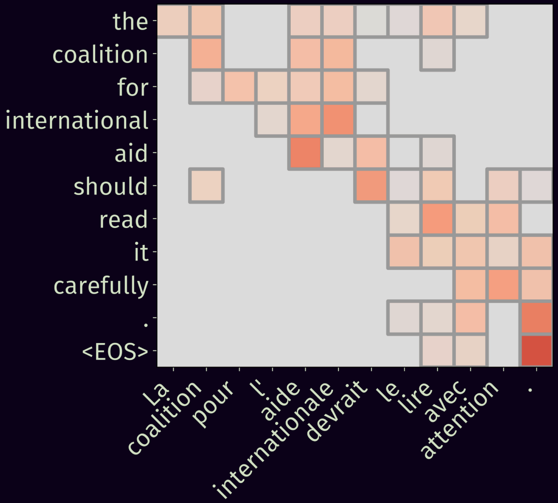
Neural Attention Mechanisms



Neural Attention Mechanisms

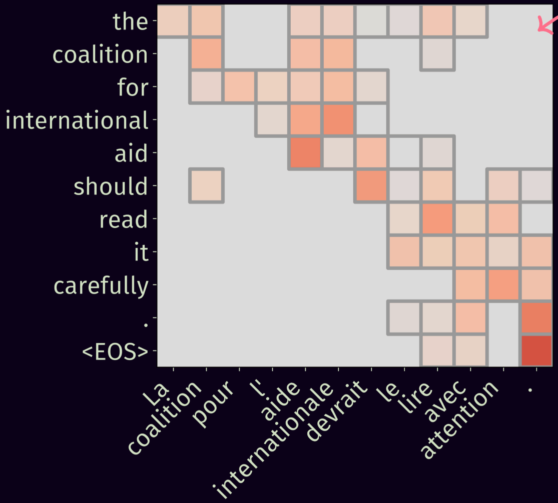


Sparse Neural Attention



sparsemax
(Martins & Astudillo, 2016)

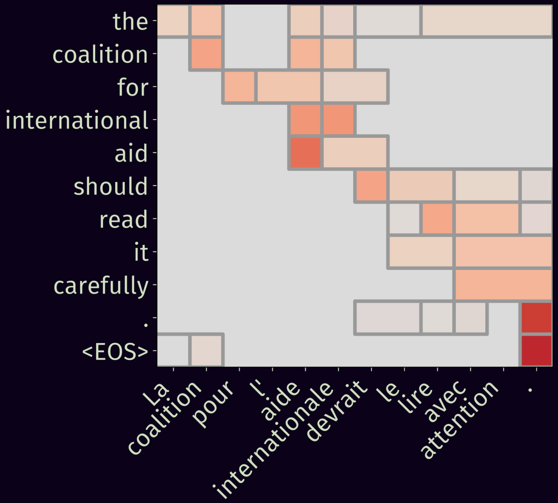
Sparse Neural Attention



$p = 0.0\%!$

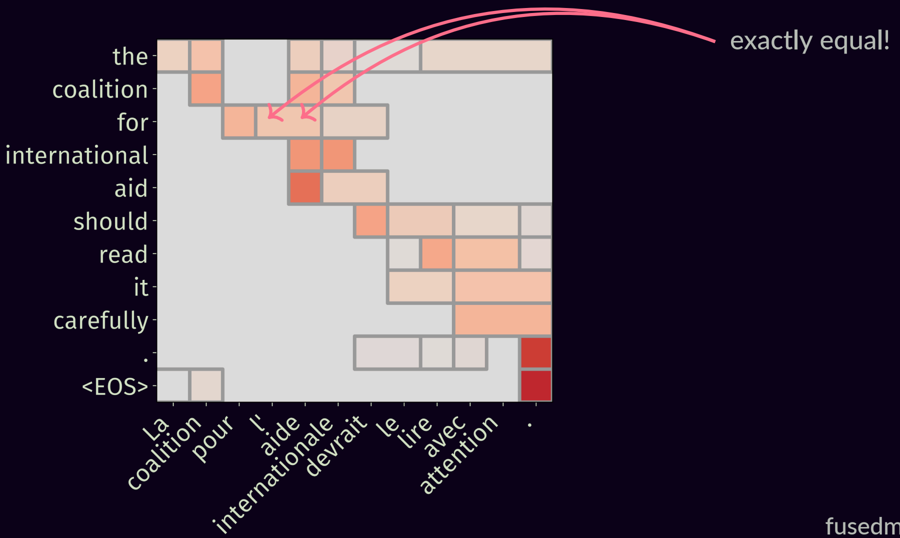
sparsemax
(Martins & Astudillo, 2016)

Structured & Sparse Attention



fusedmax
(Niculae & Blondel, 2017)

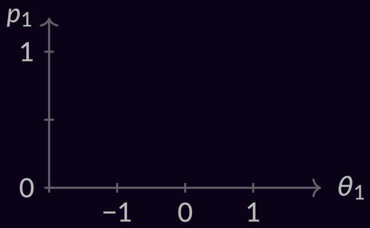
Structured & Sparse Attention

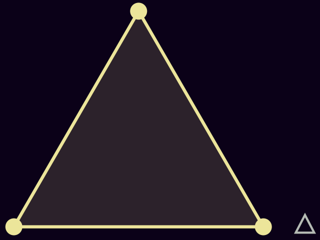


fusedmax
(Niculae & Blondel, 2017)

Smoothed Max Operators

$$\Pi_\Omega(\boldsymbol{\theta}) = \underset{\boldsymbol{p}\in\triangle}{\arg\max}\, \boldsymbol{p}^\top \boldsymbol{\theta} - \Omega(\boldsymbol{p})$$

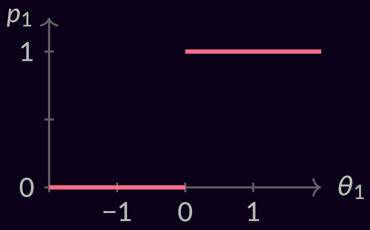


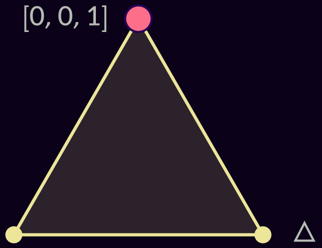


Smoothed Max Operators

$$\Pi_\Omega(\boldsymbol{\theta}) = \underset{\boldsymbol{p}\in\triangle}{\arg\max}\, \boldsymbol{p}^\top\boldsymbol{\theta} - \Omega(\boldsymbol{p})$$

- argmax: $\Omega(\boldsymbol{p}) = 0$

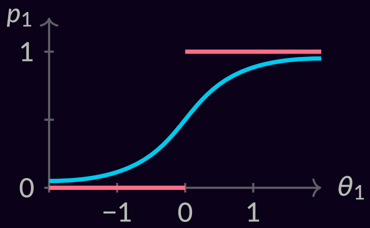


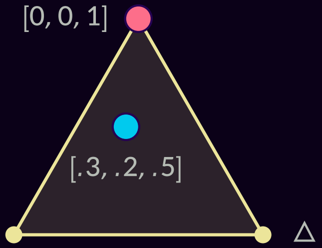


Smoothed Max Operators

$$\Pi_\Omega(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{p}\in\triangle} \boldsymbol{p}^\top\boldsymbol{\theta} - \Omega(\boldsymbol{p})$$

- argmax: $\Omega(\boldsymbol{p}) = 0$
- softmax: $\Omega(\boldsymbol{p}) = \sum_j p_j \log p_j$

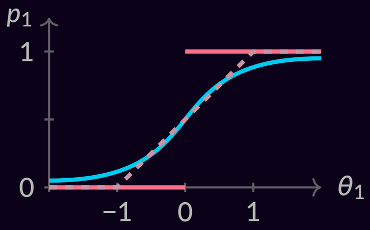


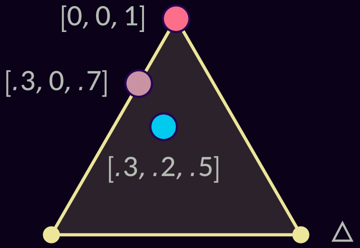


Smoothed Max Operators

$$\Pi_\Omega(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{p}\in\triangle} \boldsymbol{p}^\top \boldsymbol{\theta} - \Omega(\boldsymbol{p})$$

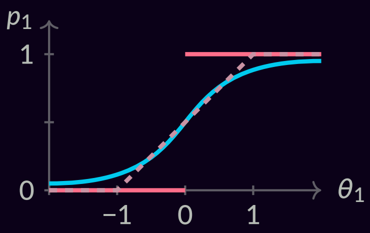
- argmax: $\Omega(\boldsymbol{p}) = 0$
- softmax: $\Omega(\boldsymbol{p}) = \sum_j p_j \log p_j$
- sparsemax: $\Omega(\boldsymbol{p}) = 1/2\|\boldsymbol{p}\|_2^2$





Smoothed Max Operators

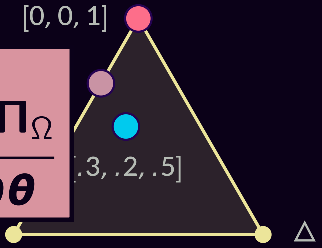
$$\Pi_\Omega(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{p}\in\triangle} \boldsymbol{p}^\top\boldsymbol{\theta} - \Omega(\boldsymbol{p})$$



- argmax: $\Omega(\boldsymbol{p}) = 0$
- softmax: $\Omega(\boldsymbol{p}) = \sum_j p_j \log p_j$
- sparsemax: $\Omega(\boldsymbol{p}) =$

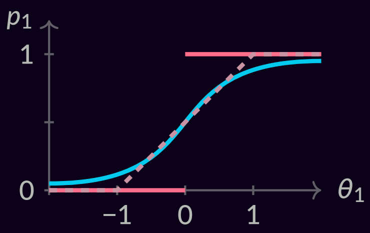
Unlike lasso,

sparse attention needs $\dfrac{\partial\, \Pi_\Omega}{\partial \boldsymbol{\theta}}$



Smoothed Max Operators

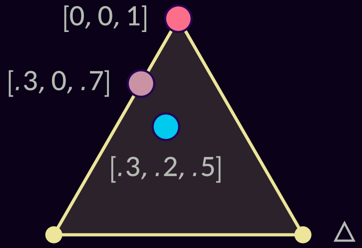
$$\Pi_\Omega(\boldsymbol{\theta}) = \underset{\boldsymbol{p}\in\triangle}{\arg\max}\, \boldsymbol{p}^\top \boldsymbol{\theta} - \Omega(\boldsymbol{p})$$



- argmax: $\Omega(\boldsymbol{p}) = 0$
- softmax: $\Omega(\boldsymbol{p}) = \sum_j p_j \log p_j$
- sparsemax: $\Omega(\boldsymbol{p}) = 1/2\|\boldsymbol{p}\|_2^2$

fusedmax: $\Omega(\boldsymbol{p}) = 1/2\|\boldsymbol{p}\|_2^2 + \sum_j |p_j - p_{j-1}|$

oscarmax: $\Omega(\boldsymbol{p}) = 1/2\|\boldsymbol{p}\|_2^2 + \sum_{i,j} \max(p_i, p_j)$

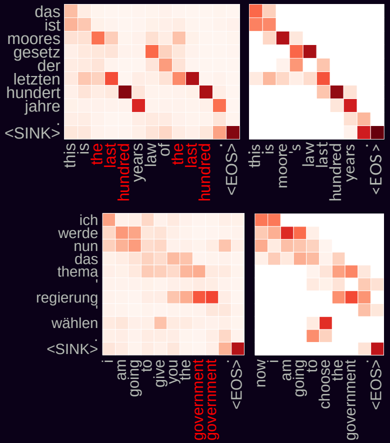


Constrained Attention

$$\underset{\substack{\boldsymbol{p}\in\Delta \\ \boldsymbol{a}\le\boldsymbol{p}\le\boldsymbol{b}}}{\arg\max}\ \boldsymbol{p}^\top\boldsymbol{\theta} - \Omega_1(\boldsymbol{p})$$

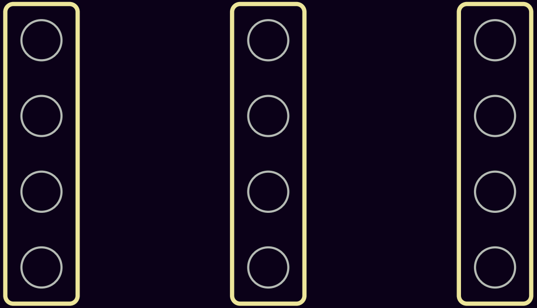
$$= \underset{\boldsymbol{p}\in\Delta}{\arg\max}\ \boldsymbol{p}^\top\boldsymbol{\theta} - \underbrace{\Omega(\boldsymbol{p})}_{:=\Omega_1+\mathrm{Id}_{[a,b]}}$$

e.g., fertility constraints for NMT

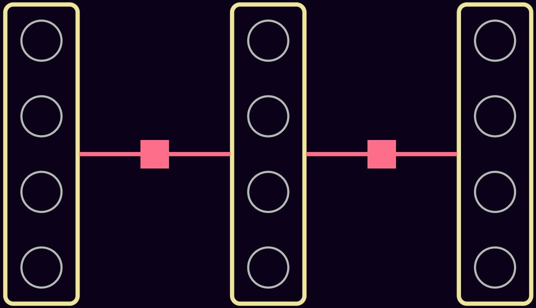


(Kreutzer & Martins, 18)
(Malaviya et al, 18)

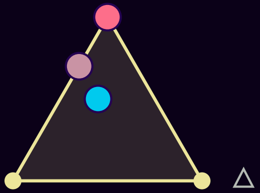
Structured Attention & Graphical Models



Structured Attention & Graphical Models



- **argmax** $\arg\max_{\boldsymbol{p}\in\triangle} \boldsymbol{p}^\top\boldsymbol{\theta}$
- **softmax** $\arg\max_{\boldsymbol{p}\in\triangle} \boldsymbol{p}^\top\boldsymbol{\theta} + \mathrm{H}(\boldsymbol{p})$
- **sparsemax** $\arg\max_{\boldsymbol{p}\in\triangle} \boldsymbol{p}^\top\boldsymbol{\theta} - 1/2\|\boldsymbol{p}\|^2$



argmax $\arg\max_{\boldsymbol{p}\in\triangle} \boldsymbol{p}^\top \boldsymbol{\theta}$

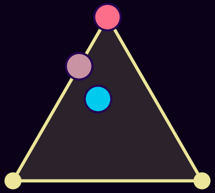
softmax $\arg\max_{\boldsymbol{p}\in\triangle} \boldsymbol{p}^\top \boldsymbol{\theta} + \mathrm{H}(\boldsymbol{p})$

sparsemax $\arg\max_{\boldsymbol{p}\in\triangle} \boldsymbol{p}^\top \boldsymbol{\theta} - 1/2\|\boldsymbol{p}\|^2$

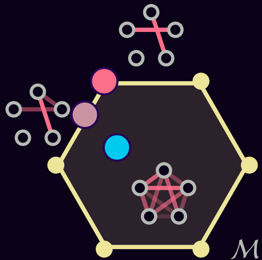
MAP $\arg\max_{\boldsymbol{\mu}\in\mathcal{M}} \boldsymbol{\mu}^\top \boldsymbol{\eta}$

marginals $\arg\max_{\boldsymbol{\mu}\in\mathcal{M}} \boldsymbol{\mu}^\top \boldsymbol{\eta} + \widetilde{\mathrm{H}}(\boldsymbol{\mu})$

SparseMAP $\arg\max_{\boldsymbol{\mu}\in\mathcal{M}} \boldsymbol{\mu}^\top \boldsymbol{\eta} - 1/2\|\boldsymbol{\mu}\|^2$



$\triangle$



$\mathcal{M}$

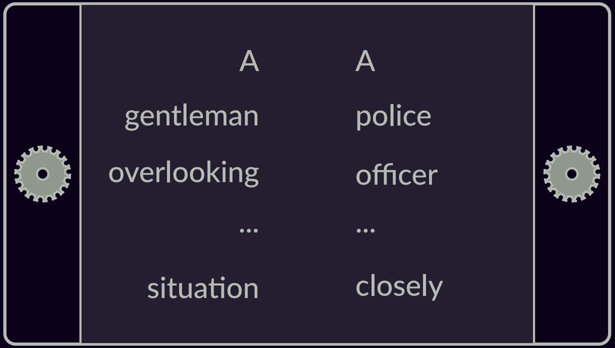
Structured Attention for Alignments

NLI

premise: A gentleman overlooking a neighborhood situation.
hypothesis: A police officer watches a situation closely.

input

(P, H)



(Model: ESIM)

output

entails
contradicts
neutral

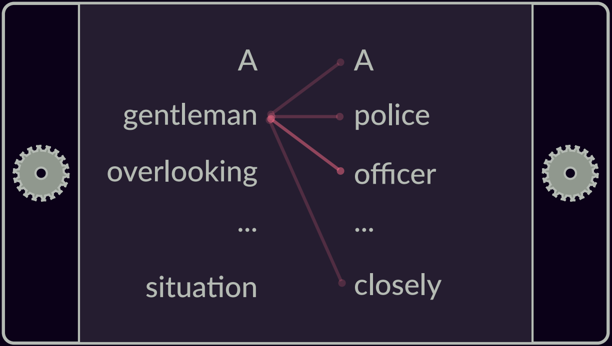
Structured Attention for Alignments

NLI

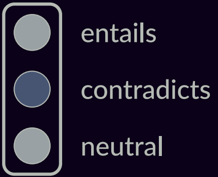
premise: A gentleman overlooking a neighborhood situation.
hypothesis: A police officer watches a situation closely.

input

(P, H)



output



(Model: ESIM)

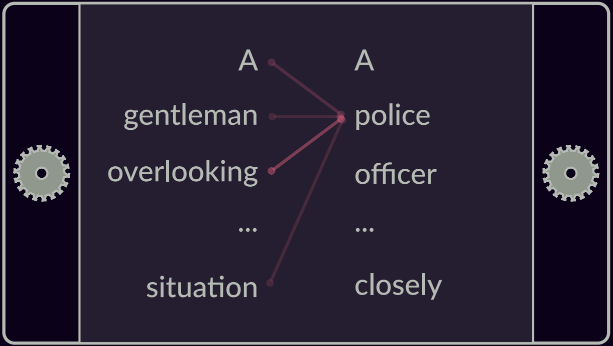
Structured Attention for Alignments

NLI

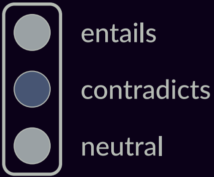
premise: A gentleman overlooking a neighborhood situation.
hypothesis: A police officer watches a situation closely.

input

(P, H)



output



(Model: ESIM)

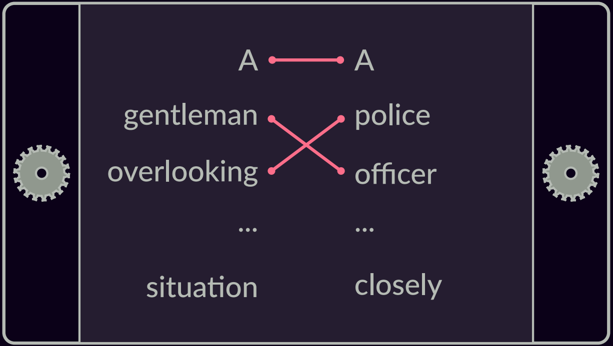
Structured Attention for Alignments

NLI

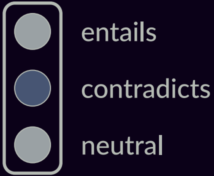
premise: A gentleman overlooking a neighborhood situation.
hypothesis: A police officer watches a situation closely.

input

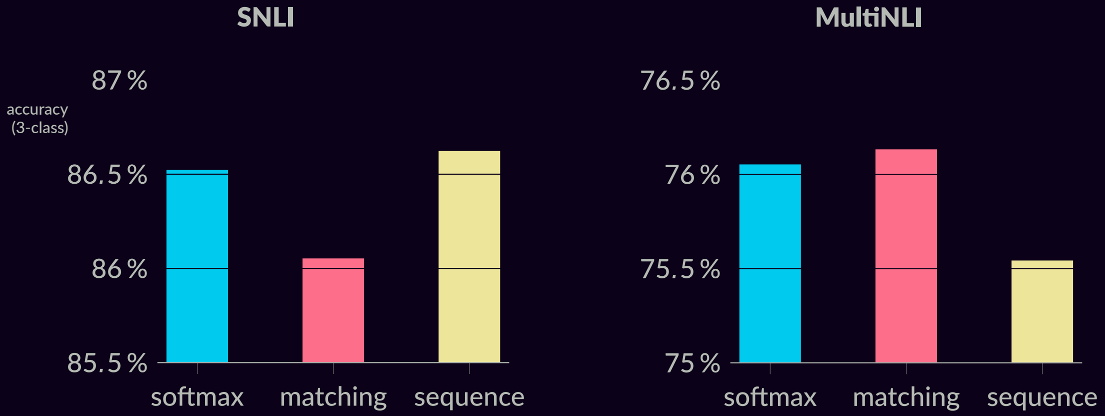
(P, H)

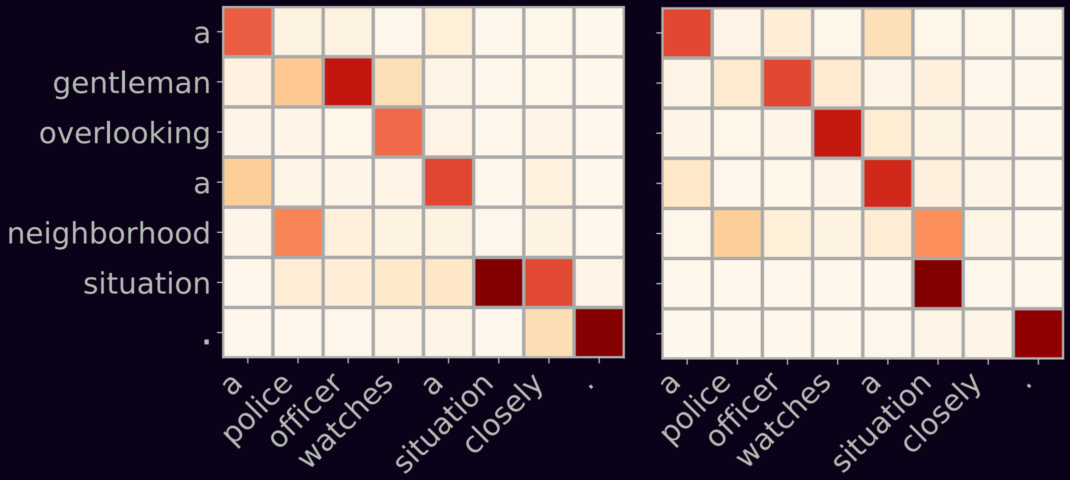


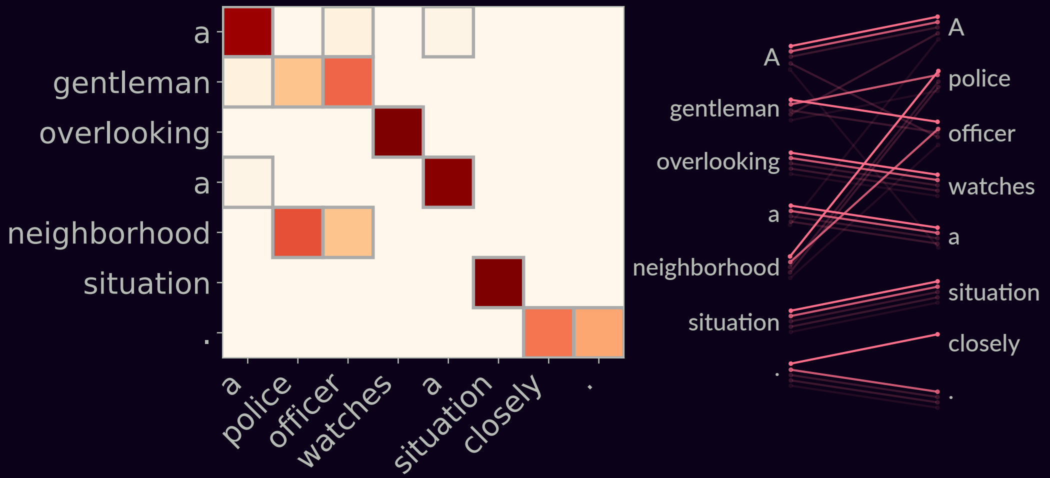
output



(Proposed model: global matching)

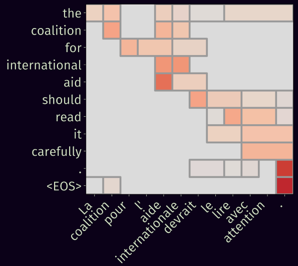

SNLI
87 %
accuracy
(3-class)
86.5 %
86 %
85.5 %
softmax
matching
sequence
MultiNLI
76.5 %
76 %
75.5 %
75 %
softmax
matching
sequence


a
gentleman
overlooking
a
neighborhood
situation
.
a
police
officer
watches
a
situation
closely
.
a
police
officer
watches
a
situation
closely
.

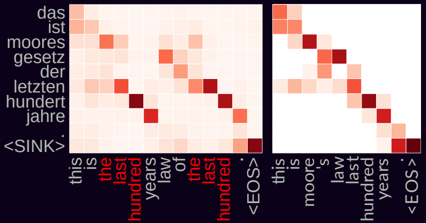


a
gentleman
overlooking
a
neighborhood
situation
.
a
police
officer
watches
a
situation
closely
.
A
gentleman
overlooking
a
neighborhood
situation
.
A
police
officer
watches
a
situation
closely
.

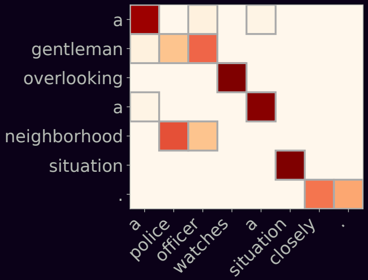
Summary: Neural attention with...



structured sparsity
(*e.g.* fusedmax)



constraints
(*e.g.* csparsemax — fertility)



structure
(*e.g.* SparseMAP alignments)

and dynamic computation graphs with structured latent variables! (Friday 15:36 in 3B)

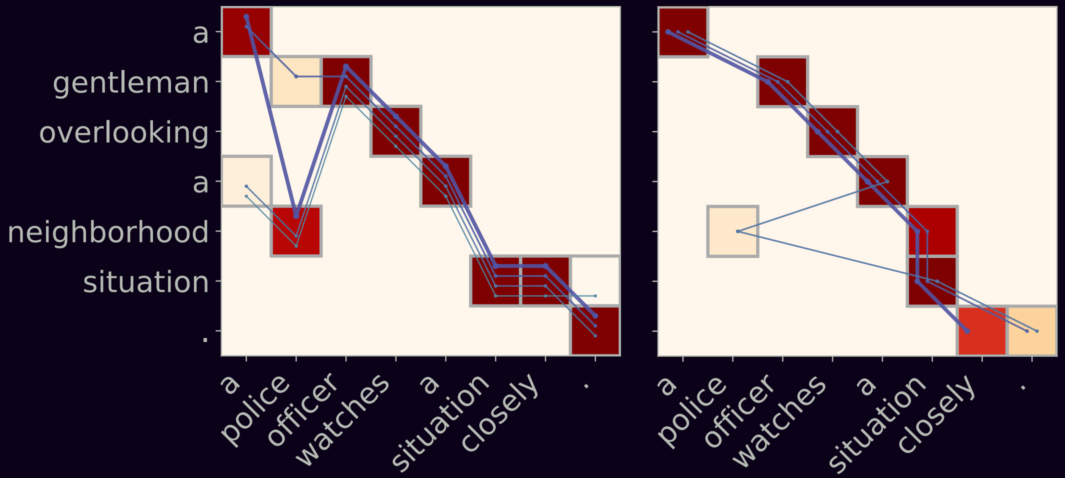
vlad@vene.ro github.com/vene/sparsemap https://vene.ro @vnfrombucharest

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.

Extra slides


a
gentleman
overlooking
a
neighborhood
situation
.
a
police
officer
watches
a
situation
closely
.
a
police
officer
watches
a
situation
closely
.