Interpretable Structure Induction Via Sparse Attention

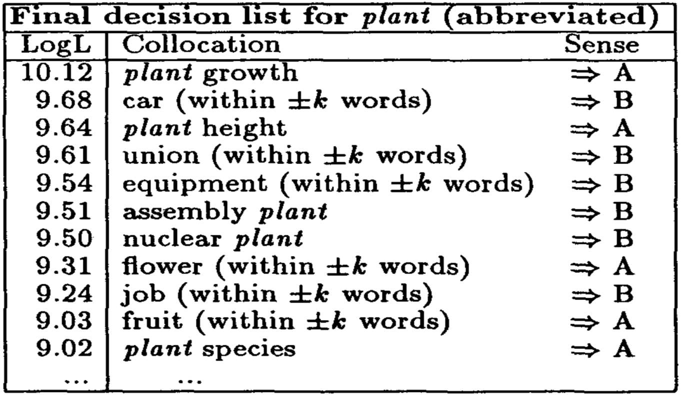
Ben Peters Instituto de Telecomunicações

→ **Vlad Niculae** IT

André Martins IT & Unbabel

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

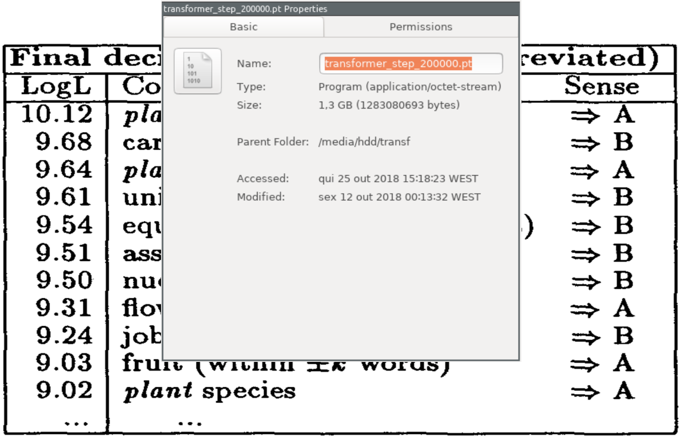
Sparse linear models are more interpretable...



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



Name: transformer_step_200000.pt

Type: Program (application/octet-stream)

Size: 1,3 GB (1283080693 bytes)

Parent Folder: /media/hdd/transf

Accessed: qui 25 out 2018 15:18:23 WEST

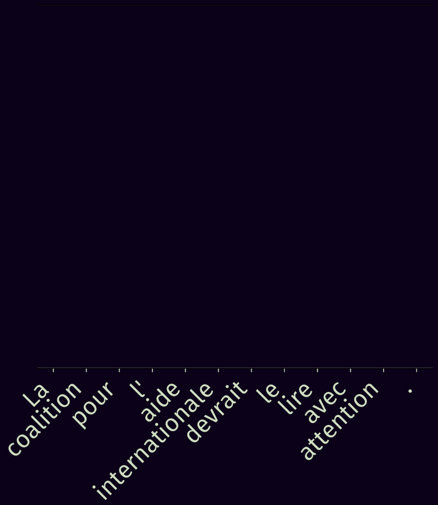
Modified: sex 12 out 2018 00:13:32 WEST

Final deci reviated)

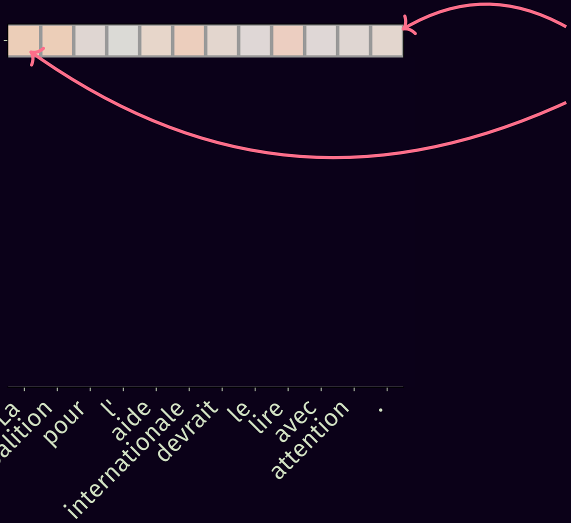
LogL	Co	Sense
10.12	*pla*	⇒ A
9.68	car	⇒ B
9.64	*pla*	⇒ A
9.61	uni	⇒ B
9.54	equ)	⇒ B
9.51	ass	⇒ B
9.50	nu	⇒ B
9.31	flo	⇒ A
9.24	job	⇒ B
9.03	fruit (within ±k words)	⇒ A
9.02	*plant* species	⇒ A
…	…	

Decision list from Yarowsky (1995)

Neural Attention Mechanisms



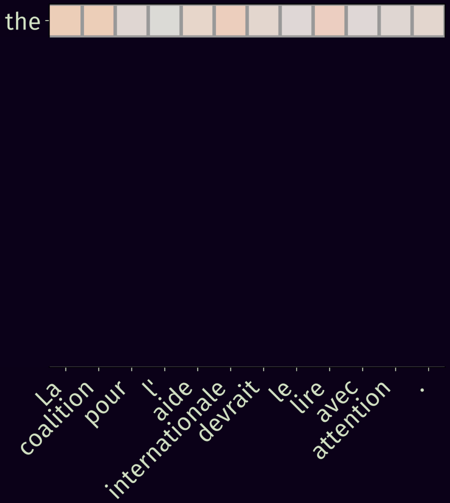
Neural Attention Mechanisms



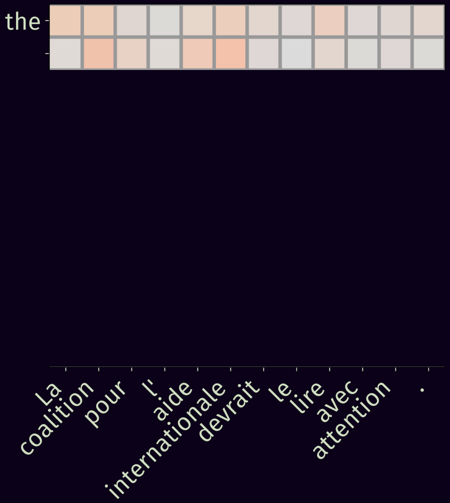
$p = 0.0003\%$

$p = 0.15\%$

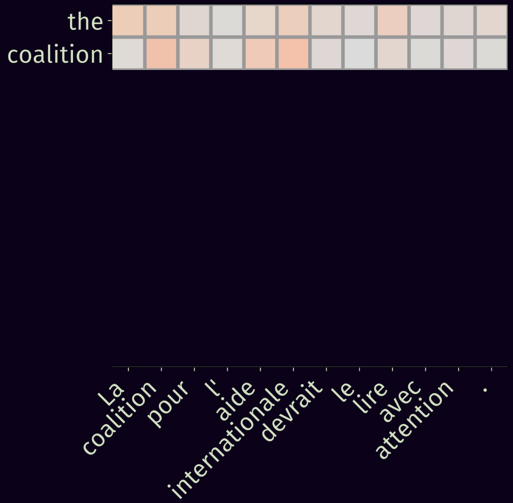
Neural Attention Mechanisms



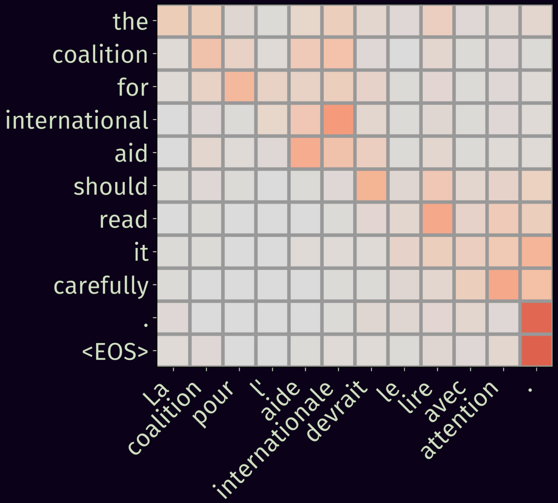
Neural Attention Mechanisms



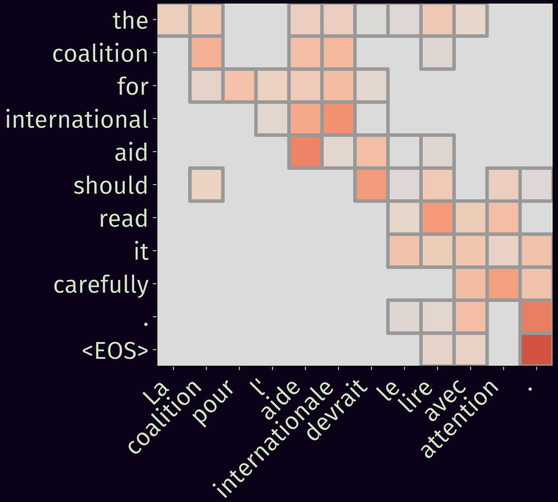
Neural Attention Mechanisms



Neural Attention Mechanisms

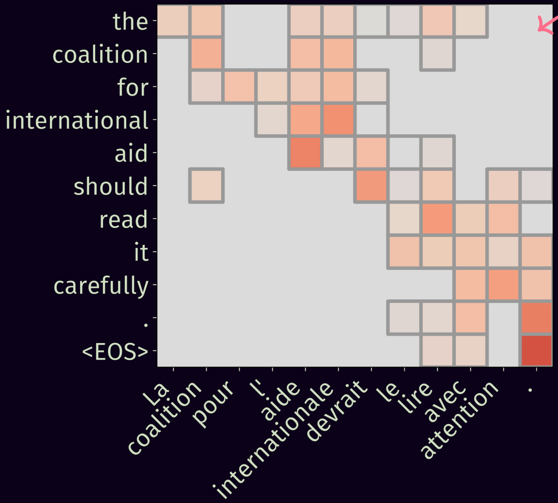


Sparse Neural Attention



sparsemax
(Martins & Astudillo, 2016)

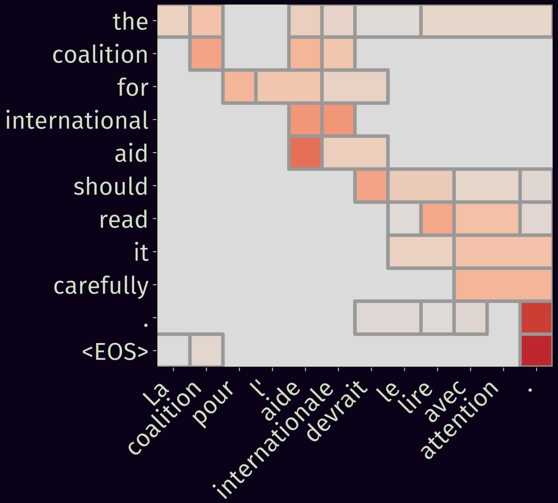
Sparse Neural Attention



$p = 0.0\%$!

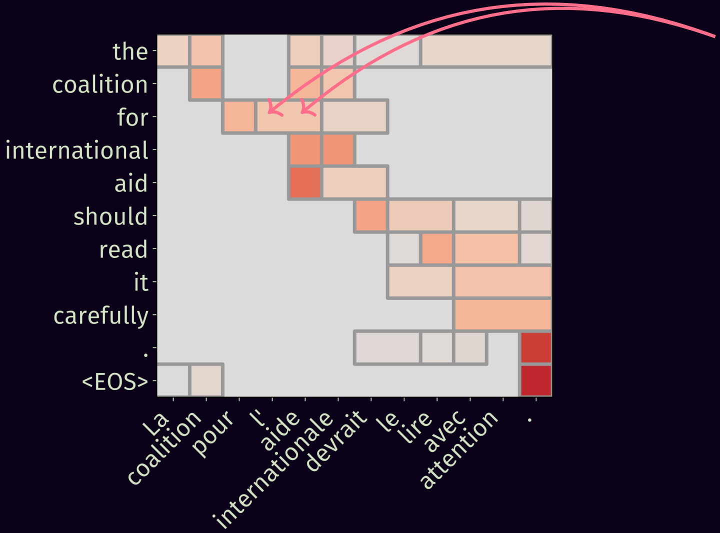
sparsemax
(Martins & Astudillo, 2016)

Structured & Sparse Attention



fusedmax
(Niculae & Blondel, 2017)

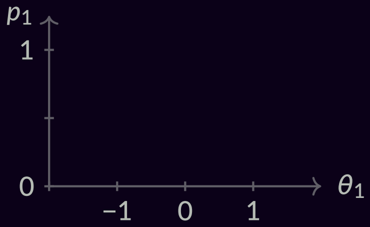
Structured & Sparse Attention

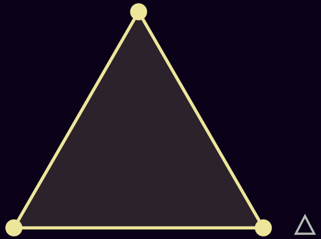


fusedmax
(Niculae & Blondel, 2017)

Smoothed Max Operators

$$\mathrm{softmax}(\boldsymbol{\theta}) = \boldsymbol{p}$$

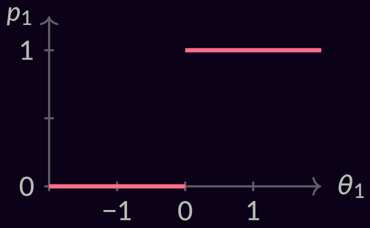


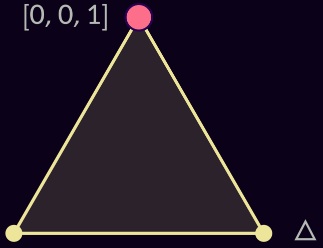


Smoothed Max Operators

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

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



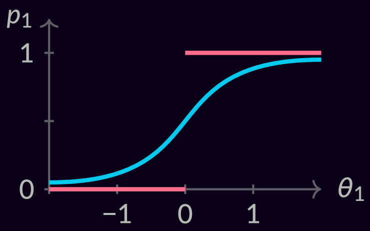


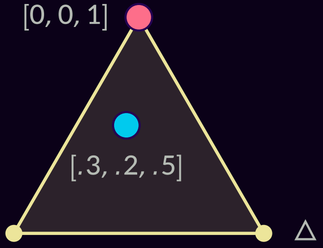
Smoothed Max Operators

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

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

● softmax: $\Omega(\boldsymbol{p}) = \sum_j p_j \log p_j$

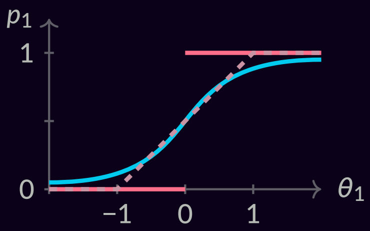


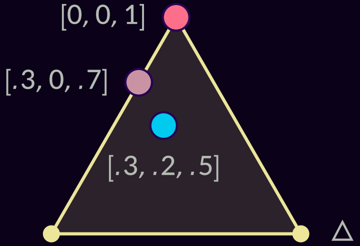


Smoothed Max Operators

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

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

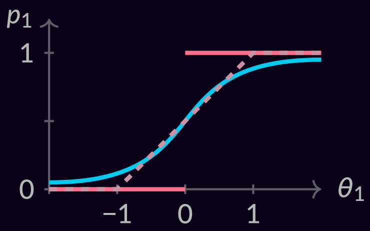


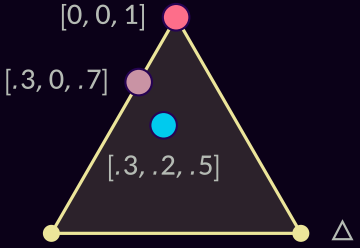


Smoothed Max Operators

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

- ● argmax: $\Omega(\boldsymbol{p}) = 0$
- ● softmax: $\Omega(\boldsymbol{p}) = \sum_j p_j \log p_j$
- ● sparsemax: $\Omega(\boldsymbol{p}) = 1/2\|\boldsymbol{p}\|_2^2$
- fusedmax: $\Omega(\boldsymbol{p}) = 1/2\|\boldsymbol{p}\|_2^2 + \sum_j |p_j - p_{j-1}|$
- oscarmax: $\Omega(\boldsymbol{p}) = 1/2\|\boldsymbol{p}\|_2^2 + \sum_{i,j} \max(p_i, p_j)$



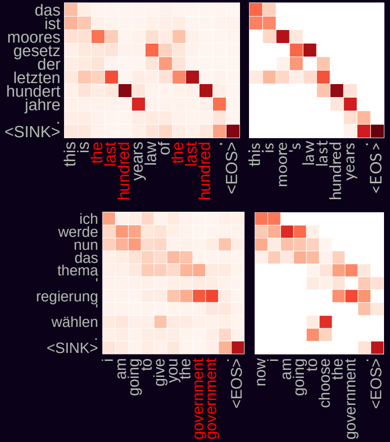


Constrained Attention

$$\arg\max_{\substack{\boldsymbol{p}\in\triangle \\ \boldsymbol{a}\le\boldsymbol{p}\le\boldsymbol{b}}} \boldsymbol{p}^\top\boldsymbol{\theta} - \Omega_1(\boldsymbol{p})$$

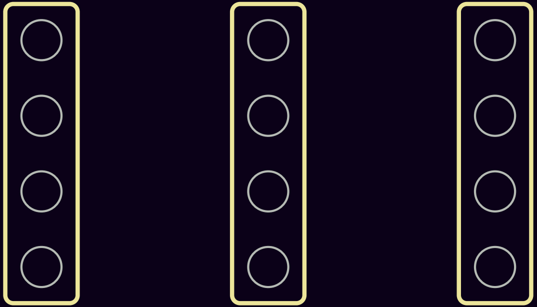
$$= \arg\max_{\boldsymbol{p}\in\triangle} \boldsymbol{p}^\top\boldsymbol{\theta} - \underbrace{\Omega(\boldsymbol{p})}_{:=\Omega_1+\mathrm{Id}_{[a,b]}}$$

e.g., fertility constraints for NMT

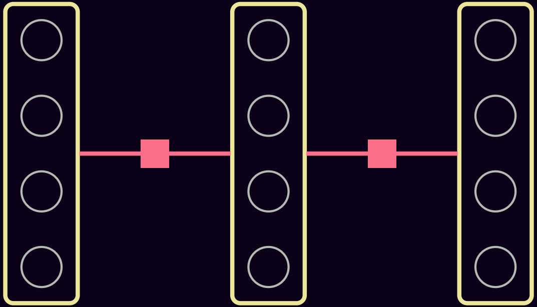


(Kreutzer & Martins, 18)
(Malaviya et al, 18)

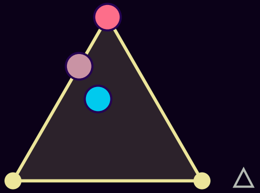
Structured Attention & Graphical Models



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

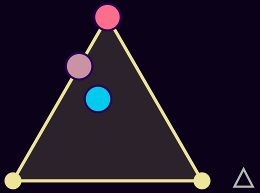
● **softmax** $\arg\max_{\boldsymbol{p}\in\triangle} \boldsymbol{p}^\top \boldsymbol{\theta} + \mathrm{H}(\boldsymbol{p})$

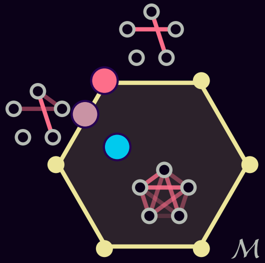
● **sparsemax** $\arg\max_{\boldsymbol{p}\in\triangle} \boldsymbol{p}^\top \boldsymbol{\theta} - 1/2\|\boldsymbol{p}\|^2$

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





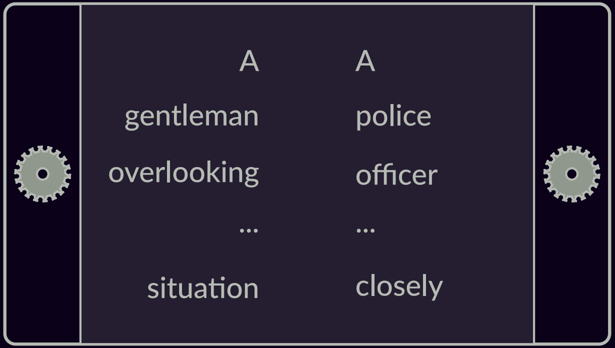
Structured Attention for Alignments

NLI

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

input

(P, H)



output

○ entails
● contradicts
○ neutral

Model: ESIM (Chen, 16)

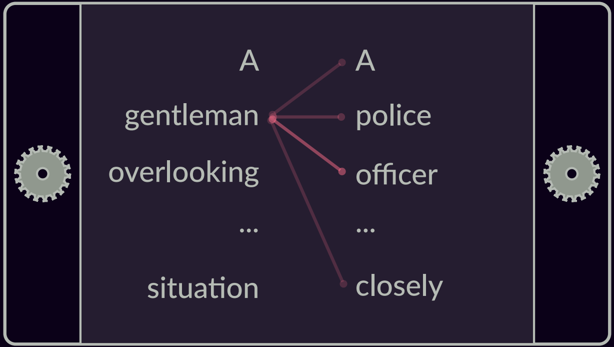
Structured Attention for Alignments

NLI

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

input

(P, H)



Model: ESIM (Chen, 16)

output

○ entails
● contradicts
○ neutral

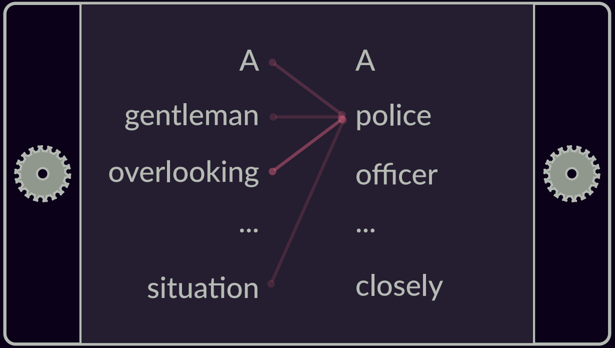
Structured Attention for Alignments

NLI

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

input

(P, H)



A
gentleman
overlooking
...
situation

A
police
officer
...
closely

output

○ entails
● contradicts
○ neutral

Model: ESIM (Chen, 16)

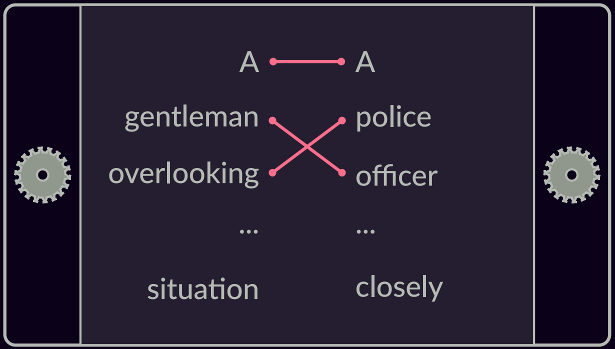
Structured Attention for Alignments

NLI

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

input

(P, H)

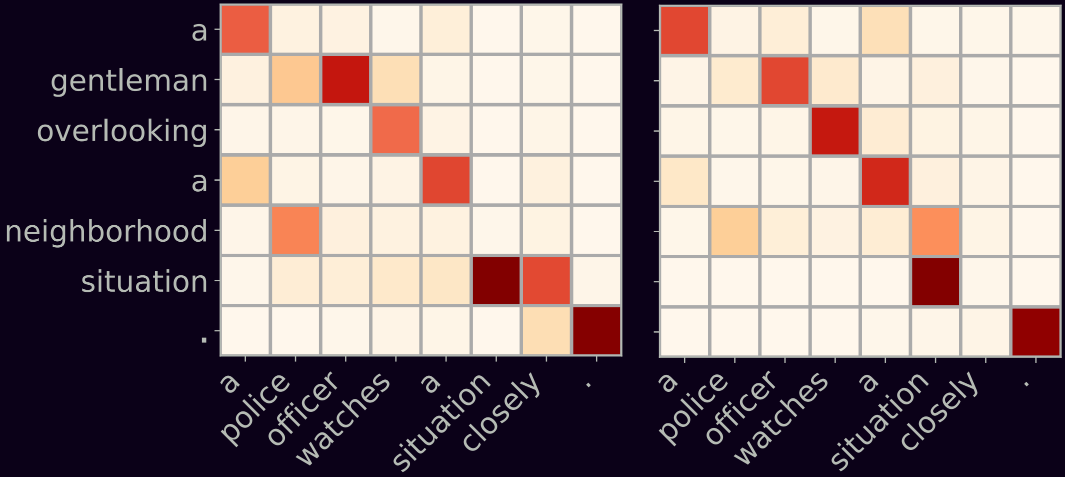


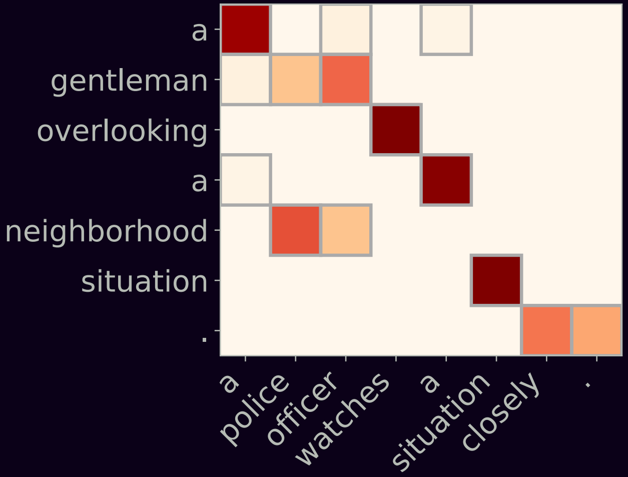
A — A
gentleman
overlooking
...
situation
police
officer
...
closely

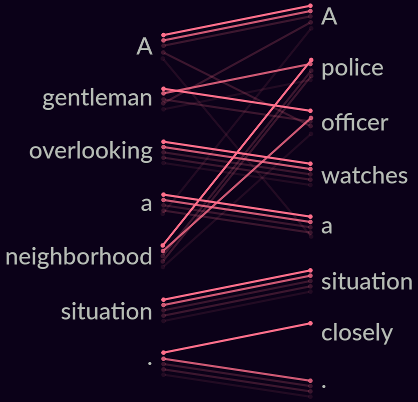
output

○ entails
● contradicts
○ neutral

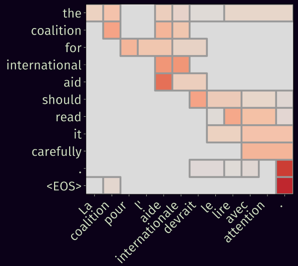
Proposed model: global matching


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

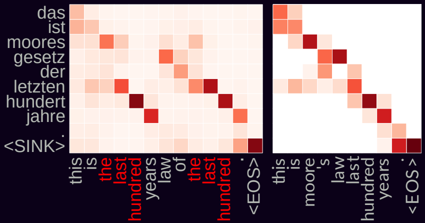

a
gentleman
overlooking
a
neighborhood
situation
.
a
police
officer
watches
a
situation
closely
.


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

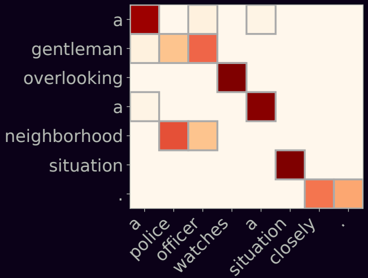
Summary: Neural attention with...



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)

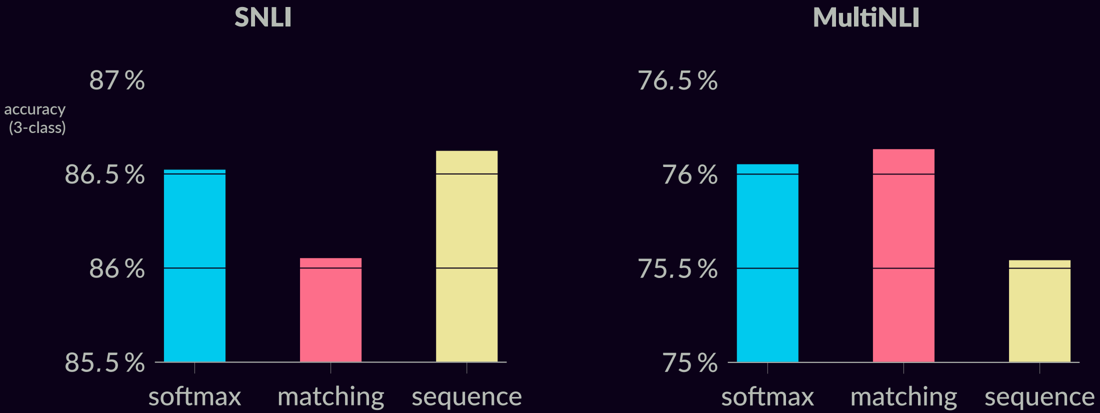
vlad@vene.ro github.com/vene/sparsemap https://vene.ro @vnfrombucharest

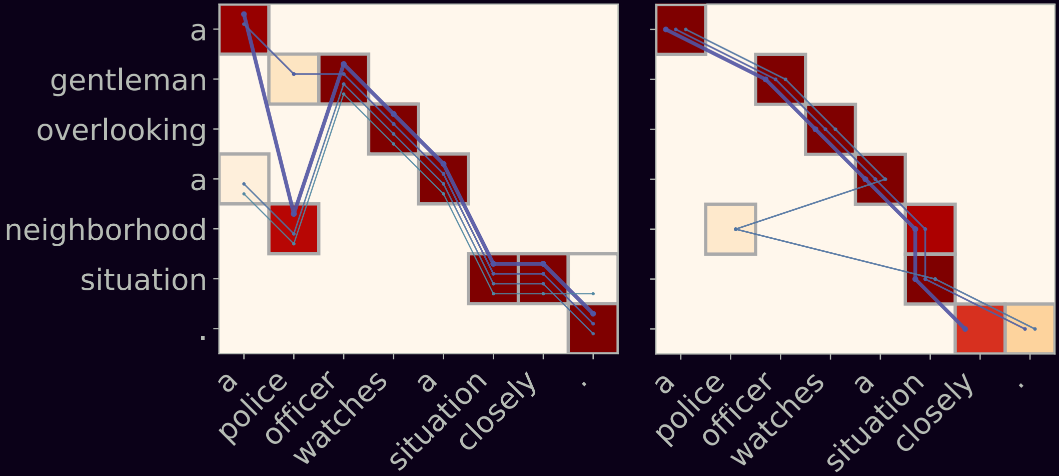
Acknowledgements

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


SNLI
MultiNLI
accuracy
(3-class)
87 %
86.5 %
86 %
85.5 %
softmax
matching
sequence
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
.