

# Analysis of “interpretable” architectures

Some architectures have components that lend themselves to inspection

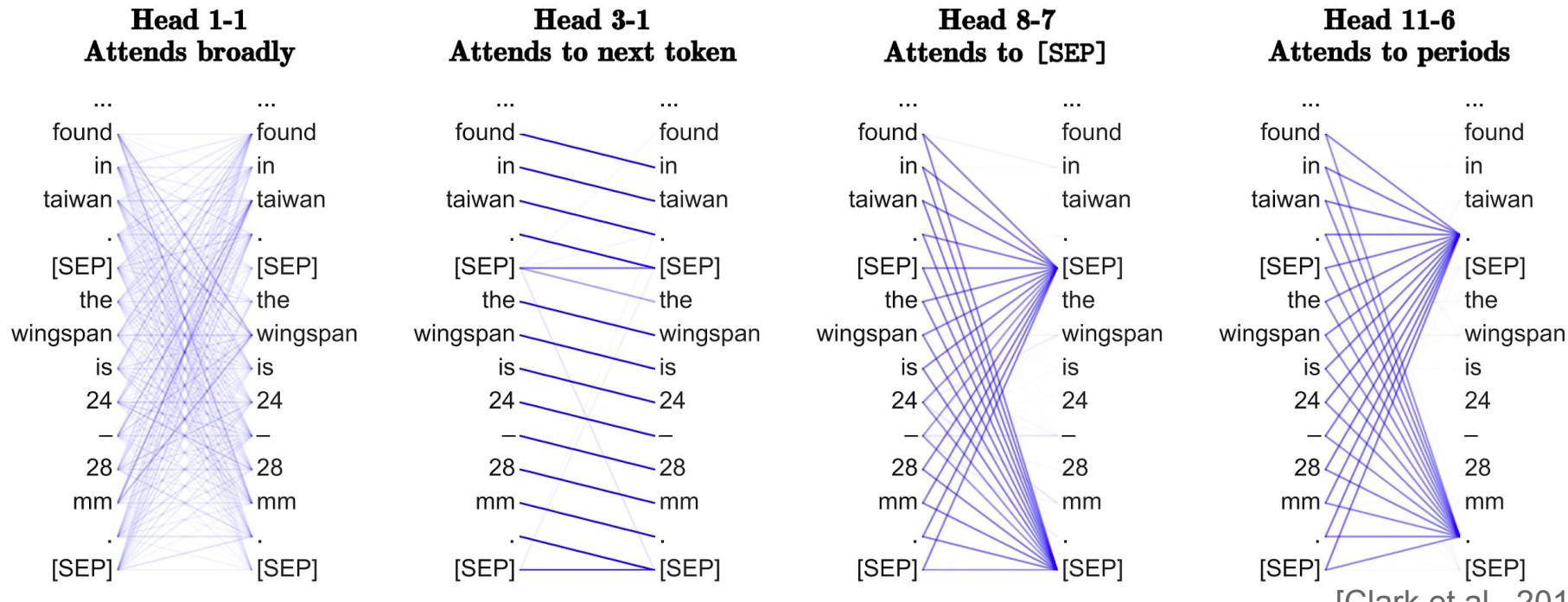
Example: Try to characterize each attention head of BERT.



# Analysis of “interpretable” architectures

Some architectures have components that lend themselves to inspection

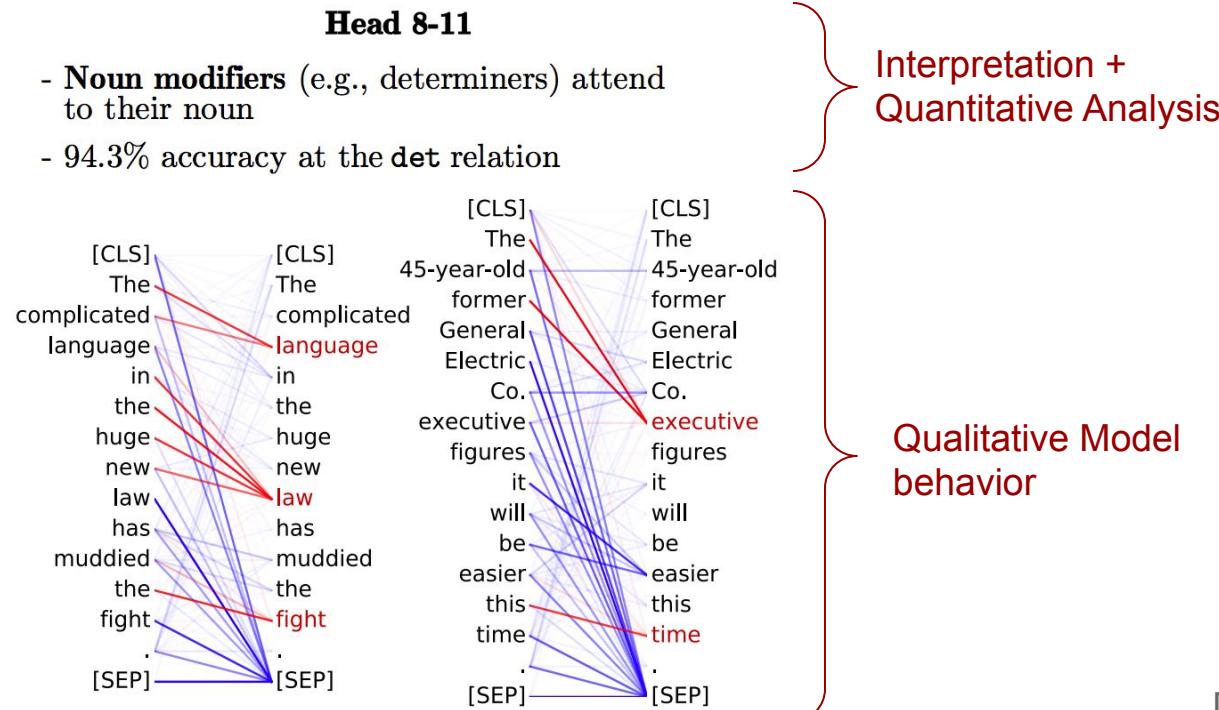
Example: Try to characterize each attention head of BERT.



# Analysis of “interpretable” architectures

Some architectures have components that lend themselves to inspection

Example: Try to characterize each attention head of BERT.



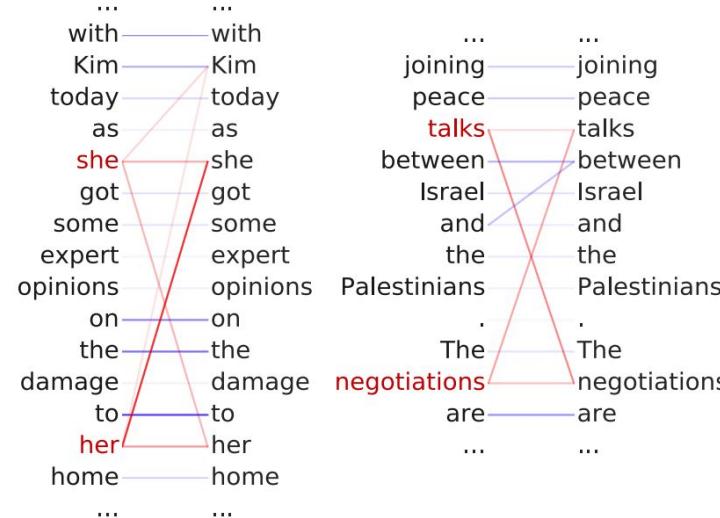
# Analysis of “interpretable” architectures

Some architectures have components that lend themselves to inspection

Example: Try to characterize each attention head of BERT.

## Head 5-4

- **Coreferent** mentions attend to their antecedents
- 65.1% accuracy at linking the head of a coreferent mention to the head of an antecedent



Interpretation +  
Quantitative Analysis

Qualitative Model  
behavior

# Understanding representations by inspection

Are individual hidden units in recurrent neural networks interpretable?

Cell sensitive to position in line:

The sole importance of the crossing of the Berezina lies in the fact  
that it plainly and indubitably proved the fallacy of all the plans for  
cutting off the enemy's retreat and the soundness of the only possible  
line of action--the one Kutuzov and the general mass of the army  
demanded--namely, simply to follow the enemy up. The French crowd fled  
at a continually increasing speed and all its energy was directed to  
reaching its goal. It fled like a wounded animal and it was impossible  
to block its path. This was shown not so much by the arrangements it  
made for crossing as by what took place at the bridges. When the bridges  
broke down, unarmed soldiers, people from Moscow and women with children  
who were with the French transport, all--carried on by vis inertiae--  
pressed forward into boats and into the ice-covered water and did not,  
surrender.



# Understanding representations by inspection

Are individual hidden units in recurrent neural networks interpretable?

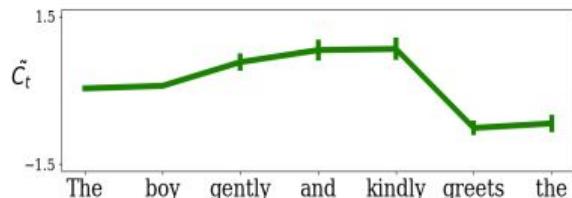
Cell that turns on inside quotes:

"You mean to imply that I have nothing to eat out of.... On the contrary, I can supply you with everything even if you want to give dinner parties," warmly replied Chichagov, who tried by every word he spoke to prove his own rectitude and therefore imagined Kutuzov to be animated by the same desire.

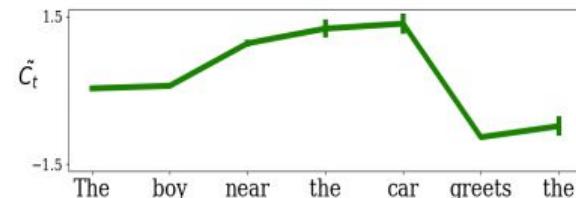
Kutuzov, shrugging his shoulders, replied with his subtle penetrating smile: "I meant merely to say what I said."

# Understanding representations by inspection

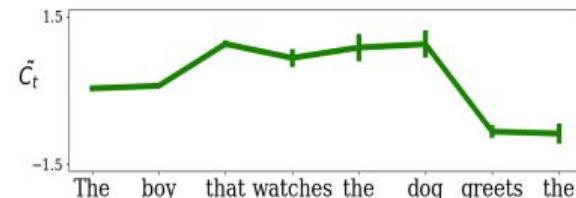
Are individual hidden units in recurrent neural networks interpretable?



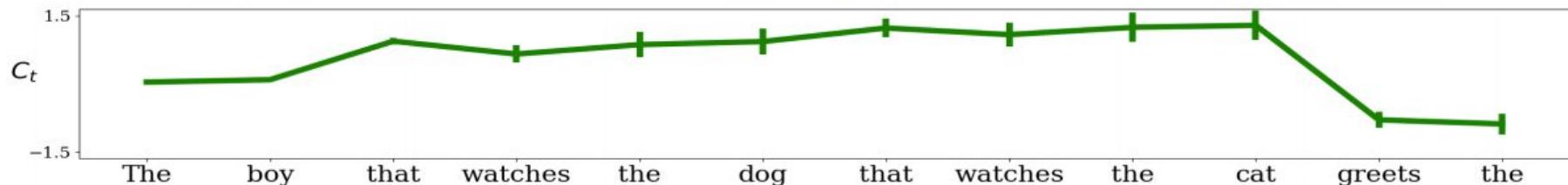
(a) 2Adv



(b) nounPP



(c) subject relative



Interpretation: this LSTM cell unit fires approximately between a subject and its verb