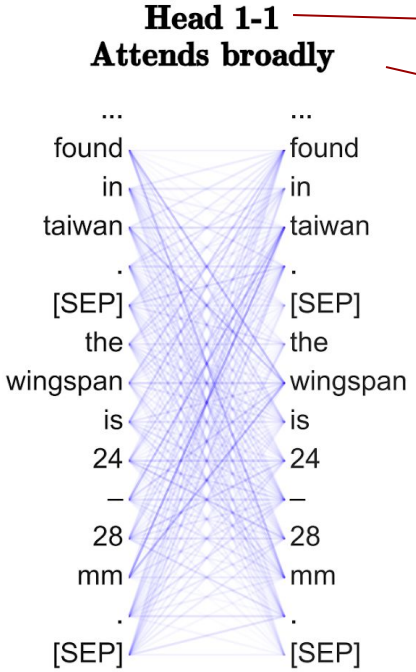
Analysis of “interpretable” architectures

Some architectures have components that lend themselves to inspection

Example: Try to characterize each attention head of BERT.



Attention head 1 of layer 1

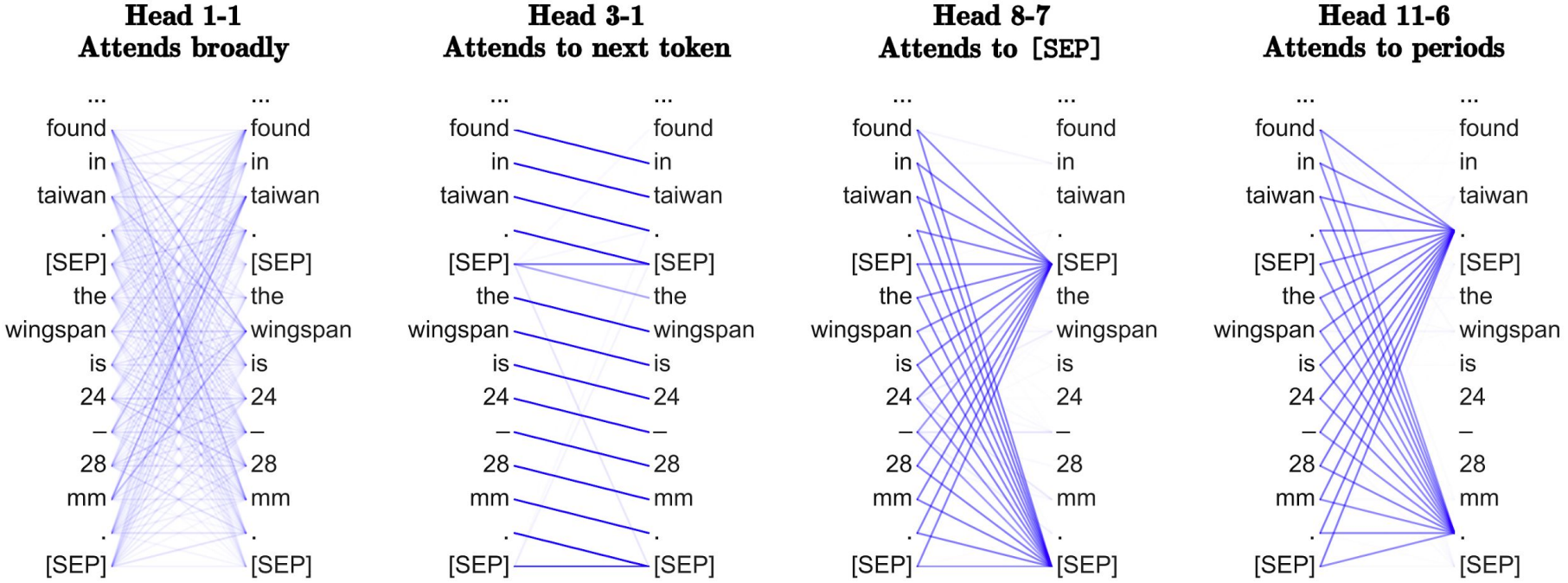
Not just on this sentence, but on most sentences, this head demonstrates a behavior like this.

[Clark et al., 2019]

Analysis of “interpretable” architectures

Some architectures have components that lend themselves to inspection

Example: Try to characterize each attention head of BERT.



[Clark et al., 2019]

Analysis of “interpretable” architectures

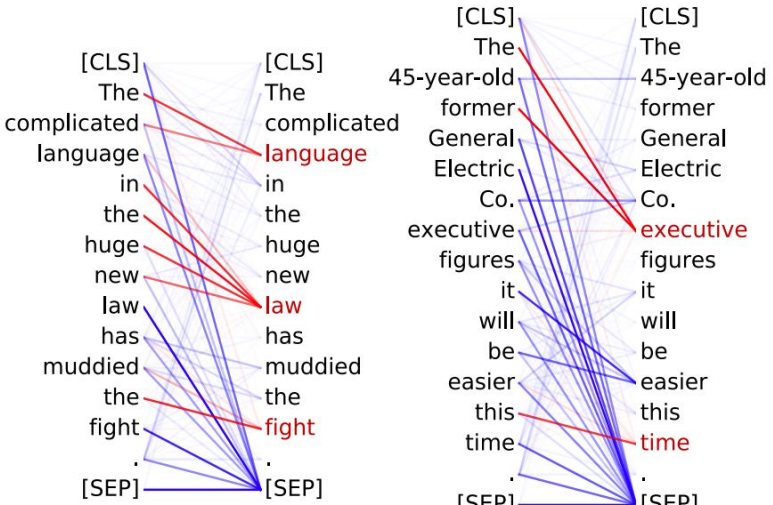
Some architectures have components that lend themselves to inspection

Example: Try to characterize each attention head of BERT.

Head 8-11

- **Noun modifiers** (e.g., determiners) attend to their noun
- 94.3% accuracy at the **det** relation

Interpretation + Quantitative Analysis



Qualitative Model behavior

[Clark et al., 2019]

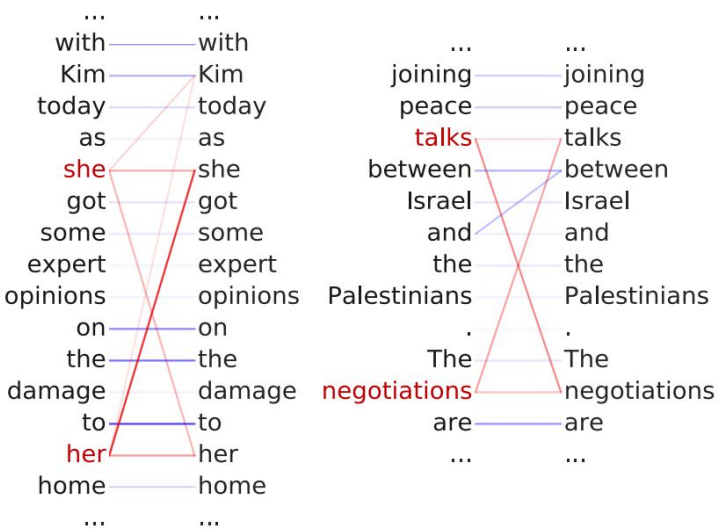
Analysis of “interpretable” architectures

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

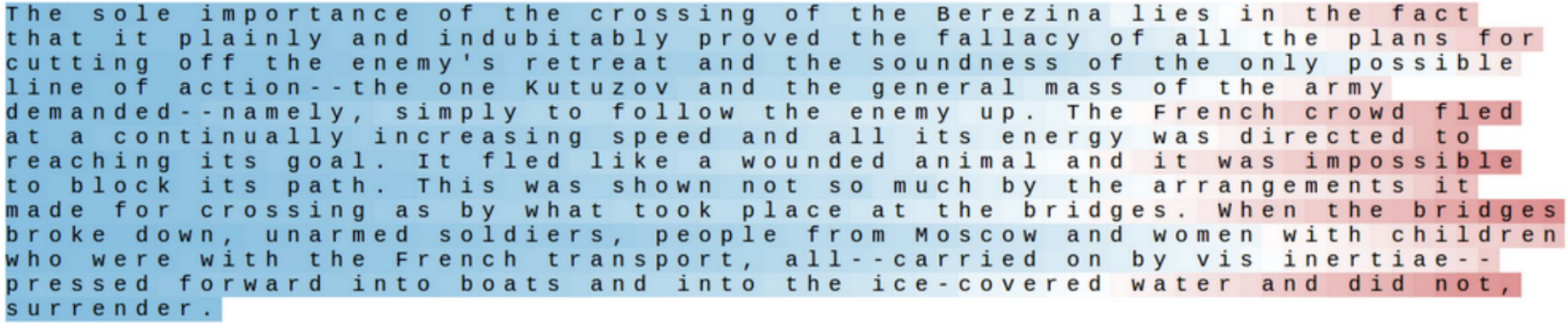
Qualitative Model behavior

[Clark et al., 2019]

Understanding representations by inspection

Are individual hidden units in recurrent neural networks interpretable?

Cell sensitive to position in line:

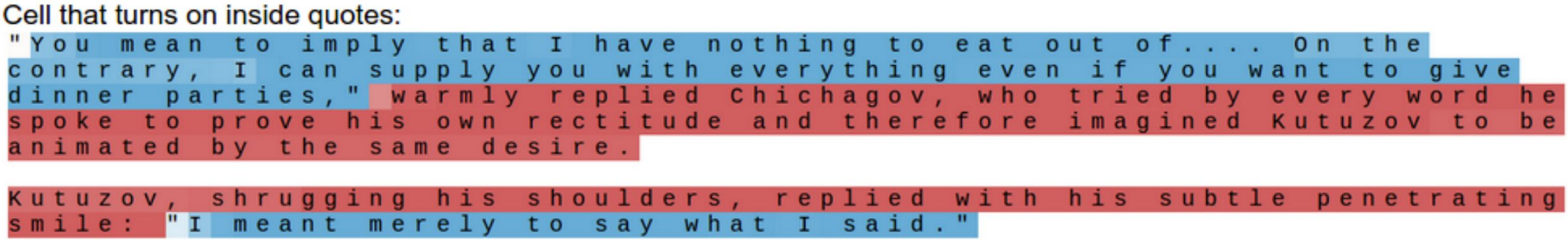


[Karpathy et al., 2016]

Understanding representations by inspection

Are individual hidden units in recurrent neural networks interpretable?

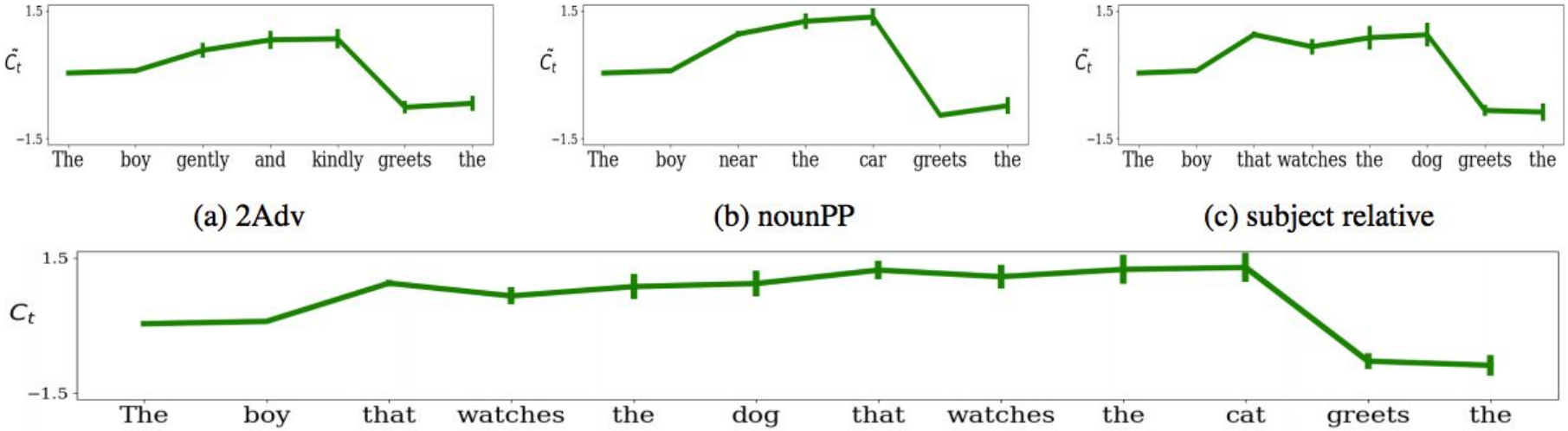
Cell that turns on inside quotes:



[Karpathy et al., 2016]

Understanding representations by inspection

Are individual hidden units in recurrent neural networks interpretable?



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

[Lakretz et al., 2019]