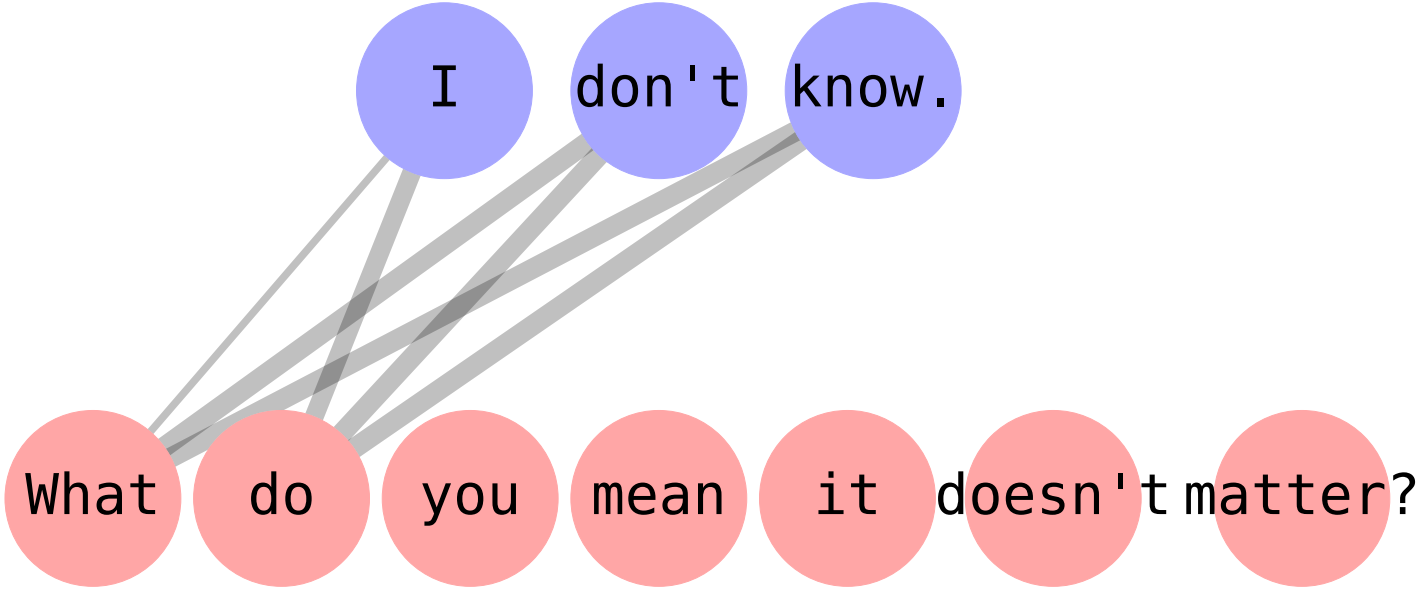
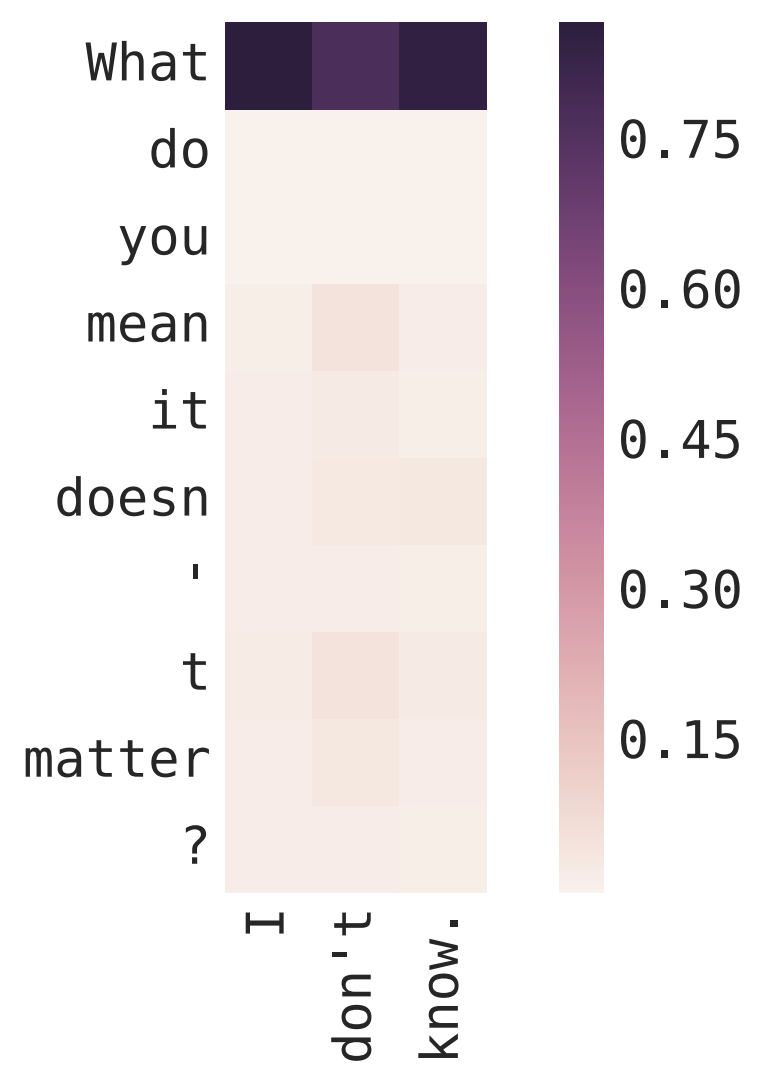
- Black-box: seq2seq + attention, trained on dialogue corpus

Interpretability for Flaw Detection in Dialogue Systems

[redacted]


I
don't
know.
What
do
you
mean
it
doesn't
matter?


What
do
you
mean
it
doesn
'
t
matter
?
I
don't
know.
0.75
0.60
0.45
0.30
0.15


What do you mean
It doesn't matter?
Neural Net
I don't know
Input
Prediction

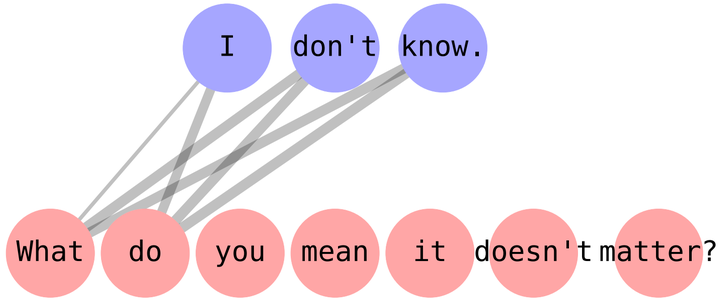
Explanation

Actual attention scores

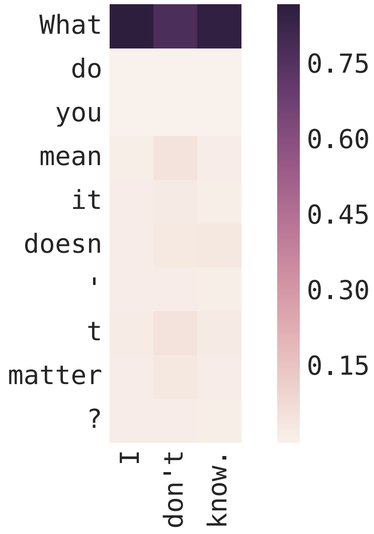
Interpretability for Flaw Detection in Dialogue Systems

- Black-box: seq2seq + attention, trained on dialogue corpus

Explanation



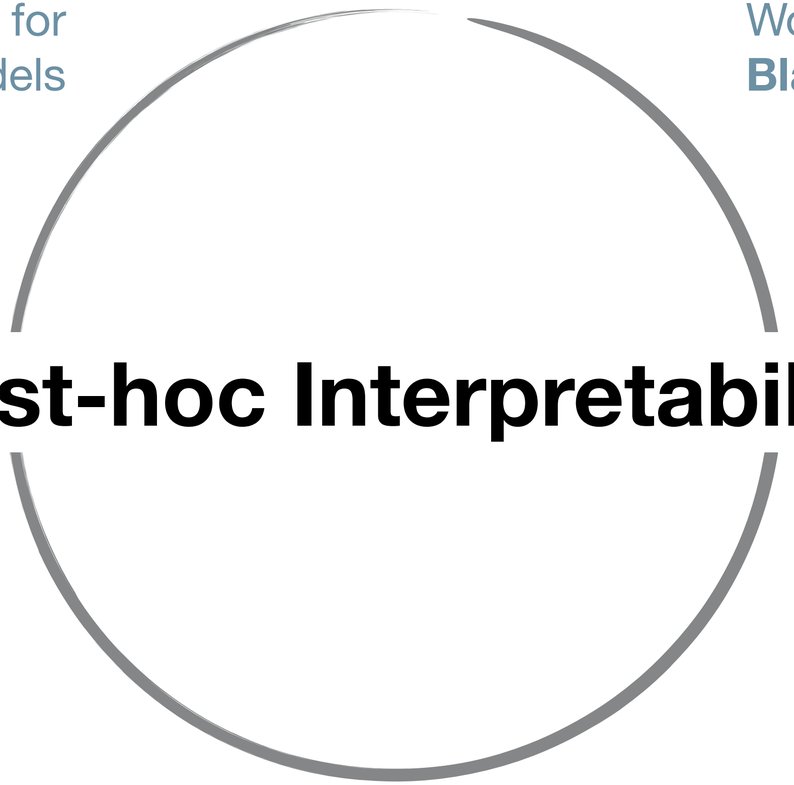
Actual attention scores



The model is flawed!

The only option for
already-trained models

Works for
Black-Box models



Post-hoc Interpretability