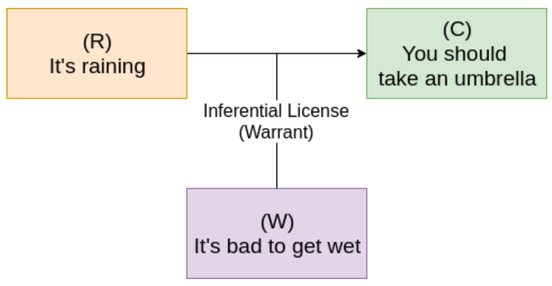
Probing Neural Network Comprehension of Natural Language Arguments

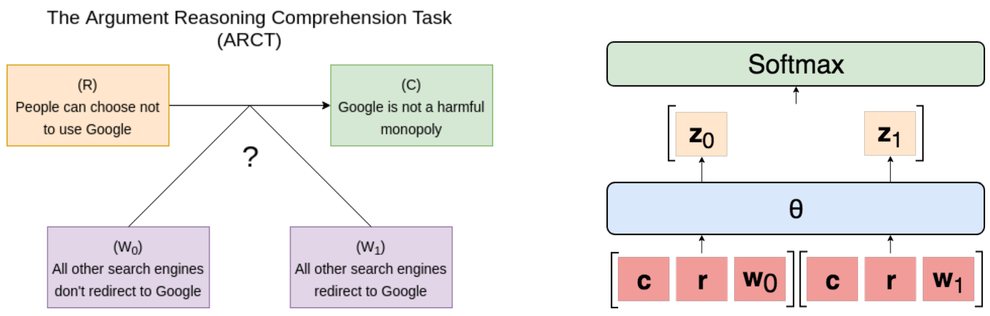
Timothy Niven and **Hung-Yu Kao**
Intelligent Knowledge Management Lab
National Cheng Kung University, Taiwan
tim.niven.public@gmail.com

How do you know two sentences connect argumentatively?



- Warrants provide a basis for making these connections
- Would have to *learn* and *reason* with warrants

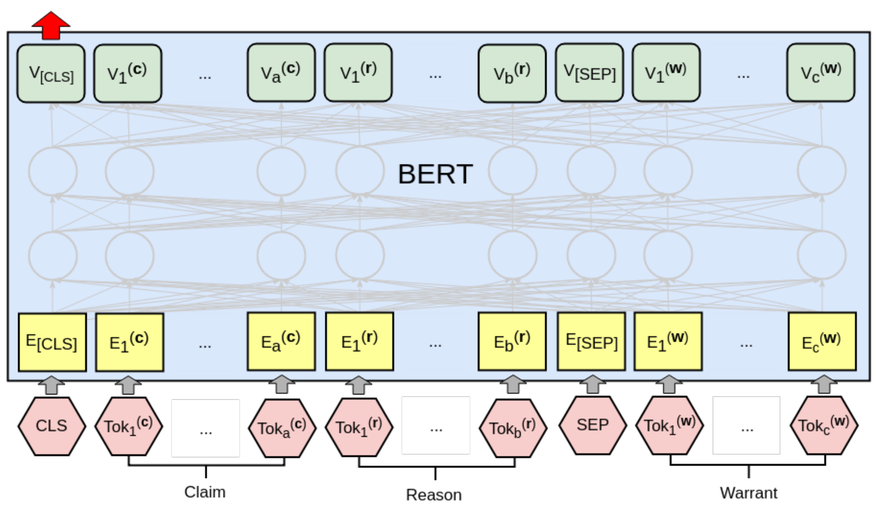
Even supplying warrants, reasoning with them is very hard



$$R \wedge W_0 \rightarrow C$$
$$R \wedge W_1 \rightarrow \neg C$$

- Binary classification with a small dataset (1,210 training data points)
- SemEval 2018 shared task systems found it hard to break 60% accuracy
- Still need significant world knowledge: how do web directs relate to the concept of monopoly in the context of search engines?

BERT is a strong learner



	Dev	Test		
	Mean	Mean	Median	Max
Human (trained)		0.909 ± 0.11		
Human (untrained)		0.798 ± 0.16		
BERT (Large)	0.701 ± 0.05	0.671 ± 0.09	**0.712**	**0.770**
Choi and Lee (2018)	$\mathbf{0.716} \pm 0.01$	$\mathbf{0.711} \pm 0.01$		
BERT (Base)	0.680 ± 0.02	0.623 ± 0.07	0.651	0.685
Botschen et al. (2018)	0.674 ± 0.01	0.568 ± 0.03		0.610
BoV	0.639 ± 0.02	0.564 ± 0.02	0.569	0.595
BiLSTM	0.658 ± 0.01	0.552 ± 0.02	0.552	0.592

- 5/20 BERT Large runs were degenerate - non-skewed mean is 0.716 ± 0.04
- BERT's maximum performance of 77% is three points behind the average (untrained) human baseline
- But this does not seem reasonable since BERT lacks the required knowledge
- What has BERT learned?

There are spurious statistical signals in the dataset

- Let k be some heuristic, such as a unigram or bigram
- $\mathbb{T}_j^{(i)}$ is the set of tokens in warrant j
- $y^{(i)} \in \{0, 1\}$ is the label

$k =$ "not"	Productivity	Coverage
Train	0.65	0.66
Validation	0.62	0.44
Test	0.52	0.77
All	**0.61**	**0.64**

$$\alpha_k = \sum_{i=1}^{n} \mathbb{1}\left[\exists j, k \in \mathbb{T}_j^{(i)} \wedge k \notin \mathbb{T}_{\neg j}^{(i)}\right]$$
$$\text{productivity} = \frac{\sum_{i=1}^{n} \mathbb{1}\left[\exists j, k \in \mathbb{T}_j^{(i)} \wedge k \notin \mathbb{T}_{\neg j}^{(i)} \wedge y^{(i)} = j\right]}{\alpha_k}$$
$$\text{coverage} = \frac{\alpha_k}{n}$$

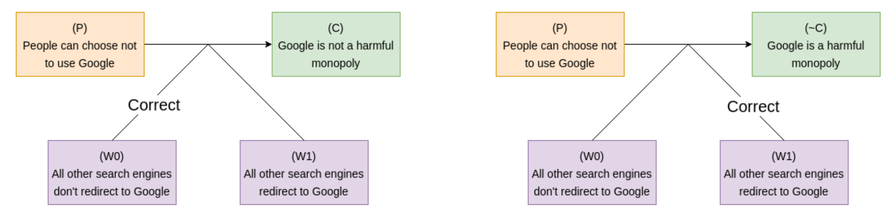
- *Productivity* measures how often you are rewarded for relying on a cue
- *Coverage* measures how strong the signal is in the data
- The strongest cue is "not"
- A great many more exist, albeit at much lower productivity for any decent coverage

All models exploit these spurious statistics; BERT best of all

	Test		
	Mean	Median	Max
BERT	$\mathbf{0.671} \pm 0.09$	**0.712**	**0.770**
BERT (W)	0.656 ± 0.05	0.675	0.712
BERT (R, W)	0.600 ± 0.10	0.574	0.750
BERT (C, W)	0.532 ± 0.09	0.503	0.732
BoV	0.564 ± 0.02	0.569	0.595
BoV (W)	0.567 ± 0.02	0.572	0.606
BoV (R, W)	0.554 ± 0.02	0.557	0.579
BoV (C, W)	0.545 ± 0.02	0.544	0.589
BiLSTM	0.552 ± 0.02	0.552	0.592
BiLSTM (W)	0.550 ± 0.02	0.547	0.577
BiLSTM (R, W)	0.547 ± 0.02	0.551	0.577
BiLSTM (C, W)	0.552 ± 0.02	0.550	0.601

- (W) just considers warrants (like a hypothesis-only NLI baseline)
- (R, W) additionally considers the reason
- (C, W) considers the claim and warrant
- Each of these setups breaks the task and reflects learning of spurious statistics
- The entirety of BERT's best 77% can be accounted for by these degenerate setups

We can eliminate the main problem in the dataset



- We create adversarial examples by negating each claim and flipping the label
- The major source of spurious signal is eliminated by mirroring around the labels

The new state of the art is random

	Test		
	Mean	Median	Max
BERT	$\mathbf{0.504} \pm 0.01$	**0.505**	**0.533**
BERT (W)	0.501 ± 0.00	0.501	0.502
BERT (R, W)	0.500 ± 0.00	0.500	0.502
BERT (C, W)	0.501 ± 0.01	0.500	0.518

Conclusion

- ARCT remains very difficult
- The adversarial dataset should be the standard in future work
- BERT is a strong learner capable of picking up on very subtle statistical cues
- As our models get stronger, the issue of spurious statistics becomes more pressing