Enhancing Common Sense Comprehension of Language Models

19 ACL & ???

이잉걸

Data Mining & Information Systems Lab.
Department of Computer Science and Engineering,
College of Informatics, Korea University

Probing Neural Network Comprehension of Natural language Arguments

Timothy Niven and Hung-Yu Kao (19 ACL short)

Introduction

- ✓ We are surprised to find that BERT's peak performance of 77% on the Argument Reasoning Comprehension Task (ARCT)
- ✓ Argumentation mining : (1) It is raining outside. (reason)
  (2) You should take an umbrella. (claim)
  (3) It is bad to get wet (warrant)
- ✓ Knowing (3) facilitates drawing the inferential connection between (1) and (2).

Introduction

- ✓ We are surprised to find that BERT's peak performance of 77% on the Argument Reasoning Comprehension Task (ARCT)
- ✓ Argumentation mining : (1) It is raining outside. (reason)
  (2) You should take an umbrella. (claim)
  (3) It is bad to get wet (warrant)
- ✓ Knowing (3) facilitates drawing the inferential connection between (1) and (2).

Claim	Google is not a harmful monopoly
Reason	People can choose not to use Google
Warrant	Other search engines don't redirect to Google
Alternative	All other search engines redirect to Google

Reason (and since) **Warrant** $\rightarrow$ **Claim**
Reason (but since) **Alternative** $\rightarrow \neg$ **Claim**

Figure 1: An example of a data point from the ARCT test set and how it should be read. The inference from R and A to $\neg C$ is by design.

Introduction

- ✓ We are surprised to find that BERT's peak performance of 77% on the Argument Reasoning Comprehension Task (ARCT)
- ✓ Argumentation mining : (1) It is raining outside. (reason)
  (2) You should take an umbrella. (claim)
  (3) It is bad to get wet (warrant)
- ✓ Knowing (3) facilitates drawing the inferential connection between (1) and (2).

Claim	Google is not a harmful monopoly
Reason	People can choose not to use Google
Warrant	Other search engines don't redirect to Google
Alternative	All other search engines redirect to Google

Reason (and since) **Warrant** $\rightarrow$ **Claim**
Reason (but since) **Alternative** $\rightarrow \neg$ **Claim**

Figure 1: An example of a data point from the ARCT test set and how it should be read. The inference from R and A to $\neg C$ is by design.

- ✓ Even supplying warrants, learners still need to rely on further world knowledge.
- ✓ For Fig 1, to correctly classify the data point, it is at least required to know how consumer choice and web re-directs relate to the concept of monopoly and that Google is a search engine.

Task Description and Baselines

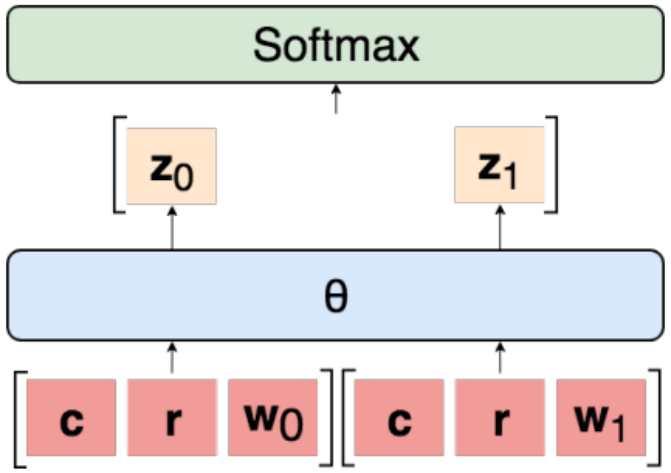
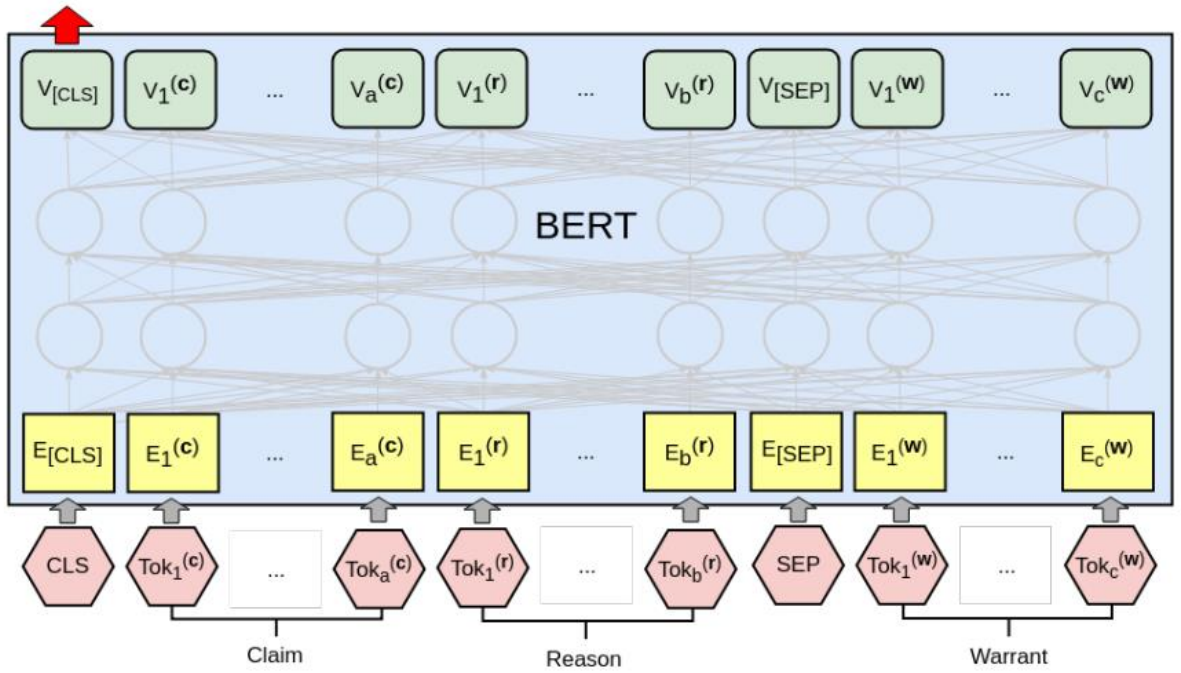


Figure 2: General architecture of the models in our experiments. Logits are independently calculated for each argument-warrant pair then concatenated and passed through softmax.



Task Description and Baselines

	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**
GIST (Choi and Lee, 2018)	**0.716** ± 0.01	**0.711** ± 0.01		
BERT (Base)	0.680 ± 0.02	0.623 ± 0.07	0.651	0.685
World Knowledge (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

Table 1: Baselines and BERT results. Our results come from 20 different random seeds ($\pm$ gives the standard deviation). The mean for BERT Large is skewed by the $5/20$ random seeds for which it failed to train, a problem noted by Devlin et al. (2018). We therefore consider the median a better measure of BERT's average performance. The mean of the non-degenerate runs for BERT (Large) is 0.716 ± 0.04.

- ✓ Only three points below the average(untrained) human baseline
- ✓ Without supplying the required world knowledge for this task it does not seem reasonable to expect it to perform so well.

- ✓ Question : what has BERT learned about argument comprehension?

Statistical Cues

- ✓ The major source of spurious statistical cues in ARCT comes from uneven distributions of linguistic artifacts over the warrants, and therefore over the labels.
- ✓ Aim to calculate how beneficial it is for a model to exploit a cue *k*, and how pervasive it is in the dataset.
- ✓ *applicability* α_k $$\alpha_k = \sum_{i=1}^{n} \mathbb{1}\left[\exists j, k \in \mathbb{T}_j^{(i)} \wedge k \notin \mathbb{T}_{\neg j}^{(i)}\right]$$
- ✓ *productivity* π_k $$\pi_k = \frac{\sum_{i=1}^{n} \mathbb{1}\left[\exists j, k \in \mathbb{T}_j^{(i)} \wedge k \notin \mathbb{T}_{\neg j}^{(i)} \wedge y_i = j\right]}{\alpha_k}$$
- ✓ *coverage* ξ_k $$\xi_k = \alpha_k / n$$

Statistical Cues

	Productivity	Coverage
Train	0.65	0.66
Validation	0.62	0.44
Test	0.52	0.77
All	**0.61**	**0.64**

Table 2: Productivity and coverage of using the presence of "not" in the warrant to predict the label in ARCT. Across the whole dataset, if you pick the warrant with "not" you will be right 61% of the time, which covers 64% of all data points.

- ✓ Unigram cue ("not") : it is just one among many such cues.
- ✓ Found a range of other unigrams, albeit with less overall productivity, mostly being high frequency words such as "is," "do," and "are."
- ✓ Bigrams that occurred with not, such as "will not" and "cannot," were also found to be highly productive.

Probing Experiments

	Test		
	Mean	**Median**	**Max**
BERT	**0.671** ± 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

- ✓ If a model is exploiting distributional cues over the labels, then if trained only on the warrants (W) it shoud perform relatively well.
- ✓ Based on this evidence our major finding is that the entirety of BERT's performance can be accounted for in terms of exploiting spurious statistical cues.

Adversarial Test Set

	Original	Adversarial
Claim	Google is not a harmful monopoly	Google is a harmful monopoly
Reason	People can choose not to use Google	People can choose not to use Google
Warrant	Other search engines do not redirect to Google	All other search engines redirect to Google
Alternative	All other search engines redirect to Google	Other search engines do not redirect to Google

Figure 4: Original and adversarial data points. The claim is negated and the warrants are swapped. The assignment of labels to W and A are kept the same. By including both, the distribution of linguistic artifacts in the warrants are thereby mirrored around the labels, eliminating the major source of spurious statistical cues in ARCT.

- ✓ This eliminates the problem by mirroring the distributions of cues around both labels.
- ✓ We trained on the original data(augmented by adding a copy of each data point with adversarial version), and validated and tested on the adversarial data.

Adversarial Test Set

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

- ✓ This eliminates the problem by mirroring the distributions of cues around both labels.
- ✓ We trained on the original data(augmented by adding a copy of each data point with adversarial version), and validated and tested on the adversarial data.
- ✓ Training on the negated data does lead to above random performance, but this is due to exploiting common statistics holding over claims and warrants.

How Reasonable are Common-Sense Reasoning Tasks: A Case-Study on the Winograd Schema Challenge and SWAG

Paul Trichelari et al. (19 EMNLP short)

Abductive Commonsense Reasoning

Chandra bhagavatula et al.