Probing Neural Network Comprehension of Natural Language Arguments

Timothy Niven and **Hung-Yu Kao**

Intelligent Knowledge Management Lab
Department of Computer Science and Information Engineering
National Cheng Kung University
Tainan, Taiwan
`tim.niven.public@gmail.com, hykao@mail.ncku.edu.tw`

Abstract

We are surprised to find that BERT's peak performance of 77% on the Argument Reasoning Comprehension Task reaches just three points below the average untrained human baseline. However, we show that this result is entirely accounted for by exploitation of spurious statistical cues in the dataset. We analyze the nature of these cues and demonstrate that a range of models all exploit them. This analysis informs the construction of an adversarial dataset on which all models achieve random accuracy. Our adversarial dataset provides a more robust assessment of argument comprehension and should be adopted as the standard in future work.

1 Introduction

Argumentation mining is the task of determining argumentative structure in natural language text - e.g., which text segments represent claims, and which comprise reasons that support or attack those claims (Mochales and Moens, 2011; Lippi and Torroni, 2016). This is a challenging task for machine learners, as it can be hard even for humans to determine when two text segments stand in argumentative relation, as evidenced by studies on argument annotation (Habernal et al., 2014).

One approach to this problem is to focus on *warrants* (Toulmin, 1958) - a form of world knowledge that permit inferences. Consider a simple argument: "(1) It is raining; therefore (2) you should take an umbrella."[1] The warrant "(3) it is bad to get wet" could license this inference. Knowing (3) facilitates drawing the inferential connection between (1) and (2). However it would be hard to find it stated anywhere since warrants are most often left implicit (Walton, 2005). Thus, on this approach, machine learners must not only reason with warrants but also discover them.

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** → **Claim**
Reason (but since) **Alternative** → ¬ **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.

The Argument Reasoning Comprehension Task (ARCT) (Habernal et al., 2018a) defers the problem of discovering warrants and focuses on inference. An argument is provided, comprising a claim C and reason R. This task is to pick the correct warrant W over a distractor, called the *alternative warrant* A. The alternative is written such that $R \wedge A \rightarrow \neg C$. An alternative warrant for our earlier example could be "(4) it is good to get wet," in which case we have $(1) \wedge (4) \rightarrow$ "($\neg 2$) you shouldn't take an umbrella." An example from the dataset is given in Figure 1.

The ARCT SemEval shared task (Habernal et al., 2018b) verified the challenging nature of this problem. Even supplying warrants, learners still need to rely on further world knowledge. For example, to correctly classify the data point in Figure 1 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. All but one participating system in the shared task could not exceed 60% accuracy (on binary classification).

It is therefore surprising that BERT (Devlin et al., 2018) achieves 77% test set accuracy with its best run (Table 1), 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. This motivates the question: what has BERT learned about argument comprehension?

---

[1]This example adapted from Black and Hunter (2012)

	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.
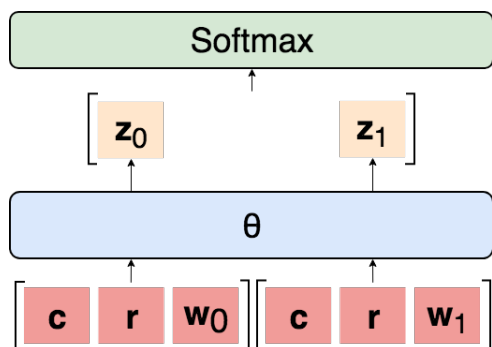

Figure 2: General architecture of the models in our experiments. Logits are independently calculated for each argument-warrant pair then concatenated and passed through softmax.

To investigate BERT's decision making we looked at data points it finds easy to classify over multiple runs. Habernal et al. (2018b) performed a similar analysis with the SemEval submissions, and consistent with their results we found that BERT exploits the presence of cue words in the warrant, especially "not." Through probing experiments designed to isolate such effects, we demonstrate in this work that BERT's surprising performance can be entirely accounted for in terms of exploiting spurious statistical cues.

However, we show that the major problem can be eliminated in ARCT. Since $R \wedge A \rightarrow \neg C$, we can add a copy of each data point with the claim negated and the label inverted. This means that the distribution of statistical cues in the warrants will be mirrored over both labels, eliminating the signal. On this adversarial dataset all models perform randomly, with BERT achieving a maximum test set accuracy of 53%. The adversarial dataset therefore provides a more robust evaluation of argument comprehension and should be adopted as the standard in future work on this dataset.

2 Task Description and Baselines

Let $i = 1, \ldots, n$ index each point in the dataset $\mathcal{D}$, where $|\mathcal{D}| = n$. The two candidate warrants in each case are randomly assigned a binary label $j \in \{0, 1\}$, such that each has an equal probability of being correct. The inputs are the representations for the claim $\mathbf{c}^{(i)}$, reason $\mathbf{r}^{(i)}$, warrant zero $\mathbf{w}_0^{(i)}$, and warrant one $\mathbf{w}_1^{(i)}$. The label $y^{(i)}$ is a binary indicator corresponding to the correct warrant.

The general architecture for all models is given in Figure 2. Shared parameters $\boldsymbol{\theta}$ are learned to classify each warrant independently with the argument, yielding the logits:

$$z_j^{(i)} = \boldsymbol{\theta}[\boldsymbol{c}^{(i)}; \boldsymbol{r}^{(i)}; \boldsymbol{w}_j^{(i)}]$$

These are then concatenated and passed through softmax to determine a probability distribution over the two warrants $\mathbf{p}^{(i)} = \text{softmax}([z_0^{(i)}, z_1^{(i)}])$. The prediction is then $\hat{y}^{(i)} = \arg\max_j \mathbf{p}^{(i)}$.

The baselines are a bag of vectors (BoV), bidirectional LSTM (Hochreiter and Schmidhuber, 1997) (BiLSTM), the SemEval winner GIST (Choi and Lee, 2018), the best model of Botschen et al. (2018), and human performance (Table 1). For all of our experiments we use grid search to select hyperparameters, dropout regularization (Srivastava et al., 2014), and Adam (Kingma and Ba, 2014) for optimization. We anneal the learning rate by $1/10$ when validation accuracy drops. The final parameters come from the epoch with maximum validation accuracy. The BoV and BiLSTM inputs are 300-dimensional GloVe embeddings trained on 640B tokens (Pennington et al., 2014). Code to reproduce all experiments, and detailing all hyperparameters, is provided on GitHub.[2]
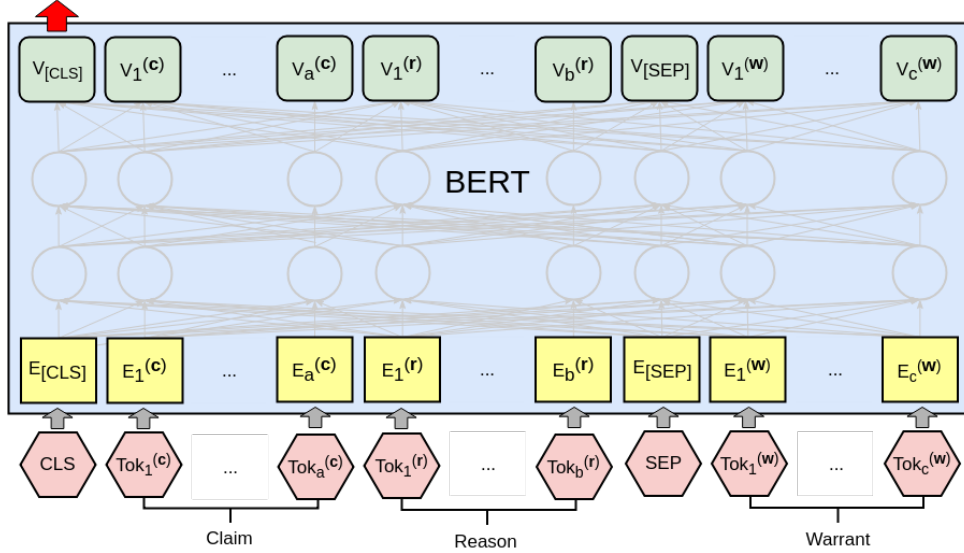
---

[2]https://github.com/IKMLab/arct2.git

4659

Figure 3: Processing an argument-warrant pair with BERT. The reason (with word pieces of length a) and claim (length b) together form the first utterance, and the warrant (length c) is the second. The final CLS vector is then passed to a linear layer to calculate the logit $z_j^{(i)}$.

3 BERT

Our BERT classifier is visualized in Figure 3. The claim and reason are joined to form the first text segment, which is paired with each warrant and independently processed. The final layer CLS vector is passed to a linear layer to obtain the logits $z_j^{(i)}$. The whole architecture is fine-tuned. The learning rate is $2e^{-5}$ and we allow a maximum of 20 training epochs, taking the parameters from the epoch with the best validation set accuracy. We use the Hugging Face PyTorch implementation.[3]

Devlin et al. (2018) report that, on small datasets, BERT sometimes fails to train, yielding degenerate results. ARCT is very small with $1,210$ training observations. In $5/20$ runs we encountered this phenomenon, seeing close to random accuracies on validation and test sets. These cases occurred where training accuracy was also not significantly above random ($< 80\%$). Removing the degenerate runs, BERT's mean is 71.6 ± 0.04., which would beat the previous state of the art - as would the median of 71.2%, which is a better average than the overall mean since it is not skewed by the degenerate cases. However, our main finding is that these results are not meaningful and should be discarded. In the following sections we focus on BERT's peak performance of 77% to make this case.

4 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. This section aims to demonstrate the presence and nature of these cues. We only consider unigrams and bigrams, although more sophisticated cues may be present. To this end, we aim to calculate how beneficial it is for a model to exploit a cue k, and how pervasive it is in the dataset (indicating the strength of the signal).

Formally, let $\mathbb{T}_j^{(i)}$ be the set of tokens in the warrant for data point i with label j. We define a cue's *applicability* α_k as the number of data points where it occurs with one label but not the other:

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

The *productivity* π_k of a cue is defined as the proportion of applicable data points for which it predicts the correct answer:

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

Finally, we define the *coverage* ξ_k of a cue as the proportion of applicable cases over the total number of data points: $\xi_k = \alpha_k/n$. In these terms, the productivity of a cue measures the benefit of exploiting it, while coverage measures the

	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.

strength of the signal it provides. With m labels, if $\pi_k > 1/m$ then the presence of a cue is going to be useful for the task and a machine learner would do well to make use of it.

The productivity and coverage of the strongest unigram cue we found ("not") is given in Table 2. It provides a particularly strong training signal. While it is less productive in the test set, it is just one among many such cues. We 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. These statistics indicate the nature of the problem. In the next section we demonstrate that our models are in fact exploiting these cues.

5 Probing Experiments

If a model is exploiting distributional cues over the labels, then if trained only on the warrants (W) it should perform relatively well. The same can be said for removing either just the claim, leaving the reason and warrant (R, W), or removing the reason (C, W). The latter setups allow the models to additionally consider cues in the reasons and claims, as well as cues holding over their combinations with the warrants. Each of these setups breaks the task since we no longer have an argument to match with a warrant.

Experimental results are given in Table 3. On warrants alone (W) BERT achieves a maximum 71% accuracy. That leaves only six percentage points to account for its peak of 77%. We find a gain of four percentage points for (R, W) over (W), and a gain of two for (C, W), accounting for the missing six points. 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.

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

Table 3: Results of probing experiments with BERT Large, and the BoV and BiLSTM baselines. These results indicate that BERT's peak 77% performance can be entirely accounted for by exploiting spurious cues. By just considering warrants (W) we can get to 71%. Adding cues over reasons (R, W) and claims (C, W) accounts for the remaining six points.

6 Adversarial Test Set

The major problem of statistical cues over labels in ARCT can be eliminated due the original design of the dataset. Given that $R \wedge A \rightarrow \neg C$, we can produce adversarial examples by negating the claim and inverting the label for each data point (Figure 4). The adversarial examples are then combined with the original data. This eliminates the problem by mirroring the distributions of cues around both labels. The ARCT authors provide a training set augmented in this way. The negation of most claims in the validation and test sets already exist elsewhere in the dataset. The remaining claims were manually negated by a native English speaker.

We tried two experimental setups. In the first, models trained and validated on the original data were evaluated on the adversarial set. All results were worse than random due to overfitting the cues in the original training set. In the second, models were trained from scratch on the adversarial training and validation sets, then evaluated on the adversarial test set. Results are given in Table 4. BERT's peak performance has reduced to 53%, with mean and median at 50%. We conclude from these results that the adversarial dataset has successfully eliminated the cues as expected, providing a more robust evaluation of machine argument comprehension. This result better apts with our intuitions about this task: with little to no understanding about the reality underlying these arguments, good performance shouldn't be feasible.

4661

	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.

7 Related Work

The most successful previous work on ARCT (Choi and Lee, 2018; Zhao et al., 2018; Niven and Kao, 2018) involved transfer learning from Natural Language Inference (NLI) datasets (Bowman et al., 2015; Williams et al., 2017), and utilized effective NLI models such as ESIM (Chen et al., 2016) and InferSent (Conneau et al., 2017). More recently, Botschen et al. (2018) added FrameNet knowledge with modest performance gains. These models should be evaluated on our adversarial dataset. In particular it will be interesting if Botschen et al.'s model stands out due to the inclusion of some of the required world knowledge.

There is much recent work focusing on statistical cues in datasets in vision (Jo and Bengio, 2017) and NLP (Sanchez et al., 2018; McCoy et al., 2019; Gururangan et al., 2018; Glockner et al., 2018; Poliak et al., 2018; Rajpurkar et al., 2018; Jia and Liang, 2017). Similar to our experiment with warrants, Poliak et al. (2018) classified NLI data based on the hypothesis only. A similar experiment to our probing task was performed by Niven and Kao (2018), but only with reasons and warrants. They found that independent warrant classification with shared parameters provides some regularization against warrant-label cues (Niven and Kao, 2018). However, this does not solve the problem since the presence of a cue is enough to increase the logits for either warrant.

The original ARCT data comes with a *training* set created in the same way as our adversarial dataset. Habernal et al. (2018a) reported experiments using this training data that led to random accuracy. They suggested it could be that high similarity between the data points made the problem too difficult for the simple models they implemented. Our work indicates the necessity of applying this transformation to the entire dataset in order to obtain a more robust evaluation by eliminating spurious statistical cues over the labels.

	Test		
	Mean	**Median**	**Max**
BERT	**0.504** ± 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

Table 4: Results for BERT Large on the adversarial test set with adversarial training and validation sets.

8 Conclusion

ARCT provides a fortuitous opportunity to see how stark the problem of exploiting spurious statistics can be. Due to our ability to eliminate the major source of these cues, we were able to show that BERT's maximum performance fell from just three points below the average untrained human baseline to essentially random. To answer our question in the introduction: BERT has learned nothing about argument comprehension.

However, our investigations confirmed that BERT is indeed a very strong learner. Analysis of easy to classify data points showed reliance on a *lower* proportion of the strongest cue word than the BoV and BiLSTM - i.e. BERT has learned when to ignore the presence of "not" and focus on different cues. This indicates an ability to exploit much more subtle joint distributional information. As our learners get stronger, controlling for spurious statistics becomes more important in order to have confidence in their apparent performance. Taken with a growing body of previous work, our results indicate the need for further research into the extent of this problem in NLP more generally.

The adversarial dataset should be adopted as the standard in future work on ARCT. We hope that providing a more robust evaluation will help to spur more productive research on this problem.

Acknowledgments

References

Elizabeth Black and Anthony Hunter. 2012. A relevance-theoretic framework for constructing and deconstructing enthymemes. *J. Log. Comput.*, 22:55–78.

Teresa Botschen, Daniil Sorokin, and Iryna Gurevych. 2018. Frame- and entity-based knowledge for common-sense argumentative reasoning. In *ArgMining@EMNLP*.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. *CoRR*, abs/1508.05326.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, and Hui Jiang. 2016. Enhancing and combining sequential and tree LSTM for natural language inference. *CoRR*, abs/1609.06038.

HongSeok Choi and Hyunju Lee. 2018. Gist at semeval-2018 task 12: A network transferring inference knowledge to argument reasoning comprehension task. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 773–777. Association for Computational Linguistics.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. *CoRR*, abs/1705.02364.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. Breaking NLI systems with sentences that require simple lexical inferences. *CoRR*, abs/1805.02266.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. *CoRR*, abs/1803.02324.

Ivan Habernal, Judith Eckle-Kohler, and Iryna Gurevych. 2014. Argumentation mining on the web from information seeking perspective. In *ArgNLP*.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018a. The argument reasoning comprehension task: Identification and reconstruction of implicit warrants. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1930–1940, New Orleans, Louisiana. Association for Computational Linguistics.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018b. Semeval-2018 task 12: The argument reasoning comprehension task. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 763–772. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. *CoRR*, abs/1707.07328.

Jason Jo and Yoshua Bengio. 2017. Measuring the tendency of cnns to learn surface statistical regularities. *CoRR*, abs/1711.11561.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Marco Lippi and Paolo Torroni. 2016. Argumentation mining: State of the art and emerging trends. *ACM Trans. Internet Technol.*, 16(2):10:1–10:25.

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. *CoRR*, abs/1902.01007.

Raquel Mochales and Marie-Francine Moens. 2011. Argumentation mining. *Artif. Intell. Law*, 19(1):1–22.

Timothy Niven and Hung-Yu Kao. 2018. NLITrans at SemEval-2018 task 12: Transfer of semantic knowledge for argument comprehension. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 1099–1103, New Orleans, Louisiana. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. *CoRR*, abs/1805.01042.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. *CoRR*, abs/1806.03822.

Ivan Sanchez, Jeff Mitchell, and Sebastian Riedel. 2018. Behavior analysis of NLI models: Uncovering the influence of three factors on robustness. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1975–1985, New Orleans, Louisiana. Association for Computational Linguistics.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958.

Stephen E. Toulmin. 1958. *The Uses of Argument*. Cambridge University Press.

Douglas N. Walton. 2005. *Informal logic: a handbook of critical argumentation*. Cambridge University Press.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. *CoRR*, abs/1704.05426.

Meiqian Zhao, Chunhua Liu, Lu Liu, Yan Zhao, and Dong Yu. 2018. Blcu_nlp at semeval-2018 task 12: An ensemble model for argument reasoning based on hierarchical attention. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 1104–1108. Association for Computational Linguistics.