An Exploratory Study of Argumentative Writing by Young Students: A Transformer-based Approach

Debanjan Ghosh, Beata Beigman Klebanov, Yi Song
Educational Testing Service
dghosh,bbeigmanklebanov,ysong@ets.org

Abstract

We present a computational exploration of argument critique writing by young students. Middle school students were asked to criticize an argument presented in the prompt, focusing on identifying and explaining the reasoning flaws. This task resembles an established college-level argument critique task. Lexical and discourse features that utilize detailed domain knowledge to identify critiques exist for the college task but do not perform well on the young students data. Instead, transformer-based architecture (e.g., BERT) fine-tuned on a large corpus of critique essays from the college task performs much better (over 20% improvement in F1 score). Analysis of the performance of various configurations of the system suggests that while children's writing does not exhibit the standard discourse structure of an argumentative essay, it does share basic local sequential structures with the more mature writers.

1 Introduction

Argument and logic are essential in academic writing as they enhance the critical thinking capacities of students. Argumentation requires systematic reasoning and the skill of using relevant examples to craft a support for one's point of view (Walton, 1996). In recent times, the surge in AI-informed scoring systems has made it possible to assess writing skills using automated systems. Recent research suggests the possibility of argumentation-aware automated essay scoring systems (Stab and Gurevych, 2017b).

Most of the current work on computational analysis of argumentative writing in educational context focuses on automatically identifying the argument structures (e.g., argument components and their relations) in the essays (Stab and Gurevych, 2017a; Persing and Ng, 2016; Nguyen and Litman, 2016) and by predicting essay scores from features derived from the structures (e.g., the number of claims and premises and the number of supported claims) (Ghosh et al., 2016). Related research has also addressed the problem of scoring a particular dimension of essay quality, such as relevance to the prompt (Persing and Ng, 2014), opinions and their targets (Farra et al., 2015), argument strength (Persing and Ng, 2015), among others.

While argument mining literature has addressed the educational context, it has so far mainly focused on analyzing college-level writing. For instance, Nguyen and Litman (2018) investigated argument structures in TOEFL11 corpus (Blanchard et al., 2013); Beigman Klebanov et al. (2017) and Persing and Ng (2015) analyzed writing of university students; Stab and Gurevych (2017b) used data from "essayforum.com", where college entrance examination is the largest forum. Computational analysis of arguments in young students' writing has not yet been done, to the best of our knowledge. Writing quality in essays by young writers has been addressed (Deane, 2014; Attali and Powers, 2008; Attali and Burstein, 2006), but identification of arguments was not part of these studies.

In this paper, we present a novel learning-and-assessment context where middle school students were asked to criticize an argument presented in the prompt, focusing on identifying and explaining the reasoning flaws. Using a relatively small pilot data collected for this task, our aim here is to automatically identify good argument critiques in the young students' writing, with the twin goals of (a) exploring the characteristics of young students' writing for this task, and (b) in view of potential scoring and feedback applications. We start with describing and exemplifying the data, as well as the argument critique annotation we performed on it (section 2). Experiments and results are pre-

Proceedings of the 15th Workshop on Innovative Use of NLP for Building Educational Applications, pages 145–150

Dear Editor,

Advertising aimed at children under 12 should be allowed for several reasons.

First, one family in my neighbourhood sits down and watches TV together almost every evening. The whole family learns a lot, which shows that advertising for children is always a good thing because it brings families together.

Second, research shows that children can't remember commercials well anyway, so they can't be doing kids any harm.

Finally, the arguments against advertising aren't very effective. Some countries banned ads because kids thought the ads were funny. But that's not a good reason. Think about it: the advertising industry spends billions of dollars a year on ads for children. They wouldn't spend all the money if the ads weren't doing some good. Let's not hurt children by stopping a good thing.

If anyone doesn't like children's ads, the advertisers should just try to make them more interesting. The ads are allowed to be shown on TV, so they shouldn't be banned.

Table 1: The prompt of the argument critique task.

sented in section 3, followed by a discussion in section 4.

2 Dataset and Annotation

The data used in this study was collected as part of a pilot of a scenario-based assessment of argumentation skills with about 900 middle school students (Song et al., 2017).[1] Students engaged in a sequence of steps in which they researched and reflected on whether advertising to children under the age of twelve should be banned. The test consists of four tasks; we use the responses to Task 3 in which students are asked to review a letter to the editor and evaluate problems in the letter's *reasoning* or use of *evidence* (see Table 1).

Students were expected to produce a written critique of the arguments, demonstrating their ability to identify and explain problems in the reasoning or use of evidence. For example, the first excerpt below shows a well-articulated critique of the hasty generalization problem in the prompt:

> (1) Just because it brings one family together to learn does not mean that it will bring all families together to learn.

> (2) The first one about the family in your neighborhood is more like an opinion, not actual information from the article.

> (3) Their claims are badly writtin [sic] and have no good arguments. They need to support their claims with SOLID evidence and only claim arguments that can be undecicive [sic].

However, many students had difficulty explaining the reasoning flaws clearly. In the second excerpt, the student thought that an argument from the family in the neighborhood is not strong, but did not demonstrate an understanding of a weak generalization in his explanation. Other common problems included students summarizing the prompt without criticizing, or providing a generic critique that does not adhere to the particulars of the prompt (excerpt (3)).

The goal of the argument critique annotation (described next) was to identify where in a response *good* critiques are made, such as the one in the first excerpt.

Annotation of Critiques: We identified 11 valid critiques of the arguments in the letter. These critiques included: (1) overgeneralizing from a single example; (2) example irrelevant to the argument; (3) example misrepresenting what actually happened; (4) misrepresenting the goal of making advertisements; (5) misunderstanding the problem; (6) neglecting potential side effects of allowing advertising aimed at children; (7) making a wrong argument from sign; (8) argument contradicting authoritative evidence; (9) argument contradicting one's own experience; (10) making a circular argument; (11) making contradictory claims. All sentences containing any material belonging to a valid critique were marked and henceforth denoted as Arg; the rest are denoted as $NoArg$. Three annotators were employed to annotate the sentences to mark them as Arg/$NoArg$. We computed κ between each pair of annotators based on the annotation of 50 essays. Inter-annotator agreement for this sentence-level Arg/$NoArg$ classification for each pair of annotators was 0.714, 0.714, and 0.811, respectively resulting in an average κ of 0.746.

Descriptive statistics: We split the data into *training* (585 response critiques) and *test* (252 response critiques). The *training* partition has 2,220 sentences (515 Arg; 1,705 $NoArg$; average number of words per sentence is 11 (std = 8.03)); *test* contains 973 sentences.

[1] The data was collected under the ETS CBAL (Cognitively Based Assessment of, for, and as Learning) Initiative.

3 Experiments and Results

3.1 Baseline

In this writing task, young students were asked to analyze the given prompt, focusing on identifying and explaining its reasoning flaws. This task is similar to a well-established task for college students previously discussed in the literature (Beigman Klebanov et al., 2017). Compared to the college task, the prompt for children appears to have more obvious reasoning errors. The tasks also differ in the types of responses they elicit. While the college task elicits a full essay-length response, the current critique task elicits a shorter, less formal response.

As our baseline, we evaluate the features that were reported as being effective for identifying argument critiques in the context of the college task. Beigman Klebanov et al. (2017) described a logistic regression classifier with two types of features:

- features capturing discourse **structure**, since it was found that argument critiques tended to occupy certain consistent discourse roles that are common in argumentative essays (such as the SUPPORT, rather than THESIS or BACKGROUND roles), as well as have a tendency to participate in roles that receive a lot of elaboration, such as a SUPPORT sentence following or preceding another SUPPORT sentence, or a CONCLUSION sentence followed by another sentence in the same role.

- features capturing **content**, based on hybrid word and POS ngrams (see Beigman Klebanov et al. (2017) for more detail).

Table 2 shows the results, with each of the two subsets of features separately and together. Clearly, the classifier performs quite poorly for detecting Arg sentences in children's data. Secondly, it seems that whatever performance is achieved is due to the content features, while the structural features fail to detect Arg. Thus, the well-organized nature of the mature writing, where essays have identifiable discourse elements such as THESIS, MAIN CLAIM, SUPPORT, CONCLUSION (Burstein et al., 2003), does not seem to carry over to young students' less formal writing.

Features	Category	Precision	Recall	F1
Content	$NoArg$	0.851	0.946	0.896
	Arg	0.611	0.338	0.436
Structure	$NoArg$	0.799	1.00	0.889
	Arg	0	0	0
Structure + Content	$NoArg$	0.852	0.940	0.894
	Arg	0.591	0.349	0.439

Table 2: Performance of baseline features. "Structure" corresponds to the *dr_pn* feature set, "Content" corresponds to the *1-3gr ppos* feature set, both from Beigman Klebanov et al. (2017).

3.2 Our system

As the training dataset is relatively small, we leverage pre-trained language models that are shown to be effective in various NLP applications. Particularly, we focus on BERT (Devlin et al., 2018), a bi-directional transformer (Vaswani et al., 2017) based architecture that has produced excellent performance on argumentation tasks such as argument component and relation identification (Chakrabarty et al., 2019) and argument clustering (Reimers et al., 2019). The BERT model is initially trained over a 3.3 billion word English corpus on two tasks: (1) given a sentence containing multiple masked words predict the identity of a particular masked word, and (2) given two sentences, predict whether they are adjacent. The BERT model exploits a multi-head attention operation to compute context-sensitive representations for each token in a sentence. During its training, a special token "[CLS]" is added to the beginning of each training utterance. During evaluation, the learned representation for this "[CLS]" token is processed by an additional layer with nonlinear activation. A standard pre-trained BERT model can be used for transfer learning when the model is "fine-tuned" during training, i.e., on the classification data of Arg and $NoArg$ sentences (i.e., *training* partition) or by first fine-tuning the BERT language-model itself on a large unsupervised corpus from a partially relevant domain, such as a corpus of writings from advanced students and then again fine-tuned on the classification data. In both the cases, BERT makes predictions via the "[CLS]" token.

Fine-tuning on classification data: We first fine-tune a pre-trained BERT model (the "bert-base-uncased" version) with the $training$ data. During training the class weights are proportional to the numbers of Arg and $NoArg$ instances. Unless stated otherwise we kept the following parameters throughout in the experiments: we utilize a batch size of 16 instances, learning_rate of 3e-5, warmup_proportion 0.1, and the Adam optimizer.

Experiment	Category	Precision	Recall	F1
BERT$_{bl}$	$NoArg$	0.884	0.913	0.898
	Arg	0.603	0.523	0.560
BERT$_{pair}$	$NoArg$	0.892	0.934	0.913
	Arg	**0.681**	0.556	0.612
BERT$_{bl+lm}$	$NoArg$	0.907	0.898	0.902
	Arg	0.610	0.636	0.623
BERT$_{pair+lm}$	$NoArg$	0.929	0.871	0.900
	Arg	0.592	**0.740**	**0.658**

Table 3: Performance of BERT transformer, various configurations. Rows 1, 2 present results of BERT fine-tuning with $training$ data only; rows 3, 4 present the effect of additional language model fine-tuning. Highest scores are **bold**.

Hyperparameters were tuned for only five epochs. This experiment is denoted as BERT$_{bl}$ in Table 3. We observe that the F1 score for Arg is 56%, resulting in a 12% absolute improvement in F1 score over the structure+content features (Table 2). This confirms that BERT is able to perform well even after fine-tuning with a relatively small training corpus with default parameters.

In the next step, we re-utilize the same pre-trained BERT model while transforming the *training* instances to *paired* sentence instances, where the first sentence is the candidate Arg or $NoArg$ sentence and the second sentence of the pair is the immediate next sentence in the essay. For instance, for the first example in section 2, "Just because . . . to learn", now the instance also contains the subsequent sentence:

> <Just because . . . to learn.>,<Second, children can't remember commercials anyway, so they can't be doing any harm," says the letter.>

A special token "FINAL_SENTENCE" is used when the candidate Arg or $NoArg$ sentence is the last sentence in the essay. This modification of the data representation might help the BERT model for two reasons. First, pairing of the candidate sentence and the next one will encourage the model to more directly utilize the next sentence prediction task. Secondly, since multi-sentence same-discourse-role elaboration was found to be common in Beigman Klebanov et al. (2017) data, BERT may exploit such sequential structures if they at all exist in our data. This is model BERT$_{pair}$ in Table 3. With the paired-sentences transformation of the instances the F1 improves to 61.2%, a boost of 5% over BERT$_{bl}$.

Fine-tuning with a large essay corpus: It has been shown in related research (Chakrabarty et al., 2019) that transfer learning by fine-tuning on a domain-specific corpus using a supervised learning objective can boost performance. We used a large proprietary corpus of college-level argument critique essays similar to those analyzed by Beigman Klebanov et al. (2017). This corpus consists of 351,363 unannotated essays, where an average essay contains 16 sentences, resulting in a corpus of 5.64 million sentences. We fine-tune the pre-trained BERT language model on this large corpus for five epochs and then again fine-tune it with the $training$ partition (BERT$_{bl+lm}$). Likewise, BERT$_{pair+lm}$ represents the model after pre-trained BERT language model is fine-tuned with the large corpus and then again fine-tuned with the paired instances of $training$. We observe that fine-tuning the language model improves F1 to 62.3% whereas BERT$_{pair+lm}$ results in the highest F1 of 65.8%, around 5% higher than BERT$_{pair}$ and over 20% higher than the feature-based model.

4 Discussion

The difference in F1 between BERT$_{bl}$, BERT$_{bl+lm}$, and BERT$_{pairs+lm}$ is almost exclusively in recall – they have comparable precision at about 0.6, with recall of 0.52, 0.64, and 0.74, respectively. Partitioning out 10% of the *training* data for a *development* set, we found that BERT$_{bl+lm}$ detected 13 more Arg sentences than BERT$_{bl}$ in the development data. These fell into two **sequential** patterns: (a) the sentence is followed by another that further develops the critique (7 cases) – see excerpts (4) and (5) below; (b) the sentence is the final sentence in the response (6 cases); excerpt (6).

> (4) They werent made to be appealing to adults. They only need kids to want the product, and beg their parents for it.

> (5) Finally, is spending billions of dollars on something that has no point a good thing? There are many arguements that all this money is just going to waste, and it could be used on more important things.

> (6) I say this because in an article I found out that children do remember advertisements that they have seen before.

Our interpretation of this finding is that BERT$_{bl+lm}$ captured organizational elements in

children's writing that *are* similar to adult patterns. Beigman Klebanov et al. (2017) found that adult writers often reiterate a previously stated critique in an extended CONCLUSION and spread critiques across consecutive SUPPORT sentences. Thus, even though alignment of critiques with "standard" discourse elements such as CONCLUSION and SUPPORT is not recognizable in children's writing (as witnessed by the failure of the structural features to detect critiques), some **basic local sequential patterns** do exist, and they are sufficiently similar to the ones in adult writing that a system with its language model tuned on adult critique writing can capitalize on this knowledge.

Interestingly, BERT$_{pairs}$ learned similar sequential patterns – indeed 7 of the 13 sentences gained by BERT$_{bl+lm}$ over BERT$_{bl}$ are also recalled by BERT$_{pairs}$. This further reinforces the conclusion that young writers exhibit certain local sequential patterns of discourse organization that they share with mature argument critique writers.

5 Conclusion and Future Work

We present a computational exploration of argument critiques written by middle school children. A feature set designed for *college-level* critique writing has poor recall of critiques when trained on children's data; a pre-trained BERT model fine-tuned on children's data does better by 18%. When BERT's language model is additionally fine-tuned on a large corpus of *college* critique essays, recall improves by further 20%, suggesting the existence of *some* similarity between young and mature writers. Performance analysis suggests that BERT capitalized on certain sequential patterns in critique writing; a larger study examining patterns of argumentation in children's data is needed to confirm the hypothesis. In future, we plan to fine-tune our models on auxiliary dataset, such as the convincing argument dataset from Habernal and Gurevych (2016).

References

Yigal Attali and Jill Burstein. 2006. Automated essay scoring with e-rater v.2. *The Journal of Technology, Learning and Assessment*, 4(3).

Yigal Attali and Don Powers. 2008. A developmental writing scale. *ETS Research Report Series*, 2008(1):i–59.

Beata Beigman Klebanov, Binod Gyawali, and Yi Song. 2017. Detecting Good Arguments in a Non-Topic-Specific Way: An Oxymoron? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 244–249, Vancouver, Canada. Association for Computational Linguistics.

Daniel Blanchard, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow. 2013. Toefl11: A corpus of non-native english. *ETS Research Report Series*, 2013(2):i–15.

Jill Burstein, Daniel Marcu, and Kevin Knight. 2003. Finding the WRITE Stuff: Automatic Identification of Discourse Structure in Student Essays. *IEEE Intelligent Systems*, 18(1):32–39.

Tuhin Chakrabarty, Christopher Hidey, Smaranda Muresan, Kathleen Mckeown, and Alyssa Hwang. 2019. Ampersand: Argument mining for persuasive online discussions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2926–2936.

Paul Deane. 2014. Using writing process and product features to assess writing quality and explore how those features relate to other literacy tasks. *ETS Research Report Series*, 2014(1):1–23.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Noura Farra, Swapna Somasundaran, and Jill Burstein. 2015. Scoring persuasive essays using opinions and their targets. In *Proceedings of the Workshop on Innovative Use of NLP for Building Educational Applications*, pages 64–74.

Debanjan Ghosh, Aquila Khanam, Yubo Han, and Smaranda Muresan. 2016. Coarse-grained argumentation features for scoring persuasive essays. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 549–554, Berlin, Germany. Association for Computational Linguistics.

Ivan Habernal and Iryna Gurevych. 2016. Which argument is more convincing? analyzing and predicting convincingness of web arguments using bidirectional lstm. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1589–1599.

Huy Nguyen and Diane Litman. 2016. Context-aware argumentative relation mining. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1127–1137.

Huy V Nguyen and Diane J Litman. 2018. Argument mining for improving the automated scoring of persuasive essays. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Isaac Persing and Vincent Ng. 2014. Modeling prompt adherence in student essays. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1534–1543.

Isaac Persing and Vincent Ng. 2015. Modeling argument strength in student essays. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 543–552.

Isaac Persing and Vincent Ng. 2016. End-to-end argumentation mining in student essays. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1384–1394.

Nils Reimers, Benjamin Schiller, Tilman Beck, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. 2019. Classification and clustering of arguments with contextualized word embeddings. *arXiv preprint arXiv:1906.09821*.

Yi Song, Paul Deane, and Mary Fowles. 2017. Examining students' ability to critique arguments and exploring the implications for assessment and instruction. *ETS Research Report Series*, 2017(16):1–12.

Christian Stab and Iryna Gurevych. 2017a. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43(3):619–659.

Christian Stab and Iryna Gurevych. 2017b. Recognizing insufficiently supported arguments in argumentative essays. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 980–990, Valencia, Spain. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Douglas N Walton. 1996. *Argumentation schemes for presumptive reasoning*. Psychology Press.