Detecting Good Arguments in a Non-Topic-Specific Way: An Oxymoron?

Beata Beigman Klebanov, Binod Gyawali, Yi Song
Educational Testing Service
660 Rosedale Road
Princeton, NJ, USA
bbeigmanklebanov,bgyawali,ysong@ets.org

Abstract

Automatic identification of good arguments on a controversial topic has applications in civics and education, to name a few. While in the civics context it might be acceptable to create separate models for each topic, in the context of scoring of students' writing there is a preference for a single model that applies to all responses. Given that good arguments for one topic are likely to be irrelevant for another, is a single model for detecting good arguments a contradiction in terms? We investigate the extent to which it is possible to close the performance gap between topic-specific and across-topics models for identification of good arguments.

1 Introduction & Related Work

Argumentation is an important skill in higher education and the workplace; students are expected to show sound reasoning and use relevant evidence (Council of Chief State School Officers & National Governors Association, 2010). The increase in argumentative writing tasks, in both instructional and assessment contexts, results in a high demand for automated feedback on and scoring of arguments.

Automated analysis of argumentative writing has mostly concentrated on argument structure – namely, presence of claims and premises, and relationships between them (Ghosh et al., 2016; Nguyen and Litman, 2016; Persing and Ng, 2016; Ong et al., 2014; Stab and Gurevych, 2014). Addressing the content of arguments in on-line debates, Habernal and Gurevych (2016) ranked arguments on the same topic by convincingness; they showed that convincingness can be automatically predicted, to an extent, in a cross-topics fashion, as they trained their systems on 31 debates and tested on a new one. Swanson et al. (2015) reported that annotation of argument quality is challenging, with inter-annotator agreement (ICC) around 0.40. They also showed that automated across-topics prediction is very hard; for some topics, no effective prediction was achieved.

Song et al. (2014) developed an annotation protocol for analyzing argument critiques in students' essays, drawing on the theory of argumentation schemes (Walton et al., 2008; Walton, 1996). According to this theory, different types of arguments invite specific types of critiques. For example, an argument from authority made in the prompt – *According to X, Y is the case* – avails critiques along the lines of whether X has the necessary knowledge and is an unbiased source of information about Y. Analyzing prompts used in an assessment of argument critique skills, Song et al. (2014) identified a number of common schemes, such as arguments from policy, sample, example, and used the argumentation schemes theory to specify what critiques would count as "good" for arguments from the given scheme. Once a prompt is associated with a specific set of argumentation schemes, it follows that those critiques that count as good under one of the schemes used in the prompt would be considered as good critiques in essays responding to that prompt. The goal of the annotation was to identify all sentences in an essay that participate in making a good critique, according to the above definition. Every sentence in an essay is annotated with the label of the critique that it raises, or "generic" if none. In the current paper, we build upon this earlier work.

In practical large-scale automated scoring contexts, new essay prompts are often introduced without rebuilding the scoring system, which is typically subject to a periodic release schedule. Therefore, the assumption that the system

Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 244–249
Vancouver, Canada, July 30 - August 4, 2017. ©2017 Association for Computational Linguistics
https://doi.org/10.18653/v1/P17-2038

will have seen essays responding to each of the prompts it could encounter at deployment time is often unwarranted. Further, not only should a system be able to handle responses to an unseen prompt, it must do it gracefully, since a large disparity in the system's performance across different prompts might raise fairness concerns.

Our practical goal is thus a development of a robust argument critique analysis system for essays. Our theoretical goal is the investigation of the extent that it is at all possible to capture aspects of argument *content* in a fashion that would *generalize across various essay topics*.

2 Annotation

We used Song et al. (2014) annotation protocol, adapting as needed to cover additional argumentation schemes. Song et al. (2017) provides a detailed exposition of the argumentation-scheme-based analysis of a number of prompts and of the annotation process. For the current study, we used a simplified version of the annotation where sentences are labeled as **non-generic** (namely, containing a good critique according to some argumentation scheme), or **generic** (all the rest of the sentences in the essay). The average inter-annotator agreement on the "generic" vs "non-generic" sentence-level classification is k=0.67.

The "non-generic" category covers all sentences that raise a good critique; everything else is "generic". The latter category thus includes, for example, sentences that rehash the argument in the prompt, provide critical but vague statements that cannot be clearly identified as a specific critique from our list (such as "The author should provide more information"), provide specific critical statements that aren't valid arguments. For example, in response to the prompt that states that the new policy led to a 10% decrease in unemployment in four years, one writer argued that "People who were unemployed could have died within the last four years and that is why there is a decrease." While trying to provide an alternative explanation to the putative effect of the policy is a reasonable move, this is not a valid argument, because it is exceedingly unlikely that unemployed people died in such disproportionate numbers to have such a big impact on the unemployment statistics.

3 Data

For this study, we use a same-topic and an across-topics sets of college-level argument critique essays. The first is used to set the bar for the performance in the context where the training and the testing essays respond to the same prompt. The second is the main dataset focused on generalization across prompts.

3.1 Same-topic

A total of 900 essays were annotated, 300 essays for each of 3 prompts. For each prompt, we train a model on 260 responses and test on 40. The training sets per prompt contain on average 2,700 sentences, of which 38% are classified as containing good argument critiques. Two of the three same-topic sets were used previously in Song et al. (2014).

3.2 Across-topics

A total of 500 essays were annotated, 50 essays for each of 10 prompts. We perform 10-fold cross validation, training on 9 prompts and testing on the 10th, modeling a scenario of generalization to an unknown topic. There are, on average, 5,492 sentences available for training, of which 3,917 (42%) are classified as containing good critiques.

4 How far do we get with pure content?

Given that making a good critique is presumably mostly about saying the right things, we expect lexical models to perform well in same-topic context and badly in the across-topics one. We evaluated 1-3grams, 1-4grams, and 1-5grams models learned using a logistic regression classifier. Differences in performance tended to be in the third or fourth decimal digit; we therefore report results for 1-3grams only. Classification accuracies are shown in row 2 of Table 1, following the majority baseline (row 1), in both same-topic and across-topics scenarios. We show average performance (**Av**) as well as the worst performance (**Min**) on 3 prompts (same-topic) and on 10 prompts (across topics). We also evaluated models built using chi-square based feature selection (**f.s.**), eliminating all features with p value above 0.05 (row 3 in Table 1).

For the same-topic context, lexical features perform at .738. As expected, lexical features are much less effective across topics, with average performance of only .645. We observe substantial gaps of 9 (.738 vs .645) and 8 (.679 vs .604) accuracy points, for average and worst case, respectively, between same-topic and across-topics scenarios for ngram models. Feature selection is ineffective in both scenarios (compare rows 2 and 3 in Table 1).

5 How far do we get with pure structure?

An approach that is perhaps better aligned with the across-topics setting is to notice that in detailing one's arguments, one tends to utilize a specially structured discourse, and that discourse role could provide a clue to the argumentative function of a sentence, without reliance on what the sentence is actually saying (beyond discourse connectives that are used to help identify the discourse role). In particular, argumentative essays often have a fairly standard structure, where a general claim (or stance, or thesis) on the issue is introduced in the beginning of the essay, followed by a sequence of main points, each elaborated using supporting statements, and finally followed by a conclusion that often re-states the thesis and provides a high-level summary of the argument. We expect the "meat" of the argument to occur mostly in the supporting statements that provide detailed exposition of the author's arguments. We use a discourse parser for argumentative essays (Burstein et al., 2003) to classify sentences into the following discourse units: Thesis, Background, MainPoint, Support, Conclusion, and Other. Row 4 (**dr**) in Table 1 shows the performance of this set of 6 binary features. Of the 6 features, Support and MainPoint have a positive weight (predict "non-generic"), the rest predict "generic".

We further hypothesize that the position of a sentence inside a discourse segment might also provide some information: A sentence surrounded by Support sentences is likely to be in the middle of exposition of an argument, as opposed to the last Supporting sentence before the next Main Point that could be summary-like, leading up to a shift to a new topic. We therefore built two sets of transition features, one for all pairs of <previous_sentence_role,current_sentence_role> (such as <Thesis,Main Point> for a sentence that is classified as Main Point and follows a Thesis sentence), and the other – for all pairs of <current_sentence_role,next_sentence_role>. We also added BeginningOfEssay and EndOfEssay discourse tags to handle the first and the last sentences of the essay. Table 2 shows the weights for some of the features.

	Model	**Same Topic**		**Across Topics**	
		Av.	Min.	Av.	Min.
1	Majority	.660	.612	.580	.399[1]
2	1-3gr	.738	.679	.645	.604
3	1-3gr f.s.	.697	.635	.633	.580
4	dr	.668	.619	.678	.634
5	dr_pn	.677	.631	.687	.649
6	dr_pn+1-3gr	.741	.690	.700	.674
7	1-3gr ppos	.728	.687	.654	.616
8	dr_pn+1-3gr ppos	**.745**	**.701**	**.706**	**.686**
9	SongEtAl2014	.756	.702	.678	.642

Table 1: Classification accuracies for generic vs non-generic sentences. Our best results for same-topic and across-topics scenarios are boldfaced.

	Discourse Transition Feature			**Weight**
	Previous	**Current**	**Next**	
1	Support	Support		0.760
2	MainPoint	Support		0.238
3	Thesis	Support		-0.028
4		Support	Support	0.716
5		Support	MainPoint	0.220
6		Support	Concl.	0.047
7		Concl.	Concl.	0.063
8		Concl.	EndOfEssay	-0.680

Table 2: Weights of the transition features.

We observe that the likelihood of the current Support sentence to carry argumentative content is higher if it follows another Support sentence (row 1) than if it follows a Main Point (row 2); if the Support sentence follows Thesis, it is actually not likely to contain argumentative content (perhaps it is more like a Main Point sentence than like a typical Support). Likewise, being followed by another Support sentence is a good sign (row 4), but being the last Support sentence before transitioning to a new Main Point has a much lower positive weight (row 5), and being the last Support before Conclusion has a still lower positive weight (row 6). Interestingly, while being the last Conclusion sentence in the essay strongly predicts "generic" (row 8), if the next sentence is still within the Conclusion segment, the prediction is actually slightly positive (row 7), suggesting that some authors rehash their arguments in substantial detail in concluding remarks, warranting a "non-generic" designation.

[1] The majority baseline for one prompt is below 50% because for that prompt the majority class is actually sentences that raise appropriate arguments, differently from the other 9 prompts.

Table 1 shows the performance of the discourse role features (**dr**), the transition pairs using previous and next discourse roles (**dr_pn**),[2] and the combination of content and discourse (row 6).

We observe that transforming the discourse role features into transitional features is effective. Second, the discourse role features are inferior to the content features for same-topic, while the opposite is true for the across-topics scenario.

Discourse structure information does in fact get us quite far in the across-topics scenario, further than the lexical information on its own. Combining the two types of information further improves performance in across-topics scenario, and reduces the gap between across-topics and same-topic contexts to 4 points on average (.741 vs .700) and 1.5 points in worst case (.690 vs .674), for a combined discourse structure and content model.

6 Can we do better?

In an attempt to further improve across-topics performance, we generalized ngrams representations and adapted feature selection to reflect the across-topics dynamic more directly.

6.1 Generalized ngrams

Suppose the prompt is arguing that some entity N should do some action V. While N and V might differ across prompts, critical sentences to the end that N should not do V are likely to occur across different prompts. In the current ngrams representation, N and V differ across prompts, and are unknown for a prompt that is unseen during training. We represent all content words (nouns, verbs, adjectives, adverbs, and cardinal numbers) in the prompt as their part-of-speech labels; we should be able to capture features such as "should not VB". Rows 7 and 8 in Table 1 show the **1-3gr ppos** model; it improves over 1-3gr in both the average and the worst cases, on its own and on top of the dr_pn features, in the across-topics scenario. The improvement over dr_pn+1-3gr is marginally significant ($p<0.1$, Wilcoxon Signed-Ranks 2-tailed test, $n=10$, $W=33$).

The single strongest lexical predictor of a generic sentence is the first person singular pronoun *I*; such sentences are likely to express stance (*I think this is a good plan*), or contain discourse-management expressions such as *I will show that the author's arguments are flawed*. Words such as *assumptions*, *evidence*, *information*, *argument*, *statistics*, *idea*, *reasons* all have negative weight, suggesting that they typically belong to generic sentences such as *The author's argument lacks evidence* that does not raise a specific critique. Lexical features for the positive class include modality as in *might*, *perhaps*, *could*, *possible that*, *potential*, *necessarily*, *if a*; negation (*not*, *will not*), as well as more specific lexica that point out, for example, outcomes of a policy (*expensive*, *increase in*, *affected the*, *fails to*). Positive features with prompt elements include *NNS does not*, *NN do not*, *many NNS*, *NN NNS are*, *NNS who VBD*, *could have VBN*, *will not VB*.

6.2 Feature Selection

We experimented with three feature selection methods. (1) We selected features with p<0.05 using χ^2 test (**p0.05**). (2) We selected features with p<0.05 for at least two out of the 9 training prompts, to find features that are likely to generalize across prompts (**p0.05_2pr**). (3) We selected features based on their mutual information with the label conditioned on values of the dr_pn features, to encourage selection of features that augment, rather than repeat, the discourse information. We calculated for each training prompt, and took the 2nd highest of the 9 values.[3] We selected features in the top 5% of this metric (**mi5%_2pr**).

Table 3 shows the results. The p0.05 mechanism is ineffective; 0.05_2pr selection is better. The mi5%_2pr mechanism performs within .002 of the original, while reducing the number of features by two orders of magnitude.

F.s.	#Features	Av.	Min.
No f.s.	∼ 200,000	.706	.686
p0.05	∼ 3,500	.687	.656
p0.05_2pr	< 500	.702	.678
mi5%_2pr	∼ 1,000	.704	.684

Table 3: Performance of feature selection, for the dr_pn+1-3gr ppos model, across topics.

7 Benchmark

We also compared our best across-topics system (dr_pn+1-3gr ppos) to the system described in

[2] Since argument critiques often span more than one sentence, we experimented with sequence labeling using Conditional Random Fields, but performance was not better than with logistic regression.

[3] Thus, the feature has at least that much informaton beyond dr_pn for at least two different prompts.

Song et al. (2014). The Song et al. (2014) system uses the following features: length of the sentence, parts of speech, overlap of words in the sentence with the prompt, relative position of the sentence in the essay, 1-3gr, and 1-3gr in previous and next sentences. The performance is shown in row 9 in Table 1. Our improvement over Song et al. (2014) is statistically significant in the across-topics scenario ($p<0.05$, Wilcoxon Signed-Ranks test, 2-tailed, n=10, W=51).

8 Conclusion

We presented experiments on classifying essay sentences as containing good argument critiques or not. While a good argument critique is a matter of content, we show that it is possible to build classifiers that are not prompt-specific, using discourse structure features and generalized lexical features that take into account reference to the text of the prompt to which the author is responding. Starting from a ngrams baseline where the performance gap between same-prompt and across-topics scenarios is 9 accuracy points on average (.738 vs .645) and 8 points in worst case (.679 vs .604), we close half the gap in average performance (.745 vs .706) and are down to only 1.5 point difference in worst case performance (.701 vs .686). This performance is preserved with only about 0.5% of the features, using a conditional mutual information criterion. The improvement in worst case performance is important for ensuring that the system does not exhibit large performance differences across different essay prompts used on the same test. We also show that our best system significantly improves over the state-of-art system for argument critique detection task on comparable essay data for the across-topics scenario.

References

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. https://doi.org/10.1109/MIS.2003.1179191.

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)*. Association for Computational Linguistics, Berlin, Germany, pages 549–554. http://anthology.aclweb.org/P16-2089.

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)*. Association for Computational Linguistics, Berlin, Germany, pages 1589–1599. http://www.aclweb.org/anthology/P16-1150.

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)*. Association for Computational Linguistics, Berlin, Germany, pages 1127–1137. http://www.aclweb.org/anthology/P16-1107.

Nathan Ong, Diane Litman, and Alexandra Brusilovsky. 2014. Ontology-based argument mining and automatic essay scoring. In *Proceedings of the First Workshop on Argumentation Mining*. Association for Computational Linguistics, Baltimore, Maryland, pages 24–28. http://www.aclweb.org/anthology/W14-2104.

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*. Association for Computational Linguistics, San Diego, California, pages 1384–1394. http://www.aclweb.org/anthology/N16-1164.

Yi Song, Paul Deane, and Beata Beigman Klebanov. 2017. Toward the automated scoring of written arguments: Developing an innovative approach for annotation. *ETS Research Report Series* https://doi.org/10.1002/ets2.12138.

Yi Song, Michael Heilman, Beata Beigman Klebanov, and Paul Deane. 2014. Applying argumentation schemes for essay scoring. In *Proceedings of the First Workshop on Argumentation Mining*. Association for Computational Linguistics, Baltimore, Maryland, pages 69–78. http://www.aclweb.org/anthology/W14-2110.

Christian Stab and Iryna Gurevych. 2014. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, pages 46–56. http://www.aclweb.org/anthology/D14-1006.

Reid Swanson, Brian Ecker, and Marilyn Walker. 2015. Argument mining: Extracting arguments from online dialogue. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics, Prague, Czech Republic, pages 217–226. http://aclweb.org/anthology/W15-4631.

Douglas N. Walton. 1996. *Argumentation schemes for presumptive reasoning*. Mahwah, NJ: Lawrence Erlbaum.

Douglas N. Walton, Chris Reed, and Fabrizio Macagno. 2008. *Argumentation schemes*. New York, NY: Cambridge University Press.