Using Argumentative Structure to Grade Persuasive Essays

Andreas Stiegelmayr and Margot Mieskes(✉)

University of Applied Sciences, Darmstadt, Germany
{andreas.stiegelmayr,margot.mieskes}@h-da.de

Abstract. In this work we analyse a set of persuasive essays, which were marked and graded with respect to their overall quality. Additionally, we performed a small-scale machine learning experiment incorporating features from the argumentative analysis in order to automatically classify good and bad essays on a four-point scale. Our results indicate that bad essays suffer from more than just incomplete argument structures, which is already visible using simple surface features. We show that good essays distinguish themselves in terms of the amount of argumentative elements (such as major claims, premises, etc.) they use. The results, which have been obtained using a small corpus of essays in German, indicate that information about the argumentative structure of a text is helpful in distinguishing good and bad essays.

1 Introduction

Writing essays is an essential part of every-day-life of pupils and students. In persuasive essays there is an additional challenge in getting argumentative structures right. Research in automated essay scoring has been looking at a wide variety of features such as text structure, vocabulary, spelling, etc. All of which are important, but considering current research in argument mining, there is a lack of research into the relationship between argument structure and essay quality. In this work, we address how various aspects of arguments (i. e., major claims, premises, etc.) relate to the quality of an essay. Additionally, we use features based on arguments in a classification task using machine learning methods. Our results indicate that persuasive essays can be reliably classified using argument-based features. This work contributes in two ways to research in the area of argument mining and essay scoring: First, we show that the argumentative structure can be used to distinguish good and bad essays in an essay scoring task. Second, to our knowledge, this is the first work to bring these two topics together based on German data.

2 Related Work

As this work is at the intersection of *argument mining* and *essay scoring* we look at relevant previous work in both areas. Reviewing the available literature in detail is beyond the scope of this paper.

G. Rehm and T. Declerck (Eds.): GSCL 2017, LNAI 10713, pp. 301–308, 2018.
https://doi.org/10.1007/978-3-319-73706-5_26

2.1 Argument Mining

Although the topic of argument mining is fairly new, it goes back to acient greece. Habernal and Gurevych (2017) provide a current, extensive overview on the area. We specifically looked at the guidelines presented by Stab and Gurevych (2014): The authors analysed three components for argument structures: Major Claim, Claim and Premise. The basis of an argument is the claim, which relates to one or more premises. This relation has two attributes: support and attack. The Major Claim is the basis for the whole essay and can be found either in the introduction or in the conclusion. In the introduction it serves as a statement, which is related to the topic of the essay. In its conclusion it summarizes the arguments of the author.

Wachsmuth et al. (2016) also based their work on Stab and Gurevych (2014), but they consider Argumentative Discourse Units (ADU). ADUs can be complete sentences or partial sentences, especially in cases where two sentences are connected via "and". The authors defined a set of features, such as n-grams, part-of-speech n-grams, etc., and analysed the flow of ADUs based on graphs.

Work on German data is (compared to English data) rare. One example is by Peldszus and Stede (2013), where artificially constructed short texts were used to determine inter-annotator aggreement on argument annotation. Kluge (2014) used web documents from the educational domain, and Houy et al. (2013) used legal cases. All authors analysed the argumentative structure of their documents.

Work on essays has been carried out for example by Faulkner (2014), but with the aim of identifying the stance of an author towards a specific claim and in the domain of summarization. Stab and Gurevych (2014) also used essays in their study, but focused on the identification of arguments.

2.2 Essay Scoring

Dong and Zhang (2016) present an overview on essay scoring, including commercial tools available. They analysed a range of features for essay scoring and used them in a deep learning approach. The authors used surface features such as the length of characters, words, etc., and linguistic features such as Part-of-Speech (POS) tags and POS-n-grams. They used words and their syonyms based on the prompts for each essay and their appearance in the resulting texts. Additionally, they used uni- and bigrams and corrected for spelling errors. They considered the task as a binary classification task, with good essays defined as "essays with a score greater than or equal to the average score and the remainder are considered as bad scoring essays". The authors report a κ-based evaluation, which achieves results "close to that of human raters".

Using arguments for essay scoring has been done by Ghosh et al. (2016), based on TOEFEL-data. Their results, based on number and length of argument chaines, indicate that essays containing many claims, connected to few or no premises score lower. They also found that length is highly correlated with the scores.

3 Data Set

We collected a corpus containing 38 essays, which are available on the internet[1]. We also tried to get real essays by contacting various schools and teachers. These would also have teachers markings. Unfortunately, this is not a viable path to follow, due to various reasons: Firstly, these essays are subject to a very strict data protection law, which puts a range of obstacles on obtaining such data. Secondly, very few schools use electronic methods and tools for writing essays. So all schools we got in touch with and which would have been willing to grant us access to their essays and markings, provided we agree to the data protection regulations, only had essays which were hand-written on paper. Digitizing them, including proof-reading, would have been beyond the scope of this work. Therefore, we took data that was available on the internet in a machine-readable format. The corpus was manually annotated using the guidelines by Stab and Gurevych (2014) using WebAnno[2]. Figure 1 shows an example structure of the resulting argument tree. The whole data set contains approximately 120,000 words, and slightly over 4,000 sentences. In total, we analysed over 1,000 argument units containing over 1,000 premises and almost 300 claims. 50% of the argument units had more than 15 words. Details can be found in Table 1.

In addition to the argument annotation, we also annotated the quality of the essays, using the German school marking system, which is based on numbers 1 to 4, where 1 represents a very good result and 4 represents a very poor result. We decided to use a reduced version of the German marking systems due to the following reasons: At universities only marks from 1 to 4 are given, with marks >4 being a *fail*[3]. Due to the data set size, using a more fine-grained marking scale would have given us very few data points for each class to train a machine learning system on, especially with respect to the already small data set size.

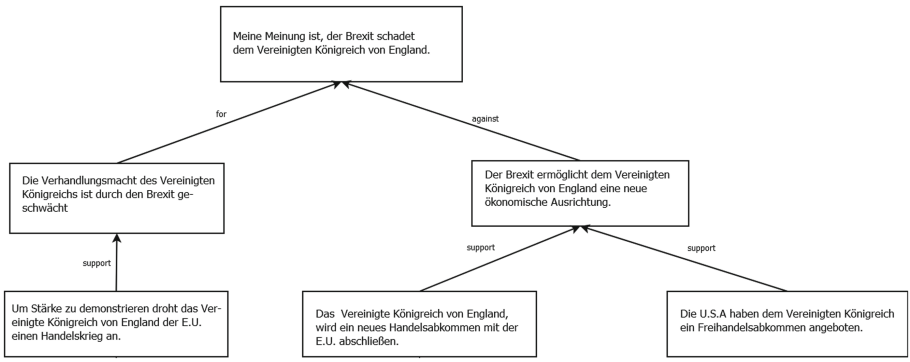

Fig. 1. Example for an argument tree as found in our data.

[1] The list of sources can be found at https://b2drop.eudat.eu/s/tR5spZeyRcW20VB.
[2] https://webanno.github.io/webanno/.
[3] One annotator marked one essays with a fail.

Table 1. Statistical information on the corpus.

Element	Count
Words	119,043
Sentences	4,047
Paragraphs	133
Argument units	1,324
Major claims	34
Claims	286
Premises	1,004
Spelling anormaly	130
Median words per argument unit	15.3

We assume, that the quality of the essay corpus is not representative of regular school essays, but rather represent the quality available on the interent. We observe, that the quality of the essays is mediocre, with many authors not explicitely stating their point of view. In some extreme cases the major claim was not detectable. This results in difficulties in deciding whether a sentence contains an argument unit or not. The distribution of the marks is therefore very skewed, with approximately 23.1% of the essays achieving good (mark 2) or very good (mark 1) marks and 77% of the essays achieving poor (mark 3) or very poor (mark 4) results. An additional problem – especially for the later automatic analysis – is the usage of metaphors, which we did not look into in this work.

About one third of the essays (13 out of 38) were graded by two persons. The percentage agreement between the two grades was 0.53. Considering a measure that is specifically designed to evaluate annotations by two coders and correcting for chance agreement (which percentage agreement does not do), we achieve a value of $S = 0.42$, which according to Landis and Koch. (1977) shows a moderate agreement. All values were calculated using DKPro Statistics[4] (Meyer et al. 2014).

4 Experimental Setup

We use DKPro Components[5], such as DKPro Core, DKPro TC and Uby for our experiments.

We defined a range of features, based on the argumentation annotation and previous work. We distinguish between *baseline* features, which have already been used in previous work and *argument* features, which are based on the argumentation annotation. The baseline features contain easy to determine features,

[4] https://dkpro.github.io/dkpro-statistics/.
[5] https://github.com/dkpro/.

such as number of tokens, number of sentences, etc. Additionally, we took into account POS-based features, which include nouns, verbs, adjectives, etc.

Based on earlier work, we included information about whether the author used overly long words or short words. We also checked for spelling errors using the LanguageTool[6]. Wachsmuth et al. (2016) observed that questions are not arguments. Therefore, we also extracted the number of questions with and without arguments. According to Stab and Gurevych (2014) one paragraph should only contain one claim. Therefore, we also counted the number of claims and the number of paragraphs in our documents. Additionally, we looked at the number of sentences with and without arguments. Finally, we examined the n-grams found in the annotated arguments. Based on Ghosh et al. (2016) we looked at the graph created by the argument structure over a document. An example can be found in Fig. 1. Tree size and grade show a strong, negative correlation (Pearsons $r = -0.57$[7]), meaning, the larger the tree, the higher the grade. Additionally, we use the argument graph to determine whether it starts with a major claim or not and which arguments are not linked to the major claim. Finally, we determined whether a person consistently uses the correct tense. The full set of features can be found in the respective .arff-files[8].

5 Results and Discussion

We experimented with various machine learning algorithms, using WEKA[9]. As we wanted to gain a qualitative insight into the results obtained through the machine learning methods, we specifically looked into decision trees (J48).

Table 2. Classification result for individual marks using the whole feature set.

Mark	p	r	f1
1	1	0.6	0.75
2	0.833	0.833	0.833
3	0.667	0.6	0.632
4	0.773	0.895	0.829
Avrg.	0.87	0.86	0.858

We observed, that the main features contributing to the results in Table 2 were `NrOfMajorClaims`, `NrOfPremises` and `RatioSentenceNonQuestion`. This supports earlier work, that the number of major claims and premises allows for detecting good essays.

[6] https://languagetool.org/de/.
[7] This correlation is significant on $\alpha = 0.01$.
[8] https://b2drop.eudat.eu/s/mTyUTJrCNyO4I3e.
[9] http://www.cs.waikato.ac.nz/ml/weka/.

Table 3. Classification result for individual marks using baseline features only.

Mark	p	r	f1
1	0	0	0
2	0	0	0
3	0.545	0.6	0.571
4	0.721	0.838	0.775
Avrg.	0.476	0.545	0.508

Using only the baseline features we observed that the lower marks (3 and 4) were still classified fairly reliably, but the better marks (1 and 2) performed very poorly. Looking into the results in detail revealed that essays marked as "2" were mostly confused with essays marked as "1", which indicates, that not so good essays suffer from more than just a lack of good argumentative structure, which is already visible with the surface features. This becomes very prominent looking at the resulting tree, where the most important features for 3 and 4 were a combination of fewer characters and a high amount of spelling errors. In order to reduce the importance of the spelling errors, we artificially introduced spelling errors to the good essays (marked 1 and 2). We tried to achieve a similar ratio as for the bad essays (marked 3 and 4). Thereby, we managed to reduce the importance of the spelling feature in the feature ranking. But the overall results (including the observations concerning the usage of major claims and premises in connection to the resulting grade) were similar to those presented in Table 2 ($p = 0.86$; $r = 0.85$ and $f1 = 0.85$) and the discussion above (Table 3).

Table 4. Classification result for individual marks using custom features only.

Mark	p	r	f1
1	1	0.6	0.75
2	0.833	0.833	0.833
3	0.700	0.700	0.700
4	0.810	0.895	0.850
Avrg.	0.87	0.86	0.858

The argumentative features allow us to clearly identify and distinguish between the various essays. A closer look at the resulting tree indicates, that good essays use premises cautiously and also keep the major claims low, which is in line with observations from previous work. Bad essays have a higher number of major claims, but also a high number of disconnected arguments (Table 4).

Overall, our results indicate that poor essays suffer from more than just poor argumentation and authors should address issues such as spelling, usage of tense, number of conjunctions and length of words. Once these issues are considerably

improved, the argumentative elements of the essays should be considered, such as a high number of major claims. For authors who already achieve good results, the focus can be put on argumentative elements, such as the number of premises, which is higher than in very good essays.

6 Conclusion and Future Work

We presented work on using argumentative structures and elements in identifying the quality of persuasive essays. We found that argumentative elements support the identification of good essays. Bad essays can be classified reliably using traditional features, indicating that these authors need to address issues such as spelling errors before improving on argumentative elements in their writing.

The next step would be to increase the data set size in order to solidify our findings. More data would also allow us to use more sophisticated machine learning methods. Additionally, we would like to incorporate a range of features previously used in the area of essays scoring, such as latent semantic analysis. Finally, we would like to have a closer look at the issue of metaphors in argumentative essays and their contribution to arguments and essay quality.

Acknowledgements. We would like to thank the German Institute for Educational Research and Educational Information (DIPF), where this work was carried out. Additionally, we would like to thank the reviewers for their helpful comments.

References

Dong, F., Zhang, Y.: Automatic features for essay scoring - an empirical study. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp. 1072–1077 (2016)

Faulkner, A.R.: Automated Classification of Argument Stance in Student Essays: A Linguistically Motivated Approach with an Application for Supporting Argument Summarization. Ph.D. thesis, Graduate Center, City University of New York (2014)

Ghosh, D., Khanam, A., Han, Y., Muresan, S.: Coarse-grained argumentation features for scoring persuasive essays. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Short Papers, vol. 2, pp. 549–554. Association for Computational Linguistics (2016). https://doi.org/10.18653/v1/P16-2089. http://aclweb.org/anthology/P16-2089

Habernal, I., Gurevych, I.: Argumentation mining in user-generated web discourse. Comput. Linguist. **43**(1), 125–179 (2017)

Houy, C., Niesen, T., Fettke, P., Loos, P.: Towards automated identification and analysis of argumentation structures in the decision corpus of the German Federal Constitutional Court. In: The 7th IEEE International Conference on Digital Ecosystems and Technologies (IEEE DEST-2013) (2013)

Kluge, R.: Searching for Arguments - Automatic analysis of arguments about controversial educational topics in web documents. AV Akademikerverlaug, Saarbrücken (2014)

Landis, R.J., Koch, G.G.: The measurement of observer agreement for categorical data. Biometrics **33**(1), 159–174 (1977)

Meyer, C.M., Mieskes, M., Stab, C., Gurevych, I.: DKPro Agreement: an open-source Java library for measuring inter-rater agreement. In: Proceedings of the 25th International Conference on Computational Linguistics (COLING), Dublin, Ireland, pp. 105–109 (2014)

Peldszus, A., Stede, M.: Ranking the annotators: an agreement study on argumentation structure. In: Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse, pp. 196–204 (2013)

Stab, C., Gurevych, I.: Identifying argumentative discourse structures in persuasive essays. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 46–56. Association for Computational Linguistics (2014). https://doi.org/10.3115/v1/D14-1006. http://aclweb.org/anthology/D14-1006

Wachsmuth, H., Al-Khatib, K., Stein, B.: Using argument mining to assess the argumentation quality of essays. In: Proceedings of the 26th International Conference on Computational Linguistics (COLING 2016), pp. 1680–1692, December 2016