Coarse-grained Argumentation Features for Scoring Persuasive Essays

Debanjan Ghosh[§], **Aquila Khanam**[†], **Yubo Han**[†] and **Smaranda Muresan**[‡]
[§]School of Communication and Information, Rutgers University, NJ, USA
[†]Department of Computer Science, Columbia University, NY, USA
[‡]Center for Computational Learning Systems, Columbia University, NY, USA
debanjan.ghosh@rutgers.edu, {ak3654,yh2635,smara@columbia.edu}

Abstract

Scoring the quality of persuasive essays is an important goal of discourse analysis, addressed most recently with high-level persuasion-related features such as thesis clarity, or opinions and their targets. We investigate whether argumentation features derived from a coarse-grained argumentative structure of essays can help predict essays scores. We introduce a set of argumentation features related to argument components (e.g., the number of claims and premises), argument relations (e.g., the number of supported claims) and typology of argumentative structure (chains, trees). We show that these features are good predictors of human scores for TOEFL essays, both when the coarse-grained argumentative structure is manually annotated and automatically predicted.

1 Introduction

Persuasive essays are frequently used to assess students' understanding of subject matter and to evaluate their argumentation skills and language proficiency. For instance, the prompt for a TOEFL (Test of English as a Foreign Language) persuasive writing task is:

> *Do you agree or disagree with the following statement? It is better to have broad knowledge of many academic subjects than to specialize in one specific subject. Use specific reasons and examples to support your answer.*

Automatic essay scoring systems generally use features based on grammar usage, spelling, style, and content (e.g., topics, discourse) (Attali and Burstein, 2006; Burstein, 2003). However, recent work has begun to explore the impact of high-level persuasion-related features, such as opinions and their targets, thesis clarity and argumentation schemes (Farra et al., 2015; Song et al., 2014; Ong et al., 2014; Persing and Ng, 2015). In this paper, we investigate whether argumentation features derived from a coarse-grained, general argumentative structure of essays are good predictors of holistic essay scores. We use the argumentative structure proposed by Stab and Gurevych (2014a): *argument components* (major claims, claims, premises) and *argument relations* (support, attack). Figure 1(i) shows an extract from an essay written in response to the above prompt, labeled with a claim and two premises. The advantage of having a simple annotation scheme is two-fold: it allows for more reliable human annotations and it enables better performance for argumentation mining systems designed to automatically identify the argumentative structure (Stab and Gurevych, 2014b).

The paper has two main contributions. First, we introduce a set of argumentation features related to three main dimensions of argumentative structure: 1) features related to *argument components* such as the number of claims in an essay, number of premises, fraction of sentences containing argument components; 2) features related to *argument relations* such as the number and percentage of supported and unsupported claims; and 3) features related to the *typology of argumentative structure* such as number of chains (see Figure 1(ii) for and example of chain) and trees (Section 3). On a dataset of 107 TOEFL essays manually annotated with the argumentative structure proposed by Stab and Gurevych (2014a) (Section 2), we show that using all the argumentation features predicts essay scores that are highly correlated with human scores (Section 3). We discuss what features are correlated with high scoring es-

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 549–554,
Berlin, Germany, August 7-12, 2016. ©2016 Association for Computational Linguistics

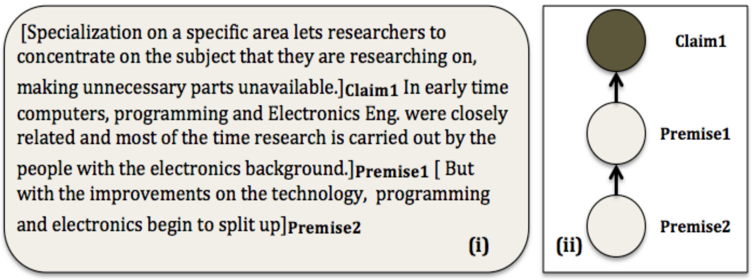
[Specialization on a specific area lets researchers to concentrate on the subject that they are researching on, making unnecessary parts unavailable.]**Claim1** In early time computers, programming and Electronics Eng. were closely related and most of the time research is carried out by the people with the electronics background.]**Premise1** [But with the improvements on the technology, programming and electronics begin to split up]**Premise2**

Figure 1: (i) Essay extract showing a claim and two premises and (ii) the corresponding argumentative structure (i.e., chain).

says vs. low scoring essays. Second, we show that the argumentation features extracted based on argumentative structures automatically predicted by a state-of-the-art argumentation mining system (Stab and Gurevych, 2014b) are also good predictors of essays scores (Section 4).[1]

2 Data and Annotation

We use a set of 107 essays from TOEFL11 corpus that was proposed for the first shared task of Native Language Identification (Blanchard et al., 2013). The essays are sampled from 2 prompts: P1 (shown in the Introduction) and P3:

> *Do you agree or disagree with the following statement? Young people nowadays do not give enough time to helping their communities. Use specific reasons and examples to support your answer.*

Each essay is associated with a score: high, medium, or low. From prompt P1, we selected 25 high, 21 medium, and 16 low essays, while for prompt P3 we selected 15 essays for each of the three scores.

For annotation, we used the coarse-grained argumentative structure proposed by Stab and Gurevych (2014a): argument components (major claim, claim, premises) and argument relations (support/attack). The unit of annotation is a clause. Our annotated dataset, $TOEFL_{arg}$, includes 107 major claims, 468 claims, 603 premises, and 641 number of sentences that do not contain any argument component. To measure the inter-annotator agreement we calculated P/R/F1 measures, which are used to account for fuzzy boundaries (Wiebe et al., 2005). The F1 measure for overlap matches (between two annotators) for argument components is 73.98% and for argument relation is 67.56%.

3 Argumentation Features for Predicting Essays Scores

A major contribution of this paper is a thorough analysis of the key features derived from a coarse-grained argumentative structure that are correlated with essay scores. Based on our annotations, we propose three groups of features (Table 1). The first group consists of features related to *argument components* (AC) such as the number of claims, number of premises, fraction of sentences containing argument components. One hypothesis is that an essay with a higher percentage of argumentative sentences will have a higher score. The second group consists of features related to *argument relations* (AR), such as the number and percentage of supported claims (i.e., claims that are supported by at least one premise) and the number and percentage of dangling claims (i.e., claims with no supporting premises). In low scoring essays, test takers often fail to justify their claims with proper premises and this phenomenon is captured by the dangling claims feature. In contrary, in high scoring essays, it is common to find many claims that are justified by premises. We also consider the number of attack relations and attacks against the major claim. Finally, the third group consists of features related to the *typology of argument structures* (TS) such as the number of argument chains ($Chain$), number of argument trees of height $=$ 1 ($Tree_{h=1}$) and the number of argument trees of height $>$ 1 ($Tree_{h>1}$). We define an argument chain when a *claim* is supported by a chain of premises. We define $Tree_{h=1}$ as a tree structure of height 1 with more than one leaves, where *the root is a claim* and the leaves are premises

[1] The annotated dataset, $TOEFL_{arg}$, is available at https://github.com/debanjanghosh/argessay_ACL2016/

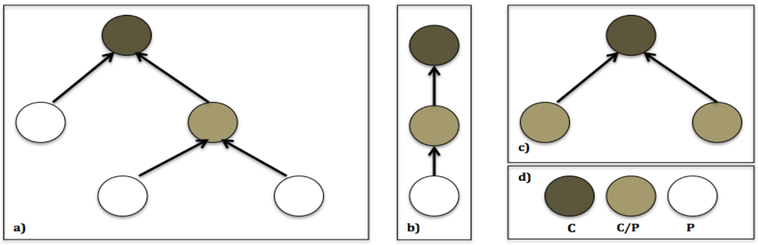

Figure 2: Typology of Argumentative Structure: Examples of (i) $Tree_{h>1}$; (ii) Chain; (iii) $Tree_{h=1}$

Feature Group	Id	Argumentation Feature Description
AC	1 2 3,4	# of Claims # of Premises # and fraction of sentences containing argument components
AR	5, 6 7, 8 9 10, 11	# and % of supported Claims # and % of dangling Claims # of Claims supporting Major Claim # of total Attacks and Attacks against Major Claim
TS	12 13 14	# of Argument $Chains$ # of Argument $Tree_{h=1}$ # of Argument $Tree_{h>1}$

Table 1: Argumentation Features

Features	Correlations
bl	0.535
AC	0.758
AR	0.671
TS	0.691
bl + AC	0.770
bl + AR	0.743
bl + TS	0.735
AC + AR + TS	**0.784**
bl + AC + AR + TS	**0.803**

Table 2: Correlation of LR (10 fold CV) with human scores.

or claims. Finally, $Tree_{h>1}$ is a tree structure of height > 1, where *the root is a claim* and the internal nodes and leaves are either supporting claims or supporting premises. Figure 2 shows examples of a $Tree_{h>1}$ structure, a $Chain$ structure, and a $Tree_{h=1}$ structure. The dark nodes represent claims (C), lighter nodes can be either claims or premises (C/P) and white nodes are premises (P). Figure 1 shows an extract from an essays and the corresponding $Chain$ structure.

To measure the effectiveness of the above features in predicting the holistic essay scores (high/medium/low) we use Logistic Regression (LR) learners and evaluate the learners using quadratic-weighted kappa (QWK) against the human scores, a methodology generally used for essay scoring (Farra et al., 2015). QWK corrects for chance agreement between the system prediction and the human prediction, and it takes into account the extent of the disagreement between labels. Table 2 reports the performance for the three feature groups as well as their combination. Our baseline feature (bl) is the number of sentences in the essay, since essay length has been shown to be generally highly correlated with essay scores (Chodorow and Burstein, 2004). We found that all three feature groups individually are strongly correlated with the human scores, much better than the baseline feature, and the AC features have the highest correlation. We also see that although the number of claims and premises can affect the score of an essay, the argumentative structures (i.e., how the claims and premises are connected in an essay) are also important. Combining all features gives the highest QWK score (0.803).

We also looked at what features are associated with high scoring essays vs. low scoring essays. Based on the regression coefficients, we observe that the high "number and % of dangling claims" are strong features for low scoring essays, whereas the "fraction of sentences containing argument components" (AC feature), "number of supported claims" (AR feature), and "number of $Tree_{h=1}$ structures" and "number of $Tree_{h>1}$ structures" (TS features) have the highest correlation with high scoring essays. For example, in a good persuasive essay, test takers are inclined to use multiple premises (e.g., reasons or examples) to support a claim, which is captured by the TS and AR features. In addition, we notice that attack relations are sparse, as was the case in Stab and Gurevych (2014b) dataset and thus the coefficients for attack relations features (#10, #11 in Table 1) are negligible.

In summary, our findings contribute to research on essay scoring, showing that argumentation features are good predictors of essay scores, besides spelling, grammar, and stylistic properties of text.

4 Automatic Extraction of Argumentation Features for Predicting Essay Scores

To automatically generate the argumentation features (Table 1), we first need to identify the argumentative structures: argument components (major claim, claim, and premise) and relations (support/attack). We use the approach proposed by Stab and Gurevych (2014b).[2] For argument component identification, we categorize clauses to one of the four classes (major claim (MC), claim (C), premise (P), and $None$). For argument relation identification, given a pair of argument clauses Arg_1 and Arg_2 the classifier decides whether the pair holds a support (S) or non-support (NS) relation (binary classification). For each essay, we extract all possible combinations of Arg_1 and Arg_2 from each paragraph as training data (654 S and 2503 NS instances; attack relations are few and included in NS). We do not consider relations that may span over multiple paragraphs to reduce number of non-support instances. For both tasks we use $Lexical$ features (e.g., unigrams, bigrams, trigrams, modal verbs, adverbs, word-pairs for relation identification), $Structural$ features (e.g., number of tokens/punctuations in argument, as well as in the sentence containing the argument, argument position in essay, paragraph position (paragraph that contains the argument)), $Syntactic$ features (e.g., production rules from parse trees, number of clauses in the argument), and $Indicators$ (discourse markers selected from the three top-level Penn Discourse Tree Bank (PDTB) relation senses: Comparison, Contingency, and Expansion (Prasad et al., 2008)).

We use two settings for the classification experiments using libSVM (Chang and Lin, 2011) for both argument component and relation identification. In the first setting, we used the dataset of 90 high quality persuasive essays from (Stab and Gurevych, 2014b) (S&G) as training and use $TOEFL_{arg}$ for testing (out-of-domain setting). In the second setting (in-domain), we randomly split the $TOEFL_{arg}$ into 80% training and 20% for testing (sampled equally from each category (MC, C, P, and $None$ for argument components; S and NS for relations)). Table 3 and 4 present the classification results for identifying argument components in the first and second setting, respectively. We ran experiments for all different features groups and observe that with the exception of the P class, the F1 scores for all the other classes is comparable to the results reported by Stab and Gurevych (2014b). One explanation of having lower performance on the P (premise) category is that the S&G dataset used for training has higher quality essays, while 2/3 of our $TOEFL_{arg}$ dataset consists of medium and low scoring essays (the writing style for providing reasons or example can differ between high and low scoring essays). When we select the top 100 features ("top100") using Information Gain (Hall et al., 2009) the F1 scores for the P class improves. The results in Table 4 show that when training and testing on same type of essays the results are better for all categories except for MC when using the "top100" setup.

Feature Type	MC	C	P	$None$
All features	50.0	44.3	48.6	97.7
top100	60.8	36.2	54.1	97.7

Table 3: F1 for argument components (out-of-domain setting)

Feature Type	MC	C	P	$None$
All features	78.6	53.2	64.0	96.1
top100	53.8	64.5	69.2	96.2

Table 4: F1 for argument components (in-domain setting)

Table 5 shows the results for relation identification in the first setting (out-of-domain). The F1 score of identifying support relations is 84.3% (or 89% using top100), much higher than reported by Stab and Gurevych (2014b). We obtain similar results when training and testing on $TOEFL_{arg}$. We observe that two specific feature groups, $Structural$ and $Lexical$, individually achieve high F1 scores and when combined with other features, they assist the classifier in reaching F1 scores in high 80s%. There can be two explanations for this: 1) essays in $TOEFL_{arg}$ have multiple short paragraphs where the position features such as position of the arguments in the essay and paragraph ($Structural$ group) are strong indicators for argument relations; and 2) due to short paragraphs, the percentage of NS instances are less than in the S&G dataset, hence the $Lexical$ features (i.e., word-pairs between Arg_1 and Arg_2) perform very well.

[2] In future work, we plan to use the authors' improved approach and larger dataset released after the acceptance of this paper (Stab and Gurevych, 2016).

Feature Type	S	NS
All features	84.3	95.0
top100	89.0	97.1

Table 5: F1 for argument relations (out-of-domain setting)

Features	Correlations
AC	0.669
AR	0.460
TS	0.311
AC + AR + TS	0.728
All features	**0.737**

Table 6: Correlation of LR (10 fold CV) with predicted results.

Based on the automatic identification of the argument components and relations, we generate the argumentation features to see whether they still predict essays scores that are highly correlated with human scores. Since our goal is to compare with the manual annotation setup, we use the first setting, where we train on the S&G dataset and test on our $TOEFL_{arg}$ dataset. We select the best system setup (top100 for both tasks; Table 3 and 5). We ran Logistic Regression learners and evaluated their performance using QWK scores. Table 6 shows that the argumentative features related to argument relations (AR) and the typology of argument structures (TS) extracted based on the automatically predicated argumentative structure perform worse compared to the scores based on manual annotations (Table 2). Our error analysis shows that this is due to the wrong prediction of argument components, specifically wrongly labeling claims as premises (Table 3). AR and TS features rely on correctly identifying the *claims*, and thus a wrong prediction affects the features in these two groups, even if the accuracy of supports relations is high. This also explains why the argument components (AC) features still have a high correlation with human scores (0.669). When we extracted the argumentation features using *gold-standard* argument components and *predicted* argument relations, the correlation of AR and TS features improved to 0.576 and 0.504, respectively and the correlation of all features reached 0.769.

5 Related Work

Researchers have begun to study the impact of features specific to persuasive construct on student essay scores (Farra et al., 2015; Song et al., 2014; Ong et al., 2014; Persing and Ng, 2013; Persing and Ng, 2015). Farra et al. (2015) investigate the impact of opinion and target features on TOEFL essays scores. Our work looks a step further by exploring argumentation features. Song et al. (2014) show that adding features related to argumentation schemes (from manual annotation) as part of an automatic scoring system increases the correlation with human scores. We show that argumentation features are good predictors of human scores for TOEFL essays, both when the coarse-grained argumentative structure is manually annotated and automatically predicted. Persing and Ng (2015) proposed a feature-rich approach for modeling argument strength in student essays, where the features are related to argument components. Our work explores features related to argument components, relations and typology of argument structures, showing that argument relation features show best correlation with human scores (based on manual annotation).

6 Conclusion

We show that argumentation features derived from a coarse-grained, argumentative structure of essays are helpful in predicting essays scores that have a high correlation with human scores. Our manual annotation study shows that features related to argument relations are particularly useful. Our experiments using current methods for the automatic identification of argumentative structure confirms that distinguishing between claim and premises is a particularly hard task. This led to lower performance in predicting the essays scores using automatically generate argumentation features, especially for features related to argument relations and typology of structure. As future work we plan to improve the automatic methods for identifying argument components similar to Stab and Gurevych (2016), and to use the dataset introduced by Persing and Ng (2015) to investigate how our argumentation features impact the argument strength score rather than the holistic essay score.

Acknowledgements

This paper is based on work supported by the DARPA-DEFT program. The views expressed are those of the authors and do not reflect the official policy or position of the Department of Defense or the U.S. Government. The authors thank the anonymous reviewers for helpful comments.

References

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

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. 2003. The e-rater® scoring engine: Automated essay scoring with natural language processing.

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27.

Martin Chodorow and Jill Burstein. 2004. Beyond essay length: Evaluating e-rater®'s performance on toefl® essays. *ETS Research Report Series*, 2004(1):i–38.

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

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. 2009. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18.

Nathan Ong, Diane Litman, and Alexandra Brusilovsky. 2014. Ontology-based argument mining and automatic essay scoring. *ACL 2014*, page 24.

Isaac Persing and Vincent Ng. 2013. Modeling thesis clarity in student essays. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 260–269.

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.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind K Joshi, and Bonnie L Webber. 2008. The penn discourse treebank 2.0. In *LREC*. Citeseer.

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*, pages 69–78.

Christian Stab and Iryna Gurevych. 2014a. Annotating argument components and relations in persuasive essays. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING 2014)*, pages 1501–1510.

Christian Stab and Iryna Gurevych. 2014b. Identifying argumentative discourse structures in persuasive essays. In *Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)(Oct. 2014), Association for Computational Linguistics, p.(to appear)*.

Christian Stab and Iryna Gurevych. 2016. Parsing argumentation structure in persuasive essays. *arXiv preprint, arxiv.org/abs/1604.07370.*

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39(2-3):165–210.