Identifying Thesis and Conclusion Statements in Student Essays to Scaffold Peer Review

Mohammad Hassan Falakmasir, Kevin D. Ashley,
Christian D. Schunn, and Diane J. Litman

Learning Research and Development Center,
Intelligent Systems Program, University of Pittsburgh
{mhf11,ashley,schunn,dlitman}@pitt.edu

Abstract. Peer-reviewing is a recommended instructional technique to encourage good writing. Peer reviewers, however, may fail to identify key elements of an essay, such as thesis and conclusion statements, especially in high school writing. Our system identifies thesis and conclusion statements, or their absence, in students' essays in order to scaffold reviewer reflection. We showed that computational linguistics and interactive machine learning have the potential to facilitate peer-review processes.

Keywords: Peer-review, high school writing instruction, discourse analysis, natural language processing, interactive machine learning.

1 Introduction

Writing is essential to communication, learning, and problem solving. However, poor achievement in high school writing is a major deficiency in the US educational system [1]. There appears to be no single best approach to teaching writing; however, some practices have been shown to be more effective than others.

One of these practices, peer-review of writing assignments, is a commonly recommended technique to improve writing skills, especially in large class settings. Peer-review not only provides students with feedback, it also gives them the opportunity to read essays of other students and improve their reflective and metacognitive skills. Several studies have found that providing feedback leads to improvement in the reviewer's writing [2], especially when the students provide constructive feedback [3] and put effort into the process [4].

While web-based peer-review systems solve logistical challenges of the review process, such as distribution of documents, providing rubrics and review criteria, and supporting successive drafts, they are still far from optimal [5]. In particular, reviewers may not focus on the core aspects of the text being evaluated [6]. In argumentative writing, a thesis statement plays a pivotal role: it communicates the author's position and opinion about the essay prompt; it anchors the framework of the essay, serving as a hook for tying the reasons and evidence presented and anticipates critiques and counterarguments [7]. The thesis statement thus has a major influence in assessing writing skills [8]. A conclusion reiterates the main idea and summarizes the

S. Trausan-Matu et al. (Eds.): ITS 2014, LNCS 8474, pp. 254–259, 2014.

entire argument in an essay. It may contain new information, such as self-reflections on the writer's position [7]. Since thesis and conclusion statements both play a critical role in the overall argument and share similar linguistic elements, in this paper we focus on automatically identifying these two core aspects.

Advances in computational linguistics enable systems to automatically and quickly analyze large text corpora. Shermis et al. [9] reviewed the features of the three most successful Automated Essay Evaluation (AEE) systems. These systems can analyze certain pedagogically significant aspects of essays as reliably as expert human graders. In particular, Burstein and Marcu [10] presented a machine learning model for detecting thesis and conclusion sentences in students' essays. Later they extended their model into a discourse analysis system as a part of ETS Criterion® software for online essay evaluation [11]. Their model uses lexical, syntactic, and rhetorical features and a complex classification framework to label different discourse elements of the essays like introductory material, thesis statement, topic sentences, and conclusion. Writing Pal (W-Pal) [12], an Intelligent Tutoring System, uses another AEE methodology to offer writing strategy instruction, game-based essay writing practice, and formative feedback to high school writers. It uses the Coh-Metrix AEE [13] to analyze student essays and provide formative feedback.

We hypothesize that AEE techniques can also improve computer-supported peer-review by calling reviewers' attention to particular features of an essay (e.g. thesis or conclusion statements) that deserve comment. Our AEE model is designed to be used as a part of the SWoRD peer-review system [14]. To the best of our knowledge, no one has used AEE techniques to support intelligent scaffolding of peer-reviews. We believe that our system has the potential to combine the strengths of both web-based peer review and automated essay evaluation systems. With an ability to identify thesis statements, the system will scaffold reviewers' consideration of these issues posing such questions as:

- *SWoRD thinks [quoted text] is [pseudonym]'s thesis statement. Do you agree?*
- *SWoRD cannot find a thesis statement for [pseudonym]'s paper. Can you?*
- *Tell [pseudonym] to add a thesis statement. What thesis statement would you recommend?*

Since the papers we assess are mainly the first drafts of high school essays that often lacking in both style and structure, the peer-review context gives us a unique opportunity to evaluate and improve our model in practice. We are planning to use the model in an interactive machine learning [15] framework. Since we use the results of our model to scaffold peer-review, the model's outputs will be evaluated first by the author of the paper, and then by a number of peer-reviewers. We can use these author and peer evaluations as feedback to improve the model, thus reducing the need for post hoc time-consuming manual text annotation. This exclusive advantage will enable the system to assess its performance in action and improve toward the desired behavior.

2 Methodology

It is important that reviewers attend to thesis statements: how well they are articulated and supported, and whether alternative interpretations/viewpoints are considered [16, 17]. We find that many peer reviewers, however, do not attend to thesis statements and focus instead on minor claims or lower level writing issues, even with review

prompts that specifically asked reviewers to comment on the logic of the argument. When students did use the term *thesis* in their reviews, the comments were not always sufficiently specific.

In this study we address two questions: 1) Can computational linguistic methods detect presence/absence of thesis and conclusion sentences in student essays in order to guide peer reviewers (i.e., at the essay level)? 2) How well does the model distinguish candidate thesis or conclusion statements from other sentences (i.e., the sentence level)? We evaluate our model both at the essay level and sentence level and compare the performance with a positional baseline and manually annotated essays.

We used 432 essays from 8 writing assignments in 2 high school courses on cultural literacy and world literature. We divided the essays into two sets, one for training and development purposes including 6 assignment prompts with 326 essays and the other for test purposes including 2 assignment prompts and 106 essays. We used the training set to build our model and extract the most predictive linguistic features of thesis and conclusion statements in student essays. Then we tested the performance of our model on the unseen test set.

Six human judges annotated our essays, with an instruction manual based on the scoring guidelines and sample responses of AP English Language and Composition courses. Each essay was coded by at least two human judges, who were asked to identify sentences that were candidate thesis or conclusion statements and to rate the candidate sentences from 1 to 3 (i.e., 1: vague or incomplete, 2: simple but acceptable, 3: sophisticated), based on criteria in the instruction manual. Table 1 shows the distribution and example sentences in each category.

Table 1. Distribution and example sentences from different ratings categories

Rating (%)	Example	Reason
1- Incomplete (%15)	There are contributing factors of our violent society but there are some possible solutions.	Too vague
2- Simple (%39)	As a result of gender roles in Africa, life for women is extremely challenging.	Does not mention the challenges.
3- Sophisticated (%46)	Including music programs in schools is beneficial because music improves students' academic, social and emotional lives.	Provides different reasons for the central claim.

We used Cohen's Kappa [18] to evaluate the agreement between judges on both sentence level (whether a sentence is a thesis/conclusion statement) and essay level (absence/presence of thesis). Kappa was calculated for all 8 writing assignments. If Kappa fell below 0.6, we asked the judges to review the instruction manual and redo the coding until their agreement was acceptable.

We used an iterative process to find the most predictive features for identifying thesis and conclusion sentences in essays. Starting with 42 basic computational linguistic features inspired by [11], such as positional, syntactic, and cue term features, we used several feature selection algorithms to select the most predictive features. We tried different combinations of the predictive features and also added some semantic and rhetorical structure features to improve the model's accuracy. Finally, we picked 19 features in three categories that were most predictive.

Positional Features: We used 3 positional features: paragraph number, sentence number in the paragraph, and type of paragraph (first, body, and last paragraph). We also used the same positional baseline as [11] in order to compare our results with their model. The positional baseline predicts all sentences in the first paragraph as a *thesis statement* and all sentences within the last paragraph as *conclusion sentences*.

Sentence Level Features: We used a number of sentence level features based on the syntactic, semantic, and dependency parsing of the sentence. Based on our feature selection process, prepositional and gerund phrases are highly predictive of thesis and conclusion sentences. The number of adjectives and adverbs within the sentence is also highly correlated with a sentence being a thesis or conclusion statement. A set of frequent words was also predictive for thesis and conclusion sentences (e.g., "although", "even though", "because", "due to", "led to", "caused"), and we used the number of occurrences of these words in a sentence as a feature in our model.

Essay Level Features: We used 4 essay level features: number of keywords among the most frequent words of the essay, number of words overlapping with the assignment prompt, and a sentence importance score based on Rhetorical Structure Theory (RST) adapted from [19]. Table 2 shows the top 5 most predictive features for each category based on the Gini Coefficient [20] attribute selection method. This method considers the prior distribution of the classes and looks for the largest class in the training set (in our case sentences that are not the thesis) and tries to isolate it from other classes, which is suitable based on the nature of our classification task.

Table 2. Top 5 most predictive features for each category based on Gini Coefficent

Ranking	Thesis	Conclusion
1	Last Sentence	Last Paragraph
2	First Paragraph	Keyword Overlap
3	Common Words	Common Words
4	Keyword Overlap	Number of Adjectives
5	Number of Noun Phrases	Number of Noun Phrases

3 Results and Discussion

After a data cleaning and pre-processing step, we created feature vectors for all of the sentences in the training set essays. Our target class had 3 labels: "thesis", "conclusion", and "other". We considered sentences rated 2 and 3 as thesis and conclusion statements and put the ones rated 1 (incomplete) into the "other" category. We evaluated our model on two levels: sentence level and essay level, and compared its performance against the positional baseline and human annotated data.

We used 3 classifiers in RapidMiner [21] in order to develop the sentence level models: Naïve Bayes, Decision Tree, and Support Vector Machine (SVM). We used 10-fold essay stratified cross validation in order to evaluate our models on sentence level. In order to evaluate the models on essay level, we aggregated the results of the sentence level model in order to predict whether an essay contains a thesis/conclusion statement or not. Table 3 shows the performance of the 3 classifiers based on average Precision (P), Recall (R), and F-measure (F) among all 10 rounds of cross-validation. We use F, the harmonic mean of P and R, as our main performance evaluation metric.

Table 3. Average performance of 3 models and the positional baseline on development set

	Thesis			Conclusion			Essay		
Classifier	P	R	F	P	R	F	P	R	F
Positional Baseline	0.53	0.89	0.50	0.51	0.89	0.46	0.61	0.78	0.54
Naïve Bayes	0.62	0.76	0.68	0.57	0.72	0.62	0.71	0.66	0.67
Decision Tree	0.75	0.68	0.71	0.62	0.43	0.51	0.75	0.71	0.73
SVM	0.85	0.66	0.74	0.67	0.41	0.51	0.69	0.64	0.66

In order to indicate how well the models generalize to new essays, we evaluated our models on an unseen test set. Table 4 shows the performance of 3 models.

Table 4. Average performance of 3 models and the positional baseline on unseen test set

	Thesis			Conclusion			Essay		
Classifier	P	R	F	P	R	F	P	R	F
Positional Baseline	0.58	0.88	0.57	0.58	0.84	0.55	0.58	0.84	0.55
Naïve Bayes	0.70	0.79	0.74	0.65	0.69	0.67	0.63	0.65	0.64
Decision Tree	0.82	0.84	0.83	0.49	0.75	0.59	0.75	0.73	0.74
SVM	0.82	0.65	0.72	0.60	0.54	0.56	0.62	0.58	0.60

The results show that all three models outperform the positional baseline. While the SVM classifier had the best precision on both development and test set at the sentence level, the Decision Tree classifier achieved higher recall and better overall performance at the essay level. Since we are not using the same training and test set as in [11], it is not valid to compare the exact value reported for P, R, and F. However, because we use the same positional baseline, and the results of the baseline can be considered as a rough estimate of the quality of the essays, we can compare the systems in terms of improvement over the baseline. In the thesis detection category, their highest reported improvement (regarding F) over the positional baseline is 0.22 while our best improvement is 0.24 on the development set and 0.26 on the unseen test set. In the conclusion detection category, their highest reported improvement is 0.23 while our best improvement is 0.16 development set and 0.12 on the unseen test set. In general, we have low performance in the conclusion category because the essays in our training set are first drafts of writing assignments and the students tend to spread the summary of their arguments across multiple sentences and our current model only works on the sentence level.

In conclusion, our study shows that even with a relatively small corpus of essays, a computational linguistic model can identify core aspects of students' essays. Our first priority was to detect the presence of thesis or conclusion statements within the student essays to provide instant feedback to authors upon submission. The second priority was to identify the particular sentences, to direct reviewers' attention so that they focus some comments on how well the author has framed and supported his/her argument.

Our next step is to embed our model into the SWoRD peer review system and evaluate its impact on the quality of student reviews. The peer-review nature of SWoRD gives us a unique opportunity benefit from both author and peer feedbacks in order to evaluate and refine our model while being used. We also plan to extend the model to detect other core elements of student essays such as topic sentences and supporting materials in order to provide feedback and scaffolding.

Acknowledgements. This research is supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A120370 to the University of Pittsburgh. The opinions expressed are those of the authors and do not represent the views of the Institute or the U.S. Department of Education.

References

1. National Center for Education Statistics, The Nation's Report Card: Writing, Institute of Education Sciences, US Department of Education, Washington, D.C. (2012)
2. Sadler, P., Good, E.: The impact of self-and peer-grading on student learning. Educational Assessment 11(1), 1–31 (2006)
3. Wooley, R., Was, C., Schunn, C., Dalton, D.: The effects of feedback elaboration on the giver of feedback. Paper presented at the 30th Annual Meeting of the Cognitive Science Society (2008)
4. Cho, K., Schunn, C.: Developing writing skills through students giving instructional explanations. In: Stein, Kucan (eds.) Instructional Explanations in the Disciplines. Springer, NY (2010)
5. Goldin, I.M., Ashley, K., Schunn, C.D.: Redesigning Educational Peer Review Interactions Using Computer Tools: An Introduction. Journal of Writing Research 4(2), 111–119 (2012)
6. Hansen, J., Liu, J.: Guiding principles for effective peer response. ELT J. 59(1), 31–38 (2005)
7. Durst, R.: Cognitive and Linguistic Demands of Analytic Writing. Research in the Teaching of English 21(4), 347–376 (1987)
8. National Assessment of Educational Progress, Writing Framework for the, National Assessment of Educational Progress (2011)
9. Shermis, M.D., Burstein, J., Higgins, D., Zechner, K.: Automated essay scoring: Writing assessment and instruction. International Encyclopedia of Education 4, 20–26 (2010)
10. Burstein, J., Marcu, D.: A machine learning approach for identification thesis and conclusion statements in student essays. Computers and the Humanities 37(4), 455–467 (2003)
11. Burstein, J., Marcu, D., Knight, K.: Finding the WRITE stuff: Automatic identification of discourse structure in student essays. IEEE Intelligent Systems 18(1), 32–39 (2003)
12. Roscoe, R.D., McNamara, D.S.: Writing pal: Feasibility of an intelligent writing strategy tutor in the high school classroom. Journal of Educational Psychology 105(4), 1010 (2013)
13. Crossley, S.A., McNamara, D.S.: Understanding expert ratings of essay quality: Coh-Metrix analyses of first and second language writing. International Journal of Continuing Engineering Education and Life Long Learning 21(2), 170–191 (2011)
14. Cho, K., Schunn, C.D.: Scaffolded writing and rewriting in the discipline: A web-based reciprocal peer review system. Computers & Education 48(3), 409–426 (2007)
15. Fails, J.A., Olsen Jr, D.R.: Interactive machine learning. In: Proceedings 8th International Conference on Intelligent User Interfaces, pp. 39–45 (2003)
16. De La Paz, S., Graham, S.: Explicit teaching strategies, skills and knowledge: Writing instruction in middle school classrooms. Journal of Educational Psychology 94(4), 687–698 (2002)
17. Durst, R., Laine, C., Schultz, L.M., Vilter, W.: Appealing Texts The Persuasive Writing of High School Students. Written Communication 7(2), 232–255 (1990)
18. Fleiss, J.L., Cohen, J., Everitt, B.S.: Large sample standard errors of kappa and weighted kappa. Psychological Bulletin 72(5), 323 (1969)
19. Marcu, D.: Discourse trees are good indicators of importance in text. Advances in Automatic Text Summarization, 123–136 (1999)
20. Kakwani, N.: On a class of poverty measures. Econometrica: Journal of the Econometric Society, 437–446 (1980)
21. Mierswa, I., Wurst, M., Klinkenberg, R., Scholz, M., Euler, T.: Yale: Rapid prototyping for complex data mining tasks. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 935–940 (2006)