Toward Using Text Summarization for Essay-Based Feedback

Jill Burstein
ETS Technologies
Rosedale Road, MS 11R
Princeton, NJ 08541
USA
jburstein@ets.org

Daniel Marcu
Information Sciences Institute/ University of Southern California
4676 Admiralty Way, Suite 1001
Marina Del Rey, CA 90292
USA
marcu@isi.edu

Abstract

We empirically study the impact of using automatically generated summaries in the context of electronic essay rating. Our results indicate that 40% and 60% discourse-based essay summaries improve the performance of the topical analysis module of *e-rater*. *E-rater* is a system that electronically scores GMAT essays. We envision using automatically generated essay summaries for instructional feedback, as a supplement to the *e-rater* score.

1. Introduction

Educational Testing Service (ETS) has been successfully using their automated essay scoring system, *e-rater*, as one of the two readers for the Graduate Management Admissions Test (GMAT) Analytical Writing Assessment (AWA). The system was designed to score essays based on holistic scoring guides (scoring rubrics) – see http://www.gmat.org for sample scoring rubrics. Holistic scoring guides instruct the human reader to assign an essay score based on the general quality of writing characteristics in an essay. For instance, the reader is to assess the overall quality of the writer's organization of ideas, syntactic variety, and appropriate vocabulary use. Accordingly, *e-rater* assigns a score based on its identification of essay features related to syntax, discourse structure, and essay topic.

E-rater is undergoing further development in a web-based essay evaluation, *Criterion*. Our aim is to eventually provide test-takers not only with the score of their essays, but also with instructional feedback that they may use to improve their writing skills. Feedback may delineate, for example, syntactic errors, inappropriate rhetorical renderings, and omitted topics. Syntax- and discourse-related instructional feedback can be provided on the basis of only the essay under scrutiny. Topic-related feedback could incorporate a synthesis across multiple essays. In giving topic-relevant feedback. To be effective, topical feedback needs to assess how well a given essay covers the important topics that are specific to a given essay question (*prompt*) and how salient these topics are in the essay under scrutiny. A good essay should not only cover most of the important topics, but should also present them in a salient manner.

We hypothesize that summaries can be used in the context of instructional feedback to determine the most important points of essays. We envision at least two possible uses of essay summaries.

1. For any given prompt one can, for example, build individual summaries of all essays of the highest score; use sentence-based similarity measures to determine the topics that occur frequently in these essays; and present these topics to a test-taker. Test-takers would then be able to assess what topics they might have included in order to be given a high score. (In some instances, the important *gold standard topics* may be provided by the test developers.) This application would be particularly useful for subject-based writing tests in which factual information is required in the essay response.
2. For any given essay, one can build a summary and can present it to the test-taker in a format that makes explicit whether the main points in the summary cover the topics that are considered important for the prompt. One way of doing this might be to present to test-takers, summaries of other essays that received a high score. Test-takers would be able to assess whether the rhetorical organization of their essays makes the important topics salient.

Currently, the College Board's Advanced Placement (AP) writing instruction software for English Literature and U.S. History use sample essays at different score points to illustrate different writing competencies to test-takers (Burstein and Boodoo, submitted). We believe that by exploiting essay summaries along the lines described above, students will be able to improve their skills with respect to selecting adequate content and presenting the content so that the main points are emphasized.

In this paper, we take the first step toward evaluating the usability of automatically generated summaries for an application that automatically provides instructional feedback. More precisely, we examine the effectiveness of using summaries for eliminating the noise in essay data, while maintaining substantive information, central to the essay response.

2. Approach

Consider for a moment the document classification task used by DARPA during the SUMMAC evaluation (SUMMAC, 1998). With respect to the classification task, document summaries were considered adequate when on their basis, one can accurately classify documents as belonging to the correct category. To assess the effectiveness of essay summaries, we take a similar approach. Previous research (Burstein, *et al*., 1998) has shown that the topical analysis component of *e-rater* is one of the strongest indicators of the score of an essay. Essays having similar scores tend to use the same vocabulary and discuss the same topics. As a result, test essays can be scored automatically on the basis of their similarity with essays in a training sample of essays that have been manually scored by human readers. If a test essay uses the same vocabulary and addresses the same topics as a large number of essays of, let's say, score 5, it is likely that the test essay should be assigned a score of 5 too. Using only the topical analysis component of *e-rater*, we can assign scores that yield exact or adjacent agreement with human judges in about 83% of the cases.

The assumption that underlies the SUMMAC classification task is that one should be able to determine the category to which a document belongs using only a summary of that document (provided the summary is adequate). The SUMMAC evaluation and other research (SUMMAC, 98; Mani and Bloedorn, 98) has shown that when summaries are used, the categorization performance decreases slightly, but the time needed to carry out the classification task decreases substantially. We hoped we could demonstrate a similar effect on essay summaries. In contrast with the SUMMAC classification task, which requires the participation of humans, in the context of essay scoring we assess the performance of a summarization system using the topical analysis component of *e-rater*. If we are able to generate summaries without degrading the performance of the topical analysis component of *e-rater*, we assume that the summaries are adequate for the purpose of determining the main points in an essay.

In this paper, we report on experiments that assess the effect of different summarization systems that operate at different compression rates for the task of eliminating the non-essential information in essays. To our knowledge, this is the first extrinsic evaluation that demonstrates that the use of summaries leads to an increase in performance that is statistically significant. For question-answering and document classification (SUMMAC, 1998; Mani and Bloedorn, 1997), the use of summaries saves time but leads to a small decrease in performance. For information retrieval (Corston-Oliver and Dolan, 1999), the use of summaries saves disk space but leads to a small decrease in recall and precision. Our experiments show that the class of summaries built from the 40% and 60% discourse-based summaries of a text can be successfully used to eliminate noise in essays.

In the rest of the paper, we will explain first how the topical analysis component of *e-rater* works. Secondly, we present the summarization algorithms that we rely upon and discuss the results of our experiments.

3. Topical Analysis By-Argument

To capture use of vocabulary (or identification of topic), *e-rater* uses content vector analyses that are based on the vector-space model, commonly found in information retrieval applications (Salton, 1989). One way that content vector analysis is performed in *e-rater* is at the level of the argument. Training essays are converted into vectors of word frequencies, and the frequencies are then transformed into word weights.[1] These weight vectors populate the training space. To score a test essay, it is converted into a weight vector, and a search is conducted to find the training vectors most similar to it, as measured by the cosine between the test and training vectors. The closest matches among the training set are used to assign a score to the test essay.

---

[1] Word (or term) weight reflects not only a word's frequency in the essay but also its distribution across essays. E-rater's formula for the weight of word *w* in essay *j* is:

$weight_{wj}=(freq_{wj}/maxfreq_{j}) * \log(nessays/essays_{w})$

where $freq_{wj}$ is the frequency of word *w* in essay *j*, $maxfreq_{j}$ is the frequency of the most frequent word in essay *j*, nessays is the total number of training essays, and $essays_{w}$ is the number of training essays that contain *w*. The first part of the formula measures the relative importance of the word in the essay. The second part gauges its specificity across essays, so that a word that appears in many essays will have a lower weight than one which appears in only a few. In the extreme case, a word that appears in all essays (e.g., "the") has a weight of 0.

All of the training essays for each score category and are used to populate the training space with just 6 "supervectors", one each for scores 1-6. (This is the standard range of scores used in *holistic* essay scoring as performed by human readers.) The text in each test essay is evaluated one argument at a time. Each argument is converted into a vector of word weights and compared to the 6 vectors in the training space. The closest vector is found and its score is assigned to the argument. This process continues until all the arguments have been assigned a score. The overall score for the test essay is an adjusted mean of the argument scores using the following formula, rounded to the nearest integer:

$$\text{Score for test essay } t = (\sum \text{argscore}_j + \text{nargs}_t)/(\text{nargs}_t + 1)$$

where j ranges over the arguments in test essay t, argscore_j is the score of argument j, and nargs_t is the number of arguments in t. Using this adjusted mean has the overall effect of reducing, slightly, the score for essays with few arguments, and of increasing somewhat the score of essays with many arguments.

4. Summarizers

For the purpose of this experiment, we relied on two summarizers.

1. The *Position-based* summarizer assumes that the most important sentences are those that occur at the beginning of an essay (Baxendale 1958, Edmundson, 1968, Lin and Hovy, 1997). This summarizer constructs extracts of compression *k* by selecting the first *k%* words in an essay. Given that essays are written by students under time pressure, we expect position-based extracts might reflect the most important information in an essay.
2. The *Discourse-based* summarizer assumes that the most important clauses in a text can be determined on the basis of the rhetorical structure of that text. The discourse-based summarizer derives first the rhetorical structure of the essay under scrutiny (Mann and Thompson, 1988, Marcu, 1997). This structure is a binary tree that associates to every text span a status, which can be *nucleus* or *satellite*. A nucleus node subsumes text that is important; a satellite node subsumes text that is subsidiary to the text subsumed by a nucleus. On the basis of this structure and the statuses of the nodes in the structure, the discourse-based summarizer associates an importance score to each clause in a text; the closer a clause is to the root of the tree, the higher the score. A *k*% extract of the essay is determined by selecting the *K* clauses of highest score, where the text subsumed by the *K* clauses represents k% of the original text. The summaries extracted in this manner reflect a global view with respect to the relation between rhetorical structure trees and importance.

For our experiments, we used an existing discourse-based summarizer (see (Marcu, 1997, 1999)). A more detailed description of the rhetorical parsing strategy from which the discourse-based summaries are derived is presented below. Figure 1 illustrates a RST parse tree.

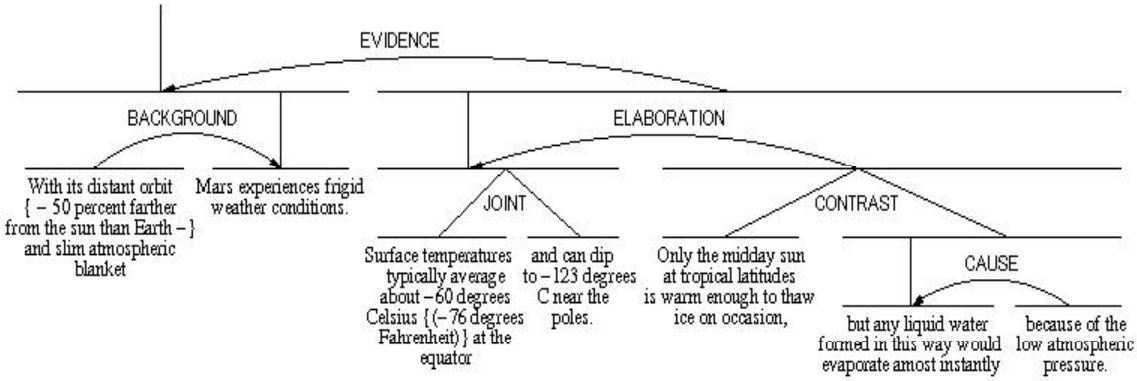

Figure 1: An example of RST parse tree.

4.1 Using Rhetorical Structure Theory to Identify Discourse Strategy in Text

According to rhetorical structure theory (RST) (Mann and Thompson, 1988), one can associate a rhetorical structure tree to any text. The leaves of the tree correspond to elementary discourse units and the internal nodes correspond to contiguous text spans. Text spans represented at the clause and sentence level. Each node in a tree is characterized by a *status* (nucleus or satellite) and *a rhetorical relation*, which is a relation that holds between two non-overlapping text spans. The distinction between nuclei and satellites comes from the empirical observation that the nucleus expresses what is more essential to the writer's intention than the satellite; and that the nucleus of a rhetorical relation is comprehensible independent of the satellite, but not vice versa. When spans are equally important, the relation is multinuclear.

Rhetorical relations reflect semantic, intentional, and textual relations that hold between text spans as is illustrated in Figure 1. For example, one text span may elaborate on another text span; the information in two text spans may be in contrast; and the information in one text span may provide background for the information presented in another text span. Figure 1 displays in the style of Mann and Thompson (1988) the rhetorical structure tree of a text fragment. In Figure 1, we can see that both texts within each text span in the contrast relation are clauses. In Figure 1, nuclei are represented using straight lines; satellites using arcs. Internal nodes are labeled with rhetorical relation names.

Intuitively, rhetorical structure representations of the kind shown in Figure 1 can be used to determine the rhetorical structure of essays. More specifically, we hypothesize that features that characterize the rhetorical structure of texts can be used as indicators for modeling the holistic assessments that concern the logical organization and rhetorical strategy in essays. Therefore, to identify essay summaries, essays are parsed using an existing RST-based parser (Marcu, 1997). Using the RST trees for essays, an algorithm is applied (above) to extract the essay summaries at desired compression rates (k%).

5. Evaluation of Essay Summaries

For the purpose of evaluating the use of summary data for its use as instructional feedback, we have used 20 sets of essay responses from 20 different GMAT test questions (prompts). Ten of these were *argument* prompts and ten were *issue* prompts. Argument prompts require a test-taker to write an essay that evaluates an argument. These essays are typically more focused and use more overt discourse cues than issue prompts. Issue prompts are more general; they concern questions that ask students to construct responses by including their personal observations, experiences, and opinions. We used an even set of each prompt type so that we would also be able to evaluate if there was an effect of prompt type. For each prompt, we used a set of 270 manually scored essays for training and a set of 500 essays for testing. In the training samples, the distribution of essays at each score point is as follows: 5 0's, 15 1's, and 50 at each score point from 2 through 6. For the sample of essays used in testing, the distribution at each score point is proportional to its average distribution in the operational testing environment.[2]

For each prompt type, we first evaluated the performance of the topical analysis component using the full texts. We then extracted 20%, 40%, and 60% summaries using both the position- and discourse-based summarizers. We compared the performance of the topical component of *e-rater* in two working modes.

1. In *full-text mode*, the score of a test essay is determined by measuring the similarity between the full texts in the training corpus and a k% summary of the test essay.
2. In *summarized-text mode*, the score of an essay is determined by measuring the similarity between the k% summaries of the texts in the training corpus and a k% summary of the test essay.

Tables 1 through 4 summarize our results across all prompts. These results show the mean performance of the topical analysis component of *e-rater*, given positional (Pos Sum) and discourse-based summarization (Disc Sum) methods that are employed both in full-text and summarized-text mode. Mean agreement performance is shown for exact agreement with human reader score, and for exact-plus-adjacent agreement (that is, where there is no more than a 1-point difference between the human reader score and the topical analysis score).[3] No effect occurred for individual prompt types (i.e., argument or issue types); therefore, our analysis is performed across all prompts.

In general, 40% and 60% summaries using either full-text or summarized-text mode improve the performance of the topical analysis component. Using 20% summaries in the discourse-based summarization methods consistently degraded the performance of the topical analysis module. Position-based summarization degraded performance of the topical analysis component across the board. Results in italicized boldface indicate an improvement in performance, using summarization.

---

[2] The approximate operational distributions at each score point are as follows: 4% at score point 1, 12% at score point 2, 25% at score point 3, 31% at score point 4, 20% at score point 5, and 8% at score point 6.

[3] When two human readers are used to score essays for a real test administration, such as GMAT, their scores are considered to be in agreement if they assign a score within one-point of each other. Two human reader scores are only considered to be discrepant if they differ by two or more points.

Table 1: Topical analysis performance results in *full-text mode*: Exact Agreement

Full-Text/Full-Text	**.3499**	
	Pos Sum	Disc Sum
Full-Text / Summarized-Text 20%	.2516	.3111
Full-Text / Summarized-Text 40%	.2702	***.3741***
Full-Text / Summarized-Text 60%	.2803	***.3661***

Table 2: Topical analysis performance results in *summarized-text mode*: Exact Agreement

Full-Text/Full-Text	**.3499**	
	Pos Sum	Disc Sum
Summarized-Text 20%/ Summarized-Text 20%	.2640	.3252
Summarized-Text 40%/ Summarized-Text 40%	.2736	***.3684***
Summarized-Text 60%/ Summarized-Text 60%	.2738	***.3650***

Table 3: Topical analysis performance results in *full-text mode*: Exact+ Adjacent Agreement

Full-Text/Full-Text	**.8321**	
	Pos Sum	Disc Sum
Full-Text/ Summarized-Text 20%	.6642	.7671
Full-Text/ Summarized-Text 40%	.6956	***.8626***
Full-Text/ Summarized-Text 60%	.7089	***.8579***

Table 4: Topical analysis performance results in *summarized-text mode*: Exact+ Adjacent Agreement

Full-Text/Full-Text	**.8321**	
	Pos Sum	Disc Sum
Summarized-Text 20%/ Summarized-Text 20%	.6811	.7721
Summarized-Text 40%/ Summarized-Text 40%	.6954	***.8560***
Summarized-Text 60%/ Summarized-Text 60%	.7116	***.8481***

We evaluated the significance of the improved performance for the 40% and 60% discourse-based summarization methods, since only these cases showed a consistent improvement in performance. For both exact, and exact-plus-adjacent performance, analysis of variance were run to assess performance effects of the topical analysis component using standard full text, and 40% and 60% discourse-based summaries, both in full-text mode and summarized-text mode. The results of the analysis of variance indicate significant effects for exact, and exact-plus-adjacent performance.

For exact agreement, there was a significant effect with summarized-text mode, where $F(2,19) = 4.52$, $p < .01$. For full-text mode, there was also a significant effect, where $F(2,19) = 6.78$, $p < .003$. For exact-plus-adjacent agreement, for summarized-text mode, a significant effect showed up, where, $F(2,19) = 8.87$, $p < .0006$, and the effect was also significant for full-text mode, $F(2,19) = 17.90$, $p < 3.33E-06$.

Overall, there is a statistically significant indication of improvement in performance of topical analysis with the use of 40% and 60% discourse-based summaries. This suggests that vocabulary in these summaries represents salient concepts in essays. The summaries might be used to generate relevant text extracts from which instructional feedback might be generated to illustrate substantive information in essays at different levels of writing competency.

6. Discussion and Conclusions

The results in this paper suggest that vocabulary in 40% and 60% discourse-based summaries contain substantive information from the text of original essays. The topical analysis by-argument component uses training samples at each score point to recognize word use in essays at different score points. Since these essays are written under a timed testing situation, there is little time for editing and smoothing out of the text. As a result, the text of an essay could have a lot of noise, such as repetitive statements and extraneous comments not necessarily central to the main arguments in the essay. It appears that by using these summaries, we are able to eliminate some of the noise and in a sense ‘clean up’ the essay so that the topical analyzer has access to the more substantive vocabulary in the essay. Information in summaries could, therefore, be useful as instructional feedback to illustrate substantive information in essays at different levels of writing competency. Test-takers might use summarized versions of their essays to evaluate how the substantive information in their own essay compares to an essay that received a higher score.

We hypothesize that this technique could be highly valuable for providing feedback for subject-based essays in which test-takers are required to provide factual information related to the question topic, such as College Board’s AP exams in U.S. History. Refinement of this technique could lead to instructional feedback that would help test-takers focus the discussion points or arguments in their essays. If the main points in higher scoring essays could be identified, they could be used to illustrate the points that were lacking from lower scoring essays.

Acknowledgements

We would like to thank Martin Chodorow for his comments on earlier versions of this paper.

References

BAXENDALE, P.B. 1958. Machine-Made Index for Technical Literature—An Experiment. *IBM Journal* (October) 354–361.

CORSTON-OLIVER S. AND W. DOLAN (1999). Less is more: Eliminating Index Terms from Subordinate Clauses. *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pp. 348-356.

BURSTEIN, J. AND BOODOO, G. (submitted). Automated Scoring for Advanced Placement Essay Responses for U.S. History and English Literature and Composition. Educational Testing Service, Princeton, NJ.: College Board Report.

BURSTEIN, J., K. KUKICH, S. WOLFF, C. LU, M. CHODOROW, L. BRADEN-HARDER, AND M. DEE HARRIS. (1998). Automated Scoring Using A Hybrid Feature Identification Technique. In the *Proceedings of the Annual Meeting of the Association of Computational Linguistics*, Montreal, Canada.

EDMUNDSON, H.P. (1968). New Methods in Automatic Extraction. *Journal of the ACM* 16(2), 264–285.

HOVY, E.H. AND LIN, C-Y. 1998. Automated Text Summarization in SUMMARIST. In M. Maybury and I. Mani (eds), *Intelligent Scalable Summarization Text Summarization*, pp. 81-94. The MIT Press.

MANI I. AND E. BLOEDORN. 1997. Multi-document Summarization by Graph Search and Matching. *Proceedings of the National Conference on Artificial Intelligence, (AAAI).* Providence, Rhode Island, pp. 622-628.

MANN, W.C. AND S.A. THOMPSON. 1988. Rhetorical Structure Theory: Toward a Functional Theory of Text Organization. *Text* 8(3), 243–281. Also available as USC/Information Sciences Institute Research Report RR-87-190.

MARCU, D. (1997). The Rhetorical Parsing of Natural Language Texts. *The Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, pp. 96-103.

MARCU, D. (1999). Discourse trees are good indicators of importance of text. In I. Mani and M. Maybury eds., *Advances in Automatic Text Summarization*, pp. 123-136. The MIT Press.

SALTON G. (1989). *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer.* Addison-Wesley Publishing Co.

SUMMAC (1998). Firmin Hand, T. and B. Sundheim (eds). TIPSTER-SUMMAC Summarization Evaluation. *Proceedings of the TIPSTER Text Phase III Workshop*. Washington.